SIAM J. CONTROL OPTIM.
Vol. 37, No. 1, pp. 1–32

SMALL PARAMETER LIMIT FOR DISCRETE-TIME PARTIALLY OBSERVED RISK-SENSITIVE CONTROL PROBLEMS*

FRANCESCA ALBERTINI† AND PAOLO DAI PRA†

Abstract. We show that risk-sensitive control problems and deterministic dynamic games can be connected, under rather mild assumptions, by a small noise limit. In order to control this limit, new techniques are developed to study propagation of large deviations through conditional probabilities.

Key words. risk-sensitive control, dynamics games, large deviations

AMS subject classifications. Primary, 93E20; Secondary, 60F10

PII. S0363012996307801

1. Introduction.

Properties of risk-sensitive control problems and their connections with dynamic games have been widely investigated in recent years [16, 18, 11, 8, 9, 12, 13, 14, 6, 4], in part inspired by seminal results for linear quadratic models contained in [10, 17, 1]. In particular, it has been shown [8, 9, 12, 13, 4] that under a suitable small parameter limit (*small noise limit*) a family of risk-sensitive stochastic control problems becomes equivalent to a deterministic dynamic game. In other words, this means that optimal risk-sensitive control with small noise and suitably rescaled risk parameter is almost equivalent to deterministic robust control (worst-case approach).

Although this result is conceptually natural, its proof usually involves rather sophisticated mathematical techniques, and fairly strong requirements on the model. For continuous-time, totally observable systems a quite satisfactory theory has been developed in [8, 9] by using viscosity solution techniques to analyze the Hamilton–Jacobi equation associated to the optimal control problem. It has been shown, in particular, that the value function of the risk-sensitive control problem converges, as the noise parameter goes to zero, to the upper value function of a related two players, zero-sum differential game.

The analysis of the small parameter limit for nonlinear, partially observed, risk-sensitive control problems has been initiated by P. Whittle [18], whose mostly nonrigorous results have inspired most of the further development. A considerable advancement in the understanding of these models is represented by the results in [12, 13], where the *information state* approach is used. This approach consists in reformulating the partially observed control problem as a completely observed one, in a way that, in a suitable sense, is "preserved" in the small parameter limit. One consequence of this method is that it provides a natural notion of information state for the limiting dynamic game. In [12] the information state approach is applied to discrete-time systems. The result obtained is parallel to the one given in [8], i.e., the connection with partially observed dynamic game is established in terms of the convergence of the value function of the equivalent totally observed model. This value function is the

*Received by the editors August 5, 1996; accepted for publication (in revised form) November 11, 1997; published electronically September 25, 1998. This research was supported in part by US Air Force grant F49620-95-1-0101.

http://www.siam.org/journals/sicon/37-1/30780.html

†Università di Padova Dipartimento di Matematica Pura e Applicata, Via Belzoni 7, 35131 Padova, Italy (albertin@math.unipd.it, daipra@math.unipd.it).

solution of a dynamic programming equation in an infinite-dimensional space. The small parameter limit of this equation is obtained by using large deviation techniques.

The corresponding result in continuous time has been obtained, at a nonrigorous level, in [13]. In this context, one has to deal with the small parameter limit of a Hamilton–Jacobi equation (Mortensen's equation) with infinite-dimensional state space; the current mathematical understanding of this problem has not allowed a complete proof yet. A different approach to the small parameter limit in continuous time is used in [3], where, rather than the convergence of the value function, it is (*rigorously*) shown the convergence of the *cost functional* for *any* control u in a suitably defined admissible class. This requires considerable work for controlling the small parameter limit of the information state, but avoids the use of Mortensen's equation.

In this paper we consider discrete-time, finite time-horizon partially observed systems, and we develop further some large deviation techniques that were introduced in [4] for totally observed systems. The models for which we can analyze the small parameter limit include those of type

$$\begin{aligned} x_{n+1} &= f_n(x_n, u_n, w_n), \\ y_n &= \phi_n(x_n, v_n) \end{aligned} \tag{1.1}$$

for $n = 0, \dots, N-1$, with $x_n \in \mathcal{X}, u_n \in \mathcal{U}, w_n \in \mathcal{W}, y_n, v_n \in \mathbb{R}^d$, where $\mathcal{X}, \mathcal{W}$ are metric spaces and $\mathcal{U}$ is a compact metric space. Moreover, w_n, v_n are independent random variables, $w_n \sim \mu_n^\epsilon$, $v_n \sim \nu_n^\epsilon$, where $(\mu_n^\epsilon)_{\epsilon>0}$, $(\nu_n^\epsilon)_{\epsilon>0}$ are families of probability measures satisfying a large deviation principle (see section 2). Some further regularity assumptions will be needed for ϕ_n and ν_n^ϵ (see section 4), while f_n is supposed only to be continuous. We associate with (1.1) a cost functional of the form

$$J^\epsilon(\mathbf{u}) = \epsilon \log E\Big\{ \exp\Big[\epsilon^{-1}\Big(\sum_{n=0}^{N-1} g_n(x_n, u_n) + g_N(x_N)\Big)\Big]\Big\} \tag{1.2}$$

defined for the control sequences $\mathbf{u} = (u_0, \dots, u_{N-1})$ that are nonanticipative functions of the output sequence $(y_0, \dots, y_N)$. Note that the parameter ϵ appears both in the noise distribution and in (1.2), where it can be interpreted as a *risk parameter*, a measure of controller's aversion to risk. We show that, as $\epsilon \to 0$, the risk-sensitive control problem (1.1), (1.2) converges, in a suitable sense, to the deterministic game with dynamics (1.1) (where w_n, v_n are thought of as deterministic but unknown disturbances) and cost

$$J(\mathbf{u}) = \sup_{\mathbf{v},\mathbf{w}} \Big[\sum_{n=0}^{N-1} \Big(g_n(x_n, u_n) - h_n(w_n) - k_{n+1}(v_{n+1})\Big) + g_N(x_N)\Big], \tag{1.3}$$

where h_n, k_n are *rate functions* (see section 2) associated with $\mu_n^\epsilon, \nu_n^\epsilon$. This convergence is expressed, similarly to [8] and [12], in terms of the convergence of the value function for an equivalent totally observed problem, so that our result can be seen as a generalization of [12]. In fact, models of type (1.1), (1.2) include the ones studied in [12], but we can deal with fairly more general noise distribution, state-space and dynamical equations. The results of this paper have been, in part, announced in [5], where, however, much stronger conditions were required.

This paper is organized as follows. In sections 2 and 3 we develop some new large deviation techniques that are suitable for the problem we deal with. Section 4, which contains the main results of this paper, is devoted to the analysis of the small parameter limit for risk-sensitive control problems.

2. Preliminary notions. In this section we recall some notions from large deviation theory that will be used throughout the paper, and introduce some new ones. We let $\mathcal{X}$ be a metric space. All measures on $\mathcal{X}$ are intended to be defined on its Borel σ-field.

DEFINITION 2.1. *A family of probability measures* $\{P^\epsilon : \epsilon > 0\}$ *on* $\mathcal{X}$ *is said to satisfy a* large deviation principle (*LDP*) *with rate function* $H : \mathcal{X} \to [0, +\infty]$ *if*

i) H *is lower semicontinuous and* $\{x : H(x) \le l\}$ *is compact for every* $l \ge 0$;

ii) *for every* $A \subset \mathcal{X}$ *measurable*

$$\begin{aligned}-\inf_{x\in\mathring{A}} H(x) &\le \liminf_{\epsilon\to 0} \epsilon \log P^\epsilon(A)\\ &\le \limsup_{\epsilon\to 0} \epsilon \log P^\epsilon(A) \le -\inf_{x\in\bar{A}} H(x),\end{aligned}$$

where $\mathring{A}, \bar{A}$ *denote respectively the interior and the closure of* A.

In [4] a modification of the above definition has been introduced for the case of a family of probability measures depending on a further parameter.

DEFINITION 2.2. *Let* Θ *be a set. A family of probability measures* $\{P^\epsilon(dx;\theta) : \epsilon > 0, \theta \in \Theta\}$ *on* $\mathcal{X}$ *is said to satisfy a* uniform large deviation principle (*ULDP*) *with rate function* $H : \mathcal{X} \times \Theta \to [0, +\infty]$ *if*

i) *for every fixed* $\theta \in T$, $H(\cdot, \theta)$ *is lower semicontinuous and* $\{x : H(x, \theta) \le l\}$ *is compact for every* $l \ge 0$,

ii) *for every* $A \subset \mathcal{X}$ *measurable and* $M > 0$,

$$\limsup_{\epsilon\to 0} \sup_{\theta\in\Theta} \left[\epsilon \log P^\epsilon(A;\theta) + \min\left(M, \inf_{x\in\bar{A}} H(x;\theta)\right)\right] \le 0$$

and

$$\liminf_{\epsilon\to 0} \inf_{\theta\in\Theta} \left[\epsilon \log P^\epsilon(A;\theta) + \inf_{x\in\mathring{A}} H(x;\theta)\right] \ge 0.$$

One of the main consequences of an LDP is the well known Varadhan's lemma [15]. In [4] the following version of Varadhan's lemma has been proved.

LEMMA 2.3. *Suppose the family* $\{P^\epsilon(dx;\theta) : \epsilon > 0, \theta \in \Theta\}$ *satisfies a ULDP. Then for every* $F : \mathcal{X} \to \mathbb{R}$ *bounded and continuous*,

$$\lim_{\epsilon\to 0} \epsilon \log \int e^{\epsilon^{-1}F(x)} P^\epsilon(dx;\theta) = \sup_{x\in X} \left[F(x) - H(x,\theta)\right] \tag{2.1}$$

uniformly for $\theta \in \Theta$.

Remark 2.1. Identity (2.1) is the crucial large deviation property in applications to risk-sensitive control. When Θ is a singleton (and therefore pointwise in θ) it is well known [7, Bryc's theorem] that, under rather mild assumptions (namely, exponential tightness, see definition below), identity (2.1) is equivalent to the LDP. It is natural to ask whether, for general Θ, (2.1) implies the ULDP. The answer is no. A simple counterexample is the following: $\mathcal{X} = \mathbb{R}$, $\Theta = [0, 1]$, $P^\epsilon(dx;\theta) = \frac{1}{2\epsilon}\chi_{[\theta-\epsilon,\theta+\epsilon]}(x)dx$, where χ denotes the characteristic function of a set. As we will see later in Remark 2.3, the family $P^\epsilon(dx;\theta)$ satisfies (2.1) with $H(x;\theta) = +\infty$ for $x \ne \theta$, and $H(\theta, \theta) = 0$. Now take $A = (-\infty, 0]$. If $P^\epsilon(dx;\theta)$ satisfied a ULDP, then $\lim_{\epsilon\to 0} \epsilon \log P^\epsilon(A;\epsilon/2) = -\infty$, since $\inf_{x\in A} H(x;\epsilon/2) = +\infty$ for all $\epsilon > 0$. However it holds that $P^\epsilon(A;\epsilon/2) = 1/4$.

We introduce now a notion which is weaker than the one of ULDF. For the rest of this section we assume Θ to be a metric space.

Definition 2.4. *A family $\{P^\epsilon(dx;\theta) : \epsilon > 0, \theta \in \Theta\}$ of positive finite measures on $\mathcal{X}$ is called a* weakly uniform large deviation family *(WULDF) with rate function $H : \mathcal{X} \times \Theta \to (-\infty, +\infty]$ if*

i) *for every fixed $\theta \in \Theta$, $H(\cdot, \theta)$ is lower semicontinuous and $\{x : H(x,\theta) \leq l\}$ is compact for every $l \in \mathbb{R}$;*

ii) *the map $\theta \to \inf_{x \in X} H(x,\theta)$ is real valued and is bounded on the compact subsets of Θ;*

iii) *for every $F : \mathcal{X} \to \mathbb{R}$ bounded and continuous*

$$\lim_{\epsilon \to 0} \epsilon \log \int e^{\epsilon^{-1}F(x)} P^\epsilon(dx;\theta) = \sup_{x \in X} \Big[F(x) - H(x,\theta)\Big] \tag{2.2}$$

uniformly for θ in the compact subsets of Θ.

Remark 2.2. For reasons that will become apparent later, we have chosen to allow P^ϵ in Definition 2.4 to be a positive finite measure, not necessarily a probability measure. For technical reasons, we will need condition ii) in Definition 2.4, which roughly says that $P^\epsilon(\mathcal{X};\theta)$ does not either go to zero or grow too fast as $\epsilon \to 0$. Indeed, by using ii) and letting $F \equiv 0$ in iii) the following statement is easy to prove: for each $K \subset \Theta$ compact, there exists $M(K) > 0$ such that, for ϵ sufficiently small,

$$e^{-\epsilon^{-1}M(K)} \leq P^\epsilon(\mathcal{X};\theta) \leq e^{\epsilon^{-1}M(K)} \tag{2.3}$$

for all $\theta \in K$. Note that, if all $P^\epsilon(dx;\theta)$ are probability measures, then ii) is automatically satisfied, since $\inf_{x \in \mathcal{X}} H(x,\theta) \equiv 0$ (see [7]).

We now state a proposition that serves both as a technical lemma for later use and as a preliminary justification of the notion of WULDF. Its proof will be given in section 3.

Proposition 2.5. *Let $\mathcal{W}$ be a metric space, $f : \Theta \times \mathcal{W} \to \mathcal{X}$ a continuous map, and $\{\mu^\epsilon : \epsilon > 0\}$ a family of probability measures on $\mathcal{W}$ that satisfy an LDP with rate function $h(w)$. Define $P^\epsilon(dx;\theta)$, a probability measure on X, by*

$$P^\epsilon(A;\theta) = \mu^\epsilon\{w : f(\theta,w) \in A\}. \tag{2.4}$$

Then $\{P^\epsilon(dx;\theta) : \epsilon > 0, \theta \in \Theta\}$ is a WULDF with rate function

$$H(x;\theta) = \inf\{h(w) : f(\theta,w) = x\}. \tag{2.5}$$

Remark 2.3. The family of probability measures in (2.4) does not necessarily satisfy a ULDP, even if Θ is compact. For example, consider $\Theta = [0,1]$, $\mathcal{X} = \mathcal{W} = \mathbb{R}$, $\mu^\epsilon(dw) = \frac{1}{2\epsilon}\chi_{[-\epsilon,\epsilon]}(w)dw$, $f(\theta,w) = \theta + w$, and we end up with the counterexample in Remark 2.1. Note that this shows that the family $P^\epsilon(dx;\theta)$ is a WULDF, since it is easy to prove that $\{\mu^\epsilon(dw) : \epsilon > 0\}$ satisfies an LDP with rate function $H(w) = +\infty$ if $w \neq 0$, and $H(0) = 0$.

We now introduce a further notion that will be useful later.

Definition 2.6. *A family $\{P^\epsilon(dx;\theta) : \epsilon > 0, \theta \in \Theta\}$ of positive finite measure on $\mathcal{X}$ is called* exponentially tight *if, for every $L > 0$ and every $K \subset \Theta$ compact, there exists $C \subset \mathcal{X}$ compact such that*

$$P^\epsilon(C^c;\theta) \leq e^{-\epsilon^{-1}L} \tag{2.6}$$

for all $\theta \in K$ and ϵ sufficiently small, where C^c is the complement of C.

Note that when the measures are probability measures and Θ is a singleton, the above definition reduces to the usual one of exponential tightness of large deviation theory [7].

We conclude this section by stating three easy lemmas that will be used in section 3.

LEMMA 2.7. *Under the assumptions of Proposition* 2.5, *let* $P^\epsilon(dx;\theta)$ *be defined by* (2.4). *If* $\{\mu^\epsilon\}$ *is exponentially tight then so is* $\{P^\epsilon(dx;\theta)\}$.

Proof. The proof is straightforward, since compactness is preserved by continuous mapping. □

LEMMA 2.8. *Suppose that* $\{P^\epsilon(dx;\theta) : \epsilon > 0, \theta \in \Theta\}$ *is a WULDF with rate function* H, *and it is exponentially tight. Then the rate function is* proper, *i.e.*, *for every* $L > 0$ *and every* $K \subset \Theta$ *compact, there exists* $C \subset \mathcal{X}$ *compact such that* $H(x;\theta) \geq L$ *for all* $(x, \theta) \in C^c \times K$.

Proof. Let $L > 0$ be given, $M > L$ and C be a compact subset of $\mathcal{X}$ such that (2.6) holds for all $\theta \in K$. Also, let C_δ denote the δ-neighborhood of C. Consider the bounded continuous function

$$F(x) = \min\left\{\frac{M}{\delta} d(x, C) - M, 0\right\}. \tag{2.7}$$

It is easily seen that $F(x) = -M$ for $x \in C$, $F(x) = 0$ on C_δ^c and $F \leq 0$. By using the definition of WULDF we have, for all $\theta \in K$,

$$\inf_{x \in C_\delta^c} H(x, \theta) \geq -\sup_{x \in \mathcal{X}}[F(x) - H(x, \theta)] = -\lim_{\epsilon \longrightarrow 0} \epsilon \log \int e^{\epsilon^{-1} F(x)} P^\epsilon(dx; \theta)$$

$$= -\lim_{\epsilon \longrightarrow 0} \epsilon \log \left[\int_C e^{\epsilon^{-1} F(x)} P^\epsilon(dx; \theta) + \int_{C^c} e^{\epsilon^{-1} F(x)} P^\epsilon(dx; \theta)\right]$$

$$\geq -\liminf_{\epsilon \longrightarrow 0} \epsilon \log \left[e^{-\epsilon^{-1} M} P^\epsilon(X; \theta) + e^{-\epsilon^{-1} L}\right] = \min\left[M - \inf_{x \in \mathcal{X}} H(x, \theta), L\right].$$

Thus, if we choose M large enough, using ii) of Definition 2.4, we have that, for all $\theta \in K$ and $\delta > 0$

$$\inf_{x \in C_\delta^c} H(x, \theta) \geq L \tag{2.8}$$

that clearly concludes the proof. □

LEMMA 2.9. *Under the assumptions of Lemma* 2.8, *let* $F^\epsilon : \mathcal{X} \to \mathbb{R}$, $\epsilon \geq 0$, *be such that* $\sup_{\epsilon \geq 0} \|F^\epsilon\|_\infty < \infty$, $F^\epsilon \to F^0$ *as* $\epsilon \to 0$ *uniformly on the compact subsets of* $\mathcal{X}$ *and* F^0 *is continuous. Then*

$$\lim_{\epsilon \longrightarrow 0} \epsilon \log \int e^{\epsilon^{-1} F^\epsilon(x)} P^\epsilon(dx; \theta) = \sup_{x \in \mathcal{X}} \left[F^0(x) - H(x, \theta)\right] \tag{2.9}$$

uniformly on the compact subsets of Θ.

Proof. If $F^\epsilon \to F^0$ uniformly in all $\mathcal{X}$ then the conclusion follows by (2.2) and by

$$\left|\epsilon \log \int e^{\epsilon^{-1} F^\epsilon(x)} P^\epsilon(dx; \theta) - \epsilon \log \int e^{\epsilon^{-1} F^0(x)} P^\epsilon(dx; \theta)\right| \leq \|F^\epsilon - F^0\|_\infty. \tag{2.10}$$

By using exponential tightness one easily reduces to this case. □

3. Propagation of WULDFs. In our analysis of the small parameter limit for risk-sensitive control problems the *filtering probabilities* and the *information states* will play a key role. They are both families of positive measures that satisfy recursive relations. To prove that they form WULDFs we show that the property of being a WULDF is preserved under four basic operations, namely, 1. *state augmentation*; 2. *composition*; 3. *contraction*; 4. *conditioning.*

In the rest of this section $\mathcal{X}, \mathcal{Y}$, and Θ are metric spaces.

PROPOSITION 3.1 (state augmentation). *Let $\{P^\epsilon(dx;\theta) : \epsilon > 0, \theta \in \Theta\}$ be an exponentially tight WULDF on $\mathcal{X}$ with rate function $H_P(x;\theta)$. Define the measures on $\mathcal{X} \times \Theta$*

$$Q^\epsilon(dx, d\zeta; \theta) = P^\epsilon(dx;\theta) \otimes \delta_\theta(d\zeta),$$

with δ denoting the Dirac measure. Then $\{Q^\epsilon(dx, d\zeta;\theta) : \epsilon > 0, \theta \in \Theta\}$ is an exponentially tight WULDF with rate function

$$H_Q(x,\zeta;\theta) = \begin{cases} H_P(x;\theta) & \text{if } \zeta = \theta, \\ +\infty & \text{if } \zeta \neq \theta. \end{cases}$$

Proof. First we prove that the measures $Q^\epsilon(dx, d\zeta;\theta)$ form a WULDF. It is easy to see that the function $H_Q(x,\zeta;\theta)$ satisfies properties i) and ii) in Definition 2.4, so we only prove that property iii) holds.

Let $K \subseteq \Theta$ be a compact set, and $F : \mathcal{X} \times \Theta \to \mathbb{R}$ be a continuous and bounded function. Let $L > 0$ be such that $|F(x,\zeta)| \leq L$, and $M(K) = \sup_{\theta \in K} |\inf_x H_P(x;\theta)|$, which is finite by ii) of Definition 2.4. Notice that, by the definition of H_Q, we have

$$\sup_{(x,\zeta)\in\mathcal{X}\times\Theta} \Big[F(x,\zeta) - H_Q(x,\zeta;\theta)\Big] = \sup_{x\in\mathcal{X}} \Big[F(x,\theta) - H_P(x;\theta)\Big].$$

Since H_P is proper (see Lemma 2.8), there exists a compact set $C \subseteq \mathcal{X}$ such that $H_P(x;\theta) \geq 3L + M(K)$ for all $x \in C^c$ and all $\theta \in K$. Moreover, for all $\theta \in K$, we have

$$\sup_{x\in\mathcal{X}} (F(x,\theta) - H_P(x;\theta)) \geq -L - M(K),$$

$$F(x,\theta) - H_P(x;\theta) \leq -2L - M(K) \qquad \text{for all } x \in C^c.$$

Thus

$$\sup_{x\in\mathcal{X}} \Big[F(x,\theta) - H_P(x;\theta)\Big] = \sup_{x\in C} \Big[F(x,\theta) - H_P(x;\theta)\Big]. \tag{3.1}$$

Let $\beta > 2L + 2M(K)$. Since $\{P^\epsilon(dx;\theta)\}$ is exponentially tight, there exists a compact set $C_M \subseteq \mathcal{X}$ such that

$$P^\epsilon(C_M^c;\theta) \leq \exp\{-\epsilon^{-1}\beta\} \tag{3.2}$$

for all $\theta \in K$. Without loss of generality, we may assume that $C \subseteq C_M$ and that for ϵ small enough $P^\epsilon(C_M;\theta) \geq e^{-2\epsilon^{-1}M(K)}$ for all $\theta \in K$ (see (2.3)). We have

$$\begin{aligned}
&\epsilon \log \int \exp\{\epsilon^{-1}F(x,\theta)\} P^\epsilon(dx;\theta) \\
&= \epsilon \log \left(\int_{C_M} \exp\{\epsilon^{-1}F(x,\theta)\} P^\epsilon(dx;\theta) + \int_{C_M^c} \exp\{\epsilon^{-1}F(x,\theta)\} P^\epsilon(dx;\theta) \right) \\
&\leq \epsilon \log \int_{C_M} \exp\{\epsilon^{-1}F(x,\theta)\} P^\epsilon(dx;\theta) + \epsilon \log \left(1 + 2\exp\{\epsilon^{-1}(2L + 2M(K) - \beta)\}\right).
\end{aligned}$$

Now choose an arbitrary $\delta > 0$. Since $2L + 2M(K) - \beta < 0$, there exists ϵ_0 such that, for all $\epsilon \le \epsilon_0$ and for all $\theta \in K$, we have

$$\left|\epsilon \log \int \exp\{\epsilon^{-1}F(x,\theta)\}P^\epsilon(dx;\theta) \;-\; \epsilon \log \int_{C_M} \exp\{\epsilon^{-1}F(x,\theta)\}P^\epsilon(dx;\theta)\right| \le \delta. \tag{3.3}$$

Since $F|_{C_M\times K}$ is uniformly continuous, we have that for each $\theta \in K$ there exists an open neighborhood U_θ of θ such that $|F(x,\theta_1) - F(x,\theta_2)| < \delta$ for all $\theta_1, \theta_2 \in U_\theta$ and all $x \in C_M$. K being compact, there exists $\theta_1, \dots, \theta_n$ such that $K \subseteq \cup_{i=1}^n U_{\theta_i}$.

Let $\theta \in K$; then there exists $\bar{i}$ such that $\theta \in U_{\theta_{\bar{i}}}$. Since $|F(x,\theta) - F(x,\theta_{\bar{i}})| < \delta$ for all $x \in C_M$, we have

$$\left|\epsilon \log \int_{C_M} \exp\{\epsilon^{-1}F(x,\theta)\}P^\epsilon(dx;\theta) - \epsilon \log \int_{C_M} \exp\{\epsilon^{-1}F(x,\theta_{\bar{i}})\}P^\epsilon(dx;\theta)\right| < \delta, \tag{3.4}$$

and

$$\left|\sup_{x\in C_M}[F(x,\theta_{\bar{i}}) - H_P(x;\theta)] - \sup_{x\in C_M}[F(x,\theta) - H_P(x;\theta)]\right| < \delta. \tag{3.5}$$

Moreover, by definition of WULDF, we also have

$$\left|\epsilon \log \int \exp\{\epsilon^{-1}F(x,\theta_{\bar{i}})\}P^\epsilon(dx;\theta) - \sup_{x\in\mathcal{X}}[F(x,\theta_{\bar{i}}) - H_P(x;\theta)]\right| < \lambda_i(\epsilon), \tag{3.6}$$

where $\lim_{\epsilon\to 0}\lambda_i(\epsilon) = 0$.

Now let $\lambda(\epsilon) = \sup_{i=1,\dots,n}\lambda_i(\epsilon)$, and note that $\lim_{\epsilon\to 0}\lambda(\epsilon) = 0$. Using (3.1), (3.3), (3.4), (3.5), and (3.6), for all $\epsilon \le \epsilon_0$, we have

$$\begin{aligned}
&\left|\epsilon \log \int \exp\{\epsilon^{-1}F(x,\zeta)\}Q^\epsilon(dx,d\zeta;\theta) - \sup_{(x,\zeta)\in\mathcal{X}\times\Theta}[F(x,\zeta) - H_Q(x,\zeta;\theta)]\right| \\
&\quad = \left|\epsilon \log \int \exp\{\epsilon^{-1}F(x,\theta)\}P^\epsilon(dx;\theta) - \sup_{x\in\mathcal{X}}[F(x,\theta) - H_P(x;\theta)]\right| \\
&\quad \le \delta + \left|\epsilon \log \int_{C_M} \exp\{\epsilon^{-1}F(x,\theta)\}P^\epsilon(dx;\theta) - \sup_{x\in C_M}[F(x,\theta) - H_P(x;\theta)]\right| \\
&\quad \le \delta + \left|\epsilon \log \int_{C_M} \exp\{\epsilon^{-1}F(x,\theta)\}P^\epsilon(dx;\theta) - \epsilon \log \int_{C_M} \exp\{\epsilon^{-1}F(x,\theta_{\bar{i}})\}P^\epsilon(dx;\theta)\right| \\
&\qquad + \left|\epsilon \log \int \exp\{\epsilon^{-1}F(x,\theta_{\bar{i}})\}P^\epsilon(dx;\theta) - \epsilon \log \int_{C_M} \exp\{\epsilon^{-1}F(x,\theta_{\bar{i}})\}P^\epsilon(dx;\theta)\right| \\
&\qquad + \left|\epsilon \log \int \exp\{\epsilon^{-1}F(x,\theta_{\bar{i}})\}P^\epsilon(dx;\theta) - \sup_{x\in\mathcal{X}}[F(x,\theta_{\bar{i}}) - H_P(x;\theta)]\right| \\
&\qquad + \left|\sup_{x\in C_M}[F(x,\theta_{\bar{i}}) - H_P(x;\theta)] - \sup_{x\in C_M}[F(x,\theta) - H_P(x;\theta)]\right| \\
&\quad \le \delta + \delta + \delta + \lambda(\epsilon) + \delta.
\end{aligned}$$

Note that to get the last inequality we used the fact that equation (3.3) still holds when we replace $F(x,\theta)$ by $F(x,\theta_{\bar{i}})$. Thus

$$\begin{aligned}
\limsup_{\epsilon\to 0}\sup_{\theta\in K}\Big|\ &\epsilon \log {\textstyle\int} \exp\{\epsilon^{-1}F(x,\zeta)\}Q^\epsilon(dx,d\zeta;\theta) \\
&- \sup_{(x,\zeta)\in\mathcal{X}\times\Theta}\big[F(x,\zeta) - H_Q(x,\zeta;\theta)\big]\Big| \le 4\delta.
\end{aligned}$$

Since δ is arbitrary, the previous limit must be zero. Thus we have proved that the measures $Q^\epsilon(dx, d\zeta;\theta)$ form a WULDF. It remains to show that this family is also exponentially tight. Let $K \subseteq \Theta$ be a compact set. Since P^ϵ is exponentially tight, for every $L > 0$ there exists $C \subseteq \mathcal{X}$ compact such that

$$P^\epsilon(C^c;\theta) \leq e^{-\epsilon^{-1}L}$$

for every $\theta \in K$. Let $\tilde{C} = C \times K$. Clearly $\tilde{C} \subseteq \mathcal{X} \times \Theta$ is compact, and

$$Q^\epsilon(\tilde{C}^c;\theta) = P^\epsilon(C^c;\theta).$$

The exponential tightness is therefore easily proved. □

The next corollary restates in a different but equivalent way the result of the previous proposition. We give it explicitly for further reference.

Corollary 3.2. *Let $\{P^\epsilon(dx;y,\theta) : \epsilon > 0, (y,\theta) \in \mathcal{Y} \times \Theta\}$ be an exponentially tight WULDF on $\mathcal{X}$ with rate function $H_P(x;y,\theta)$. Define the measures on $\mathcal{X} \times \mathcal{Y}$, $Q^\epsilon(dx, dz; y, \theta) = P^\epsilon(dx; y, \theta) \otimes \delta_y(dz)$, with δ denoting the Dirac measure. Then $\{Q^\epsilon(dx, dz; y, \theta) : \epsilon > 0, (y, \theta) \in \mathcal{Y} \times \Theta\}$ is an exponentially tight WULDF with rate function*

$$H_Q(x, z; y, \theta) = \begin{cases} H_P(x;y,\theta) & \text{if } z = y, \\ +\infty & \text{if } z \neq y. \end{cases}$$

As a simple application of Proposition 3.1, we give the proof of Proposition 2.5.

Proof of Proposition 2.5. By Proposition 3.1 the following identities are easily obtained:

$$\begin{aligned}\lim_{\epsilon\to 0} \epsilon \log \int e^{\epsilon^{-1}F(x)} P^\epsilon(dx;\theta) &= \lim_{\epsilon\to 0} \epsilon \log \int e^{\epsilon^{-1}F(f(\theta,w))} \mu^\epsilon(dw) \\ &= \lim_{\epsilon\to 0} \epsilon \log \int e^{\epsilon^{-1}F(f(\gamma,w))} \mu^\epsilon \otimes \delta_\theta(dw, d\gamma) \\ &= \sup_{w\in\mathcal{W}} \Big[F(f(\theta,w)) - h(w)\Big] = \sup_{x\in\mathcal{X}} \Big[F(x) - H(x;\theta)\Big],\end{aligned}$$

where the limit is uniform in the compact subsets of Θ. Moreover, property i) of Definition 2.4 is easily shown for $H(x;\theta)$, while property ii) comes automatically from the fact that the P^ϵ are probability measures. □

The next lemma presents an easy technical fact that we will need in the proof of Proposition 3.4.

Lemma 3.3. *Let $F : \mathcal{X} \times \Theta \to \mathbb{R}$ be a continuous and bounded map, and $H : \mathcal{X} \times \Theta \to \mathbb{R}^+$ be the rate function of an exponentially tight WULDF whose elements are probability measures. Moreover, suppose H satisfies the following properties.*

(i) *Let $A = \{(x,\theta) : H(x;\theta) < +\infty\}$. Then for every $(x,\theta) \in A$ and every sequence $\theta_n \to \theta$ there exists a sequence $x_n \to x$ such that $H(x_n;\theta_n) \to H(x;\theta)$.*

(ii) *H is lower semicontinuous as a function of (x,θ).*

Then

$$G(\theta) = \sup_{x\in\mathcal{X}} [F(x,\theta) - H(x;\theta)] \tag{3.7}$$

is a bounded and continuous function.

Proof. Assume that $|F(x,\theta)| \leq L$ for all (x,θ). Since H is the rate function for a family of probability measures, we have (see Remark 2.2) $\inf_x H(x,\theta) = 0$ for all $\theta \in T$. This easily implies $|G(\theta)| \leq L$ for all $t \in \Theta$.

We now show that G is upper semicontinuous. First of all we note that, for any $\theta \in \Theta$, there exists $x \in \mathcal{X}$ such that $G(\theta) = F(x,\theta) - H(x,\theta)$, i.e., the supremum in (3.7) is attained. In fact, that supremum can be equivalently taken for $x \in C$, where $C = \{x : H(x,\theta) \leq 3L\}$. Since C is compact and $F - H$ is upper semicontinuous, then it follows that $F(\cdot,\theta) - H(\cdot,\theta)$ has maximum in C. Now let $\theta_n \to \theta$, and x_n be such that $G(\theta_n) = F(x_n,\theta_n) - H(x_n,\theta_n)$. We have to prove that

$$\limsup G(\theta_n) \leq G(\theta). \tag{3.8}$$

Since the lim sup is the limit along a subsequence, we can assume, without loss of generality, that the sequence $G(\theta_n)$ has a limit. Due to the fact that $\{\theta_n : n \geq 0\} \cup \{\theta\}$ is compact, and the properness of H, it follows that the sequence x_n is relatively compact, so it has a convergent subsequence $x_{n_k} \to x$. Thus

$$\lim G(\theta_n) = \lim G(\theta_{n_k}) = \lim \Big[F(x_{n_k},\theta_{n_k}) - H(x_{n_k},\theta_{n_k})\Big]$$

$$\leq F(x,\theta) - H(x,\theta) \leq G(\theta),$$

where we have used the (joint) upper semicontinuity of $F - H$.

Now we prove lower semicontinuity, i.e., that $\liminf G(\theta_n) \geq G(\theta)$. Let x be such that $G(\theta) = F(x,\theta) - H(x;\theta)$. By property (i), there exists a sequence $x_n \to x$, with $H(x_n;\theta_n) \to H(x,\theta)$. So we get

$$\liminf G(\theta_n) \geq \liminf\, [F(x_n,\theta_n) - H(x_n;\theta_n)] = G(\theta). \qquad \square$$

PROPOSITION 3.4 (composition). *Let* $\{P^\epsilon(dx;y,\theta) : \epsilon > 0, (y,\theta) \in \mathcal{Y} \times \Theta\}$ *and* $\{Q^\epsilon(dy;\theta) : \epsilon > 0, \theta \in \Theta\}$ *be two exponentially tight WULDF in* $\mathcal{X}$ *and* $\mathcal{Y}$, *respectively, with rate functions* $H_P(x;y,\theta)$ *and* $H_Q(y;\theta)$. *Assume that the measures* P^ϵ *are all probability measures. Moreover, assume that the rate function* $H_P(x;y,\theta)$ *satisfies assumptions* (i)–(ii) *of Lemma* 3.3 (*with* $\mathcal{Y} \times \Theta$ *in place of* Θ). *Then* $\{R^\epsilon(dx,dy;\theta) : \epsilon > 0, \theta \in \Theta\}$ *defined by*

$$\int f(x,y) R^\epsilon(dx,dy;\theta) = \int \left[\int f(x,y) P^\epsilon(dx;y,\theta)\right] Q^\epsilon(dy;\theta)$$

is an exponentially tight WULDF with rate function

$$H_R(x,y;\theta) = H_P(x;y,\theta) + H_Q(y;\theta).$$

Proof. First we prove that R^ϵ is a WULDF. Since $H_P(\cdot;y,\theta)$ is positive and has minimum zero, property ii) of Definition 2.4 for H_R is easily derived from the corresponding property for H_Q. We now show that i) of Definition 2.4 holds. Since lower semicontinuity is obvious, we need only to prove that for all $L \in \mathbb{R}$ and for each θ, the set $Z = \{(x,y) \,|\, H_R(x,y;\theta) \leq L\}$ is compact. Notice that if $(x,y) \in Z$ then $y \in W = \{y \,|\, H_Q(y;\theta) \leq L\}$, and W is compact. Since H_P is proper, there exists a compact set $V \subseteq \mathcal{X}$ such that $H_P(x;y,\theta) \geq L + 1$ for all $x \in V^c$ and all $y \in W$ (note that $W \times \{\theta\}$ is compact). Thus we have that

$$Z \subseteq V \times W,$$

and so Z is compact, as desired.

Now we must show that also iii) of Definition 2.4 holds for R^ϵ. Let $F(x,y)$ be a continuous and bounded function, and let $K \subseteq \Theta$ be compact. We need to prove that

$$\lim_{\epsilon \to 0} \sup_{\theta \in K} \left[\epsilon \log \int \exp\{\epsilon^{-1} F(x,y)\} R^\epsilon(dx, dy; \theta) - \sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} [F(x,y) - H_R(x,y;\theta)] \right] = 0. \tag{3.9}$$

Notice that

$$\int \exp\{\epsilon^{-1} F(x,y)\} R^\epsilon(dx, dy; \theta) \;=\; \int \exp\{\epsilon^{-1} G^\epsilon(y,\theta)\} Q^\epsilon(dy, \theta),$$

where

$$G^\epsilon(y,\theta) \;=\; \epsilon \log \int \exp\{\epsilon^{-1} F(x,y)\} P^\epsilon(dx; y, \theta).$$

Clearly, the functions G^ϵ are uniformly bounded. By Corollary 3.2

$$G^\epsilon(y,\theta) \to G(y,\theta) \equiv \sup_{x \in \mathcal{X}} [F(x,y) - H_P(x; y, \theta)]$$

uniformly on the compact subsets of $Y \times \Theta$. Moreover, by Lemma 3.3, G is a bounded continuous function. Thus (3.9) follows as an application of Lemma 2.9 and Proposition 3.1.

It remains to show that the family R^ϵ is exponentially tight. Let $K \subseteq \Theta$ be a compact set, and $M > 0$. Since Q^ϵ is exponentially tight, there exists a compact set $C_1 \subseteq \mathcal{Y}$ such that for all $\theta \in K$ and for all ϵ small enough,

$$Q^\epsilon(C_1^c; \theta) \leq \frac{e^{-\epsilon^{-1} M}}{2}.$$

Moreover, since $P^\epsilon(dx; y, \theta)$ is also exponentially tight, there exists a compact set $C_2 \subseteq \mathcal{X}$ such that

$$P^\epsilon(C_2^c; y, \theta) \leq \frac{e^{-\epsilon^{-1} M}}{2}$$

for all $(y, \theta) \in C_1 \times K$ and for all ϵ small enough. Let $C = C_1 \times C_2$. Then

$$R^\epsilon(C^c; \theta) \leq R^\epsilon(C_2^c \times C_1; \theta) + R^\epsilon(C_2 \times C_1^c; \theta) \leq \sup_{y \in C_1} P^\epsilon(C_2^c; y, \theta) + Q^\epsilon(C_1^c; \theta).$$

Thus, for all ϵ small enough, and for all $\theta \in K$, we get

$$R^\epsilon(C^c; \theta) \leq e^{-\epsilon^{-1} M},$$

which completes the proof. □

Lemma 3.5. *Let $\{P^\epsilon(dx; \theta) : \epsilon > 0, \theta \in \Theta\}$ be a WULDF, with rate function $H(x; \theta)$, and $f : \mathcal{X} \to \mathcal{Y}$ be a continuous function. Then $\{P_f^\epsilon(dy; \theta) : \epsilon > 0, \theta \in \Theta\}$ defined by*

$$P_f^\epsilon(B; \theta) = P^\epsilon(f^{-1}(B); \theta), \quad B \subseteq \mathcal{Y},$$

is again a WULDF with rate function

$$H_f(y;\theta) \;=\; \inf\{H(x;\theta) :\; f(x) = y\}.$$

Moreover if the family $P^\epsilon(dx;\theta)$ is exponentially tight then also the family $P^\epsilon_f(dy;\theta)$ is exponentially tight.

The proof of Lemma 3.5 is easy, and is omitted. From Lemma 3.5 the following Proposition follows.

PROPOSITION 3.6 (contraction). *Let $\{R^\epsilon(dx,dy;\theta) : \epsilon > 0, \theta \in \Theta\}$ be a WULDF on $\mathcal{X} \times \mathcal{Y}$ with rate function $H_R(x,y;\theta)$. Then $\{P^\epsilon(dx;\theta) : \epsilon > 0, \theta \in \Theta\}$, defined by:*

$$P^\epsilon(A;\theta) \;:=\; R^\epsilon(A \times \mathcal{Y};\theta),$$

is a WULDF with rate function

$$H_P(x;\theta) \;=\; \inf_{y\in\mathcal{Y}} H_R(x,y;\theta).$$

Moreover if the family $R^\epsilon(dx,dy;\theta)$ is exponentially tight then also the family $P^\epsilon(dx;\theta)$ is exponentially tight.

PROPOSITION 3.7 (conditioning). *Let $\{P^\epsilon(dx;\theta) : \epsilon > 0, \theta \in \Theta\}$ and $\{Q^\epsilon(dy;x) : \epsilon > 0, x \in \mathcal{X}\}$ be two exponentially tight WULDF's on $\mathcal{X}$ and $\mathcal{Y}$, respectively, with rate functions $H_P(x;\theta)$ and $H_Q(y;x)$. Assume that the measures $Q^\epsilon(dy;x)$ are all probability measures and that both families of kernels are exponentially tight. Moreover assume that the rate function $H_Q(y;x)$ is always finite and continuous, and that the following properties hold.*

1. *The measure $Q^\epsilon(dy;x)$ is of the form*

   $$Q^\epsilon(dy;x) \;=\; q^\epsilon(y;x)\alpha(dy),$$

   where $q^\epsilon(y;x) > 0$ and the measure $\alpha(dy)$ satisfies

   $$\inf_{y\in K} \alpha\left(B(y,\gamma)\right) > 0$$

   for every $K \subset \mathcal{Y}$ compact and $\gamma > 0$, where $B(y,\gamma)$ is the ball centered at y with radius γ.
2. *For any compact sets $K \subseteq \mathcal{Y}$, $C \subseteq \mathcal{X}$ and any $\delta > 0$ there exist $\delta_1 > 0$ and $\epsilon(\delta)$ such that*

   $$|\epsilon \log q^\epsilon(y_1;x) \;-\; \epsilon \log q^\epsilon(y_2;x)| \;<\; \delta$$

   for all y_1, $y_2 \in K$ such that $d(y_1,y_2) < \delta_1$, for all $\epsilon \le \epsilon(\delta)$, and for all $x \in C$.
3. *For any compact sets $K \subseteq \mathcal{Y}$, $C \subseteq \mathcal{X}$ there exists $n_{K,C} > 0$ such that*

   $$\epsilon \log q^\epsilon(y;x) \ge -n_{K,C}$$

   for all $y \in K$, $x \in C$, and $\epsilon > 0$.
4. *For any compact set $K \subseteq \mathcal{Y}$, there exists $N_K > 0$ such that*

   $$\epsilon \log q^\epsilon(y;x) \le N_K$$

   for all $y \in K$, for all $x \in \mathcal{X}$, and for all $\epsilon > 0$.

Then the measures on $\mathcal{X}$

$$R^\epsilon(dx; y, \theta) \;=\; q^\epsilon(y;x) P^\epsilon(dx;\theta)$$

form an exponentially tight WULDF with rate function

$$H_R(x;y,\theta) \;=\; H_Q(y;x) + H_P(x;\theta).$$

Proof. First we prove that the family R^ϵ is exponentially tight. Let $\tilde{K} \subseteq \mathcal{Y} \times \Theta$ be a compact set, and denote by K_1 and K_2 its projection on Y and Θ respectively. By property 4 there exists a constant N_{K_1} such that $q^\epsilon(y;x) \leq e^{\epsilon^{-1} N_{K_1}}$ for all $y \in K_1$, all $x \in \mathcal{X}$, and all $\epsilon > 0$. Given any $M > 0$, since P^ϵ is exponentially tight, there exists a compact set $C \subseteq \mathcal{X}$ such that

$$\sup_{\theta \in K_2} P^\epsilon(C^c;\theta) \;\leq\; e^{-\epsilon^{-1}(M + N_{K_1})}.$$

We have

$$\begin{aligned}\sup_{(y,\theta)\in\tilde{K}} R^\epsilon(C^c;y,\theta) &= \sup_{(y,\theta)\in\tilde{K}} \int_{C^c} q^\epsilon(y;x) P^\epsilon(dx;\theta) \\ &\leq e^{\epsilon^{-1} N_{K_1}} \sup_{\theta\in K_2} P^\epsilon(C^c;\theta) \leq e^{-\epsilon^{-1} M},\end{aligned}$$

which proves exponential tightness. To show that the family $\{R^\epsilon(dx;y,\theta)\}$ is a WULDF with rate function H_R, we define

$$M^P_\theta \;=\; \inf_{x \in X} H_P(x;\theta).$$

Notice that, since the measures $Q^\epsilon(dy;x)$ are all probability measures, it holds that $H_Q(y;x) \geq 0$. Then it is clear that $H_R(x;y,\theta) \geq M^P_\theta$, so the map H_R satisfies property ii) of Definition 2.4. Moreover, since

$$\{\, x \mid H_R(x;y,\theta) \leq L \,\} \subseteq \{\, x \mid H_P(x,\theta) \leq L \,\},$$

property i) of Definition 2.4 holds also.

Now we need to establish property iii) of Definition 2.4. First we prove an intermediate step, which consists of approximating the density $q^\epsilon(y;x)$ with an average on the form

$$\frac{1}{\alpha(B(y,\delta_1))} \int_Y e^{\epsilon^{-1} h(\eta)} q^\epsilon(\eta;x) \alpha(d\eta),$$

$h(\eta)$ being a function suitably concentrated about $\eta = y$.

For any $y \in \mathcal{Y}$, $\delta > 0$, $0 < \tilde{\delta} < \delta$, and $M > 0$, let $g_{y,\delta,\tilde{\delta},M}(\eta)$ be a continuous and bounded function such that

1. $g_{y,\delta,\tilde{\delta},M}(\eta) \leq 0$ for all $\eta \in \mathcal{Y}$, and $g_{y,\delta,\tilde{\delta},M}(\eta) = 0$ for all $\eta \in B(y,\tilde{\delta})$;
2. $g_{y,\delta,\tilde{\delta},M}(\eta) = -M$ if $\eta \notin B(y,\delta)$.

For each given $y \in \mathcal{Y}$, $0 < \tilde{\delta} < \delta$, and $M > 0$, the existence of a function $g_{y,\delta,\tilde{\delta},M}(\cdot)$ satisfying the previous requirements is easily proved. For example one may take

$$g_{y,\delta,\tilde{\delta},M}(\eta) = -\frac{M}{\delta - \tilde{\delta}} \left(\min\{\mathrm{dist}(\eta, B(y,\tilde{\delta})), \delta - \tilde{\delta}\} \right). \qquad \square$$

Claim 1. *For any $K \subseteq \mathcal{Y}$, and $C \subseteq \mathcal{X}$ compact sets, and any $\delta > 0$, $M > 0$, let $\delta_1 > 0$ and $\epsilon(\delta) > 0$ be such that*

$$q^\epsilon(y;x)e^{-\epsilon^{-1}\delta} \le q^\epsilon(\eta;x) \le q^\epsilon(y;x)e^{\epsilon^{-1}\delta} \tag{3.10}$$

for all $x \in C$, $y, \eta \in K$ such that $\eta \in B(y,\delta_1)$ and for all $\epsilon \le \epsilon(\delta)$ (use property 2). Fix any $\tilde{\delta} < \delta_1$, and let $h(\eta) = g_{y,\delta_1,\tilde{\delta},M}(\eta)$. Then there exists a constant $C(\delta_1) > 0$ such that

$$\begin{aligned} &e^{-\epsilon^{-1}\delta}\big(1 + C(\delta_1)^{-1}e^{-\epsilon^{-1}(M-N_K+\delta)}\big)^{-1}\frac{1}{\alpha(B(y,\delta_1))}\int_{\mathcal{Y}} e^{\epsilon^{-1}h(\eta)}q^\epsilon(\eta;x)\alpha(d\eta) \\ &\qquad \le q^\epsilon(y;x) \le e^{\epsilon^{-1}\delta}\frac{1}{\alpha(B(y,\tilde{\delta}))}\int_{\mathcal{Y}} e^{\epsilon^{-1}h(\eta)}q^\epsilon(\eta;x)\alpha(d\eta) \end{aligned} \tag{3.11}$$

for all $y \in K$, for all $x \in C$, $\forall\, \tilde{\delta} < \delta_1$, and $\forall\, \epsilon \le \epsilon(\delta)$, where N_K is defined in Property 4 in the assumptions.

Proof of the claim. Let $C(\delta_1) = \inf_{y\in K}\alpha(B(y,\delta_1))$. It is easy to see that the following inequalities hold:

$$\int_{\mathcal{Y}} e^{\epsilon^{-1}h(\eta)}q^\epsilon(\eta;x)\alpha(d\eta) \ge \int_{B(y,\tilde{\delta})} e^{-\epsilon^{-1}\delta}q^\epsilon(y;x)\alpha(d\eta) = e^{-\epsilon^{-1}\delta}q^\epsilon(y;x)\alpha\left(B(y,\tilde{\delta})\right),$$

$$\begin{aligned} \int_{\mathcal{Y}} e^{\epsilon^{-1}h(\eta)}q^\epsilon(\eta;x)\alpha(d\eta) &\le \int_{B(y,\delta_1)} e^{\epsilon^{-1}\delta}q^\epsilon(y;x)\alpha(d\eta) \\ &\quad + \int_{B(y,\delta_1)^c} e^{-\epsilon^{-1}M}q^\epsilon(\eta;x)\alpha(d\eta) \\ &\le e^{\epsilon^{-1}\delta}q^\epsilon(y;x)\alpha\left(B(y,\delta_1)\right)\left[1 + C(\delta_1)^{-1}e^{-\epsilon^{-1}(M-N_K+\delta)}\right], \end{aligned}$$

from which (3.11) follows easily. So Claim 1 is proved.

Fix any $K \subseteq \mathcal{Y}$, and $C \subseteq \mathcal{X}$ compact sets, and any $\delta > 0$, $M > 0$. Combining equations (3.11) and (3.10) we get that for all $x \in C$, and all $\tilde{y} \in K$ such that $|\tilde{y} - y| < \delta_1$:

$$\begin{aligned} &e^{-\epsilon^{-1}2\delta}\frac{\gamma(\delta_1)^{-1}}{\alpha(B(y,\delta_1))}\int_{\mathcal{Y}} e^{\epsilon^{-1}h(\eta)}q^\epsilon(\eta;x)\alpha(d\eta) \\ &\qquad \le q^\epsilon(\tilde{y};x) \le e^{\epsilon^{-1}2\delta}\frac{1}{\alpha(B(y,\tilde{\delta}))}\int_{\mathcal{Y}} e^{\epsilon^{-1}h(\eta)}q^\epsilon(\eta;x)\alpha(d\eta) \end{aligned} \tag{3.12}$$

where $h(\eta) = g_{y,\delta_1,\tilde{\delta},M}(\eta)$ and $\gamma(\delta_1) = 1 + \frac{1}{C(\delta_1)}e^{-\epsilon^{-1}(M-N_K+\delta)}$. Since $K \subset \cup_{y\in K}B(y,\delta_1)$ and K is compact, there exist $h_1(\eta),\ldots,h_l(\eta)$ all of the type $g_{y_i,\delta_1,\tilde{\delta},M}(\eta)$ for some $y_i \in K$, such that for all $y \in K$ and for all $x \in C$ there exists an index $i \in \{1,\ldots,l\}$ such that

$$\begin{aligned} &e^{-\epsilon^{-1}2\delta}\frac{\gamma(\delta_1)^{-1}}{\alpha(B(y_i,\delta_1))}\int_{\mathcal{Y}} e^{\epsilon^{-1}h_i(\eta)}q^\epsilon(\eta;x)\alpha(d\eta) \le q^\epsilon(\tilde{y};x) \\ &\qquad \le e^{\epsilon^{-1}2\delta}\frac{1}{\alpha(B(y_i,\tilde{\delta}))}\int_{\mathcal{Y}} e^{\epsilon^{-1}h_i(\eta)}q^\epsilon(\eta;x)\alpha(d\eta), \end{aligned} \tag{3.13}$$

where (3.13) holds for all $\tilde{\delta} < \delta_1$, $\epsilon \le \epsilon(\delta)$.

Now we prove that iii) of Definition 2.4 also holds. Fix a compact set $\tilde{K} \subseteq \mathcal{Y} \times \Theta$. Let $K_1 \subseteq \mathcal{Y}$ be its first projection (i.e., $K_1 = \Pi_1(\tilde{K})$) and $K_2 \subseteq \Theta$ be its second projection. Moreover, let N_{K_1} be the positive constant given by property 4. For any continuous and bounded function $F(x)$, we need to show that

$$\lim_{\epsilon \to 0} \sup_{(y,\theta) \in \tilde{K}} \left[\epsilon \log \int_{\mathcal{X}} e^{\epsilon^{-1} F(x)} R^{\epsilon}(dx; y, \theta) - \sup_{x \in \mathcal{X}} \left(F(x) - H_R(x; y, \theta) \right) \right] = 0.$$

We let

- $|F(x)| \leq L_1$,
- $x_\theta \in \mathcal{X}$ be such that $H^P(x_\theta; \theta) = M^P_\theta = \inf_x H_P(x; \theta)$,
- $M^P(K_2)$ be such that $|M^P_\theta| \leq M^P(K_2)$ for all $\theta \in K_2$,
- $C_1 \subseteq \mathcal{X}$ be a compact set such that $x_\theta \in C_1$ for all $\theta \in K_2$ (such compact set exists by properness of the rate function),
- L_2 be such that $|H_Q(y; x)| \leq L_2$ for all $y \in K_1$ and all $x \in C_1$ (notice that this constant exists since H_Q is continuous).

Now consider a compact set $C_2 \subset \mathcal{X}$ and a constant $\Lambda > 0$ such that, for ϵ sufficiently small,

$$P^{\epsilon}(C_2; \theta) \geq e^{-\epsilon^{-1} \Lambda} \tag{3.14}$$

for every $\theta \in K_2$. Note that this can be done by (2.3) and exponential tightness of P^ϵ. Moreover, by property 3, there is a constant n such that

$$-n \leq \epsilon \log q^{\epsilon}(y; x) \tag{3.15}$$

for all $x \in C_2, y \in K_1$.

Fix any positive constants M, T and $\tilde{M}$ such that:

$$\begin{aligned} &M > 2L_1 + L_2 + 2M^P(K_2) + N_{K_1}, \qquad T > 2L_1 + \Lambda + n, \\ &\tilde{M} > 2L_1 + L_2 + M^P(K_2) + M + \Lambda. \end{aligned} \tag{3.16}$$

Since H_P is proper and P^ϵ, R^ϵ are exponentially tight, we have that there is a compact set $C_3 \subseteq \mathcal{X}$, which satisfies the following inequality for ϵ small enough:

$$H_P(x; \theta) > \tilde{M} \quad \text{for all } x \in C_3^c \text{ and all } \theta \in K_2, \tag{3.17}$$

$$P^{\epsilon}(C_3^c; \theta) \leq e^{-\epsilon^{-1} \tilde{M}} \quad \text{for all } \theta \in K_2, \tag{3.18}$$

$$R^{\epsilon}(C_3^c; y, \theta) \leq e^{-\epsilon^{-1} T} \quad \text{for all } (y, \theta) \in \tilde{K}. \tag{3.19}$$

Notice that without loss of generality, we may assume that $C_1, C_2 \subseteq C_3$. Fix any $\delta > 0$. Let $\delta_2 > 0$ be such that

$$|H_Q(y; x) - H_Q(y'; x')| \leq \delta \tag{3.20}$$

for all x, $x' \in C_3$, and y, $y' \in K_1$, such that $\operatorname{dist}(x, x') < \delta_2$ and $\operatorname{dist}(y, y') < \delta_2$.

Now, fix $(y, \theta) \in \tilde{K}$, $x \in C_3$. We have seen that there exist $\delta_1 \leq \delta$, $\epsilon(\delta) > 0$, $y_1, \ldots, y_l \in K_1$ and $i \in \{1, \ldots, l\}$ such that (3.13) holds for all $x \in C_3$, $y \in K_1$, $\tilde{\delta} \leq \delta_1$, and $\epsilon \leq \epsilon(\delta)$.

Upper bound.

$$\begin{aligned}
&\epsilon \log \int e^{\epsilon^{-1}F(x)} R^\epsilon(dx; y, \theta) \\
&\quad = \epsilon \log \int_{C_3} e^{\epsilon^{-1}F(x)} R^\epsilon(dx; y, \theta) + \epsilon \log \left[1 + \frac{\int_{C_3^c} e^{\epsilon^{-1}F(x)} R^\epsilon(dx; y, \theta)}{\int_{C_3} e^{\epsilon^{-1}F(x)} R^\epsilon(dx; y, \theta)} \right] \\
&\quad \leq \epsilon \log \int_{C_3} e^{\epsilon^{-1}F(x)} R^\epsilon(dx; y, \theta) + \epsilon \log \left(1 + e^{\epsilon^{-1}(2L_1 + \Lambda + n - T)} \right),
\end{aligned}$$

where we have used the inequalities

$$\int_{C_3^c} e^{\epsilon^{-1}F(x)} R^\epsilon(dx; y, \theta) \leq e^{\epsilon^{-1}(L_1 - T)},$$

$$\int_{C_3} e^{\epsilon^{-1}F(x)} R^\epsilon(dx; y, \theta) \geq \int_{C_2} e^{\epsilon^{-1}F(x)} R^\epsilon(dx; y, \theta) \geq e^{-\epsilon^{-1}(L_1 + \Lambda + n)}$$

for every $(y, \theta) \in \tilde{K}$. Therefore, using (3.13),

$$\begin{aligned}
&\epsilon \log \int e^{\epsilon^{-1}F(x)} R^\epsilon(dx; y, \theta) \\
&\quad \leq \epsilon \log \left(1 + e^{\epsilon^{-1}(2L_1 + \Lambda + n - T)} \right) + 2\delta \\
&\qquad - \epsilon \log \alpha(B(y_i, \tilde{\delta})) + \epsilon \log \int_{C_3 \times Y} e^{\epsilon^{-1}(F(x) + h_i(\eta))} Q^\epsilon(d\eta; x) P^\epsilon(dx; \theta) \\
&\quad \leq \epsilon \log \left(1 + e^{\epsilon^{-1}(2L_1 + \Lambda + n - T)} \right) + 2\delta - \epsilon \log C(\tilde{\delta}) \\
&\qquad + \epsilon \log \int_{X \times Y} e^{\epsilon^{-1}(F(x) + h_i(\eta))} Q^\epsilon(d\eta; x) P^\epsilon(dx; \theta) \\
&\quad \leq \epsilon \log \left(1 + e^{\epsilon^{-1}(2L_1 + \Lambda + n - T)} \right) + 2\delta - \epsilon \log C(\tilde{\delta}) \\
&\qquad + \sup_{x \in \mathcal{X}, \eta \in \mathcal{Y}} [F(x) + h_i(\eta) - H_Q(\eta; x) - H_P(x; \theta)] + \lambda_i(\epsilon)
\end{aligned}$$

with $\lambda_i(\epsilon) \to 0$ as $\epsilon \to 0$. Note that for the last inequality Proposition 3.4 has been used. Now observe that

$$\begin{aligned}
&\sup_{x \in \mathcal{X}, \eta \in \mathcal{Y}} \Big[F(x) + h_i(\eta) - H_Q(\eta; x) - H_P(x; \theta) \Big] \\
&\qquad \geq F(x_\theta) + h_i(y_i) - H_Q(y_i; x_\theta) - H_P(x_\theta; \theta) \geq -L_1 - L_2 - M^P(K_2).
\end{aligned} \tag{3.21}$$

On the other hand, for $x \notin C_3$

$$F(x) + h_i(\eta) - H_Q(\eta; x) - H_P(x; \theta) \leq L_1 - \tilde{M}, \tag{3.22}$$

and, for $\operatorname{dist}(\eta, y_i) > \delta_1$,

$$F(x) + h_i(\eta) - H_Q(\eta; x) - H_P(x; \theta) \leq L_1 - M + M^P(K_2). \tag{3.23}$$

By (3.16), it follows that the right-hand side (RHS) of (3.22) and (3.23) are smaller than the RHS of (3.21). Therefore

$$\begin{aligned}
&\sup_{x\in\mathcal{X},\eta\in\mathcal{Y}}\Big[F(x)+h_i(\eta)-H_Q(\eta;x)-H_P(x;\theta)\Big]\\
&\quad=\sup_{x\in C_3,\eta\in B(y_i,\delta_1)}\Big[F(x)+h_i(\eta)-H_Q(\eta;x)-H_P(x;\theta)\Big]\\
&\quad\le\sup_{x\in C_3}\Big[F(x)-H_Q(y;x)-H_P(x;\theta)\Big]+\delta\\
&\quad=\sup_{x\in\mathcal{X}}\Big[F(x)-H_Q(y;x)-H_P(x;\theta)\Big]+\delta
\end{aligned}$$

where the last equality comes from an argument analogous to (3.21)–(3.23). Summing up,

$$\begin{aligned}
&\epsilon\log\int e^{\epsilon^{-1}F(x)}R^\epsilon(dx;y,\theta)\\
&\quad\le\epsilon\log\Big(1+e^{\epsilon^{-1}(2L_1+\Lambda+n-T)}\Big)+2\delta-\epsilon\log C(\tilde\delta)\\
&\qquad+\sup_{x\in\mathcal{X}}\Big[F(x)-H_Q(y;x)-H_P(x;\theta)\Big]+\lambda(\epsilon)
\end{aligned}$$

with $\lambda(\epsilon)=\max_{i=1,\dots,l}\lambda_i(\epsilon)$. By using again (3.16), the last inequality implies

$$\begin{aligned}
&\limsup_{\epsilon\to0}\sup_{(y,\theta)\in\tilde K}\epsilon\log\int e^{\epsilon^{-1}F(x)}R^\epsilon(dx;y,\theta)\\
&\qquad-\sup_{x\in\mathcal{X}}\Big[F(x)-H_Q(y;x)-H_P(x;\theta)\Big]\le0.
\end{aligned}\tag{3.24}$$

Lower bound. By (3.13)

$$\begin{aligned}
\epsilon\log\int e^{\epsilon^{-1}F(x)}R^\epsilon(dx;y,\theta)&\ge\epsilon\log\int_{C_3}e^{\epsilon^{-1}F(x)}R^\epsilon(dx;y,\theta)\\
&\ge-2\delta-\epsilon\log\gamma(\delta_1)-\epsilon\log\alpha(B(y_i,\delta_1))\\
&\quad+\epsilon\log\int_{C_3\times\mathcal{Y}}e^{\epsilon^{-1}(F(x)+h_i(\eta))}Q^\epsilon(d\eta;x)P^\epsilon(dx;\theta).
\end{aligned}\tag{3.25}$$

Note that

$$\int_{C_3^c\times\mathcal{Y}}e^{\epsilon^{-1}(F(x)+h_i(\eta))}Q^\epsilon(d\eta;x)P^\epsilon(dx;\theta)\le e^{\epsilon^{-1}(L_1-\tilde M)}$$

and

$$\int_{\mathcal{X}\times\mathcal{Y}}e^{\epsilon^{-1}(F(x)+h_i(\eta))}Q^\epsilon(d\eta;x)P^\epsilon(dx;\theta)\ge e^{-\epsilon^{-1}(L_1+M+\Lambda)},$$

which implies

$$\begin{aligned}
&\epsilon\log\int_{C_3\times\mathcal{Y}} e^{\epsilon^{-1}(F(x)+h_i(\eta))}Q^\epsilon(d\eta;x)P^\epsilon(dx;\theta)\\
&\quad=\epsilon\log\int_{\mathcal{X}\times\mathcal{Y}} e^{\epsilon^{-1}(F(x)+h_i(\eta))}Q^\epsilon(d\eta;x)P^\epsilon(dx;\theta)\\
&\qquad+\epsilon\log\left[1-\frac{\int_{C_3^c\times\mathcal{Y}} e^{\epsilon^{-1}(F(x)+h_i(\eta))}Q^\epsilon(d\eta;x)P^\epsilon(dx;\theta)}{\int_{\mathcal{X}\times\mathcal{Y}} e^{\epsilon^{-1}(F(x)+h_i(\eta))}Q^\epsilon(d\eta;x)P^\epsilon(dx;\theta)}\right]\\
&\quad\ge\epsilon\log\int_{\mathcal{X}\times\mathcal{Y}} e^{\epsilon^{-1}(F(x)+h_i(\eta))}Q^\epsilon(d\eta;x)P^\epsilon(dx;\theta)\\
&\qquad+\epsilon\log\left[1-e^{\epsilon^{-1}(2L_1+M+\Lambda-\tilde{M})}\right].
\end{aligned}$$

Thus

$$\begin{aligned}
&\epsilon\log\int e^{\epsilon^{-1}F(x)}R^\epsilon(dx;y,\theta)\ge-2\delta-\epsilon\log\gamma(\delta_1)-\epsilon\log\alpha(B(y_i,\delta_1))\\
&\quad+\epsilon\log\int_{\mathcal{X}\times\mathcal{Y}} e^{\epsilon^{-1}(F(x)+h_i(\eta))}Q^\epsilon(d\eta;x)P^\epsilon(dx;\theta)+\epsilon\log\left[1-e^{\epsilon^{-1}(2L_1+M+\Lambda-\tilde{M})}\right].
\end{aligned}$$

After having noticed that, by (3.16), $2L_1+M+\Lambda-\tilde{M}<0$, the proof of the lower bound proceeds by repeating the arguments in the proof of the upper bound, yielding

$$\begin{aligned}
&\liminf_{\epsilon\to0}\inf_{(y,\theta)\in\tilde{K}}\epsilon\log\int e^{\epsilon^{-1}F(x)}R^\epsilon(dx;y,\theta)\\
&\qquad-\sup_{x\in\mathcal{X}}\left[F(x)-H_Q(y;x)-H_P(x;\theta)\right]\ge0,
\end{aligned}\tag{3.26}$$

which, together with (3.24), completes the proof. □

Note that, if in Proposition 3.7 we interpret $q^\epsilon(y;x)\alpha(dy)P^\epsilon(dx;\theta)$ as a measure in $\mathcal{X}\times\mathcal{Y}$, the measure R^ϵ has the meaning of *unnormalized* conditional measure. An analogous statement for the normalized version, whose proof follows easily from Proposition 3.7, is given below.

COROLLARY 3.8. *Under the assumptions of Proposition* 3.7, *define*

$$R^\epsilon(dx;y,\theta)=\frac{q^\epsilon(y;x)P^\epsilon(dx;\theta)}{\int_{\mathcal{X}} q^\epsilon(y;x)P^\epsilon(dx;\theta)}.$$

Then $\{R^\epsilon(dx;y,\theta):\epsilon>0,(y,\theta)\in\mathcal{Y}\times\Theta\}$ *is an exponentially tight WULDF with rate function*

$$H_R(x;y,\theta)\;=\;H_Q(y;x)+H_P(x;\theta)-\inf_{x\in\mathcal{X}}\left[H_Q(y;x)+H_P(x;\theta)\right].$$

4. The small parameter limit for partially observed, risk-sensitive control problems.

4.1. The model. Let $(\Omega,\mathcal{F},P)$ be a probability space, $\mathcal{X},\mathcal{Y}$ metric spaces, and $\mathcal{U}$ a compact metric space. Moreover, let $(\mathcal{F}_n)_{n=0}^N$, $(\mathcal{G}_n)_{n=0}^N$ be given filtrations on $(\Omega,\mathcal{F},P)$. We construct a controlled, partially observed stochastic system with state space $\mathcal{X}$, observation space $\mathcal{Y}$, and control space $\mathcal{U}$. The (discrete) time will vary in $\{0,1,\dots,N\}$.

Now, let $\mathcal{M}_1(\mathcal{X})$ $(\mathcal{M}_1(\mathcal{Y}))$ denote the set of probability measures on $\mathcal{X}$ $(\mathcal{Y})$, provided with the weak topology and the corresponding Borel σ-field. Suppose that, for $n = 0, \dots, N-1$ and $\epsilon > 0$, we are given measurable functions (*probability kernels*)

$$\begin{array}{rcl} \mathcal{X} \times \mathcal{U} & \to & \mathcal{M}_1(\mathcal{X}), \\ (x, u) & \to & P_n^\epsilon(\cdot; x, u), \end{array} \tag{4.1}$$

$$\begin{array}{rcl} \mathcal{X} & \to & \mathcal{M}_1(\mathcal{Y}), \\ x & \to & Q_{n+1}^\epsilon(\,\cdot\,; x). \end{array} \tag{4.2}$$

We assume there exists a σ-finite measure α on $\mathcal{Y}$ such that for all $x \in \mathcal{X}$, $Q_{n+1}^\epsilon(\cdot; x)$ has a density with respect to α,

$$Q_{n+1}^\epsilon(dy; x) = q_{n+1}^\epsilon(y; x)\alpha(dy), \tag{4.3}$$

and we assume $q_{n+1}^\epsilon(y; x)$ to be strictly positive everywhere.

For $n = 0, \dots, N-1$ we let $u_n : \mathcal{Y}^{n+1} \to \mathcal{U}$ be a measurable function. The sequence $(u_0, \dots, u_{N-1})$ will be denoted by $\mathbf{u}$, and the set of such sequences (*admissible controls*) will be denoted by $ad(\mathcal{U})$.

For every given $\mathbf{u} \in ad(\mathcal{U})$ we now define $(X_n^{\epsilon,\mathbf{u}})_{n=0}^N$, $(Y_n^{\epsilon,\mathbf{u}})_{n=0}^N$ to be, respectively, $\mathcal{X}$- and $\mathcal{Y}$-valued stochastic processes, defined on $(\Omega, \mathcal{F}, P)$, having the following properties:

i) $X_n^{\epsilon,\mathbf{u}}$ is $\mathcal{F}_n$-measurable, and $Y_n^{\epsilon,\mathbf{u}}$ is $\mathcal{G}_n$-measurable;

ii) for $n = 0, \dots, N-1$,

$$P\{X_{n+1}^{\epsilon,\mathbf{u}} \in \cdot\, | \mathcal{F}_n \vee \mathcal{G}_n\} = P_n^\epsilon(\,\cdot\,; X_n^{\epsilon,\mathbf{u}}, u_n(Y_0^{\epsilon,\mathbf{u}}, \dots, Y_n^{\epsilon,\mathbf{u}})); \tag{4.4}$$

iii) for $n = 1, \dots, N$,

$$P\{Y_n^{\epsilon,\mathbf{u}} \in \cdot\, | \mathcal{F}_n \vee \mathcal{G}_{n-1}\} = Q_n^\epsilon(\,\cdot\,; X_n^{\epsilon,\mathbf{u}}). \tag{4.5}$$

Note that i), ii), and iii) completely determine the law of the processes $X_n^{\epsilon,\mathbf{u}}$ and $Y_n^{\epsilon,\mathbf{u}}$ up to the initial condition $X_0^{\epsilon,\mathbf{u}}, Y_0^{\epsilon,\mathbf{u}}$. For simplicity, we assume $X_0^{\epsilon,\mathbf{u}} = \xi$, $Y_0^{\epsilon,\mathbf{u}} = \eta$, deterministic and $(\epsilon, \mathbf{u})$-independent. It is clear that for given probability kernels as in (4.1) and (4.2), one can construct on a suitable probability space a stochastic process satisfying i), ii), and iii). The dependence on ϵ of the probability kernels in (4.1)(4.2) will be specified later. From now on, the index $(\epsilon, \mathbf{u})$ in $X_n^{(\epsilon,\mathbf{u})}$ and $Y_n^{(\epsilon,\mathbf{u})}$ will be omitted, and we write u_n for $u_n(Y_0^{\epsilon,\mathbf{u}}, \dots, Y_n^{\epsilon,\mathbf{u}})$.

Now we define the cost functional for the optimal control problem. Suppose we are given bounded measurable functions

$$g_n : \mathcal{X} \times \mathcal{U} \to \mathbb{R}, \quad n = 0, \dots, N-1, \quad g_N : \mathcal{X} \to \mathbb{R}. \tag{4.6}$$

For $\mathbf{u} \in ad(\mathcal{U})$ define

$$J^\epsilon(\mathbf{u}) = \epsilon \log E\Big\{ \exp\Big[\epsilon^{-1}\Big(\sum_{n=0}^{N-1} g_n(X_n, u_n) + g_N(X_N)\Big)\Big]\Big\}. \tag{4.7}$$

The optimal control problem associated with J^ϵ consists in computing $J_*^\epsilon = \inf\{J^\epsilon(\mathbf{u}) : \mathbf{u} \in ad(\mathcal{U})\}$ and determining a $\mathbf{u}^* \in ad(\mathcal{U})$ such that $J^\epsilon(\mathbf{u}^*) = J_*^\epsilon$.

4.2. Information vector, information measure, and dynamic programming. In this section the dependence on ϵ of the objects defined in section 2.1 is not relevant. So the index ϵ will be dropped.

It is a standard procedure in stochastic control to analyze optimal control problems with partial observation through a redefinition of the model as a completely observed one. Let $n = 0, \dots, N$. The *information vector* at time n is defined by

$$Z_n = (Y_0, \dots, Y_n, u_0, \dots, u_{n-1}) \in \mathcal{Y}^{n+1} \times \mathcal{U}^n \equiv \mathcal{Z}_n. \tag{4.8}$$

In the following we often identify Z_{n+1} with the triple (Z_n, u_n, Y_{n+1}).

Note that an admissible control at time n can be thought of as a function of Z_n. The stochastic dynamics of (X_n, Y_n) described in i)–iii) induces the following stochastic dynamics for the information vector:

$$P\{Z_{n+1} \in \cdot \,|\mathcal{G}_n\} = \delta_{Y_0} \otimes \cdots \otimes \delta_{Y_n} \otimes P_n^O(dy_{n+1}; Z_n, u_n) \otimes \delta_{u_0} \otimes \cdots \otimes \delta_{u_n}. \tag{4.9}$$

The probability kernels

$$P_n^O : \mathcal{Z}_n \times \mathcal{U} \to \mathcal{M}_1(\mathcal{Y}) \tag{4.10}$$

can be recursively constructed following the procedure below. We also construct an auxiliary sequence of kernels

$$P_n^f : \mathcal{Z}_n \to \mathcal{M}_1(\mathcal{X}) \quad \text{(filtering probabilities)}. \tag{4.11}$$

a) Initialize $P_0^f = \delta_\xi$.

b) Define P_n^O by

$$P_n^O(A; z_n, u_n) = \int \Big(\int Q_{n+1}(A; x_{n+1}) P_n(dx_{n+1}; x_n, u_n) \Big) P_n^f(dx_n; z_n) \tag{4.12}$$

for $A \subset \mathcal{Y}$ measurable.

c) Define P_{n+1}^f by

$$P_{n+1}^f(B; z_{n+1}) = \frac{\int_{\mathcal{X}} \Big(\int_B q_{n+1}(y_{n+1}, x_{n+1}) P_n(dx_{n+1}; x_n, u_n) \Big) P_n^f(dx_n; z_n)}{\int_{\mathcal{X}} \Big(\int_{\mathcal{X}} q_{n+1}(y_{n+1}, x_{n+1}) P_n(dx_{n+1}; x_n, u_n) \Big) P_n^f(dx_n; z_n)} \tag{4.13}$$

for $B \subset \mathcal{X}$ measurable.

Remark 4.1. Equation (4.13) is the well-known *discrete Zakai equation* for the filtering probability. Indeed, P_n^f is a version of the conditional probability of X_n given Z_n, and P_n^O is a version of the conditional probability of Y_{n+1} given Z_n. We assume implicitly that all integrals in (4.13) are finite; this will be guaranteed by later assumptions on the model (Assumption A, section 4.3) where the function $q_n(y_n; \cdot)$ is assumed to be bounded.

Note that P_{n+1}^O has a density with respect to α given by

$$\begin{aligned} \rho_n(y_{n+1}; z_n, u_n) &= \rho_n(z_{n+1}) \\ &= \int \Big(\int q_{n+1}(y_{n+1}, x_{n+1}) P_n(dx_{n+1}; x_n, u_n) \Big) P_n^f(dx_n; z_n). \end{aligned} \tag{4.14}$$

The next step consists of writing the cost function $J(\mathbf{u})$ in terms of the information vector Z_n. The main tool is provided by what we define to be the *information measure.* The information measure at time n is a map

$$P_n^I : \mathcal{Z}_n \to \mathcal{M}(\mathcal{X}), \tag{4.15}$$

where $\mathcal{M}(\mathcal{X})$ is the space of positive finite measures on $\mathcal{X}$. The maps P_n^I are recursively defined as follows:

$$\begin{aligned} &P_0^I(dx_0) = \delta_\xi, \\ &P_{n+1}^I(A; z_{n+1}) \\ &\quad = \frac{\int_{\mathcal{X}} \Big(\int_A e^{g_n(x_n,u_n)} q_{n+1}(y_{n+1}, x_{n+1}) P_n(dx_{n+1}; x_n, u_n) \Big) P_n^I(dx_n; z_n)}{\rho_n(z_{n+1})}. \end{aligned} \tag{4.16}$$

An important property of the information measure is given in the next lemma.

LEMMA 4.1. *The following identity holds for* $n = 0, \dots, N$, *for any* $f : \mathcal{X} \times \mathcal{Z}_n \to \mathbb{R}$ *bounded and measurable and every* $\mathbf{u} \in \mathcal{U}$:

$$E\Big\{ \int f(x_n, Z_n) P_n^I(dx_n; Z_n) \Big\} = E\Big\{ f(X_n, Z_n) \exp\Big[\sum_{k=0}^{n-1} g_k(X_k, u_k) \Big] \Big\}. \tag{4.17}$$

Proof. For $n = 0$ there is nothing to prove. The inductive step is proved as follows:

(4.18)

$$\begin{aligned} &E\left\{ \int f(x_{n+1}, Z_{n+1}) P_{n+1}^I(dx_{n+1}; Z_{n+1}) \right\} \\ &\quad = E\left\{ E\left\{ \int f(x_{n+1}, Z_{n+1}) P_{n+1}^I(dx_{n+1}; Z_{n+1}) \middle| Z_n, u_n \right\} \right\} \\ &\quad = E\left\{ E\left\{ \int f(x_{n+1}, Z_n, u_n, Y_{n+1}) P_{n+1}^I(dx_{n+1}; Z_n, u_n, Y_{n+1}) \middle| Z_n, u_n \right\} \right\} \\ &\quad = E\Big\{ \int \Big[\int f(x_{n+1}, Z_n, u_n, y_{n+1}) P_{n+1}^I(dx_{n+1}; Z_n, u_n, y_{n+1}) \Big] \\ &\qquad\qquad \times \rho_n(y_{n+1}, Z_n, u_n) \alpha(dy_{n+1}) \Big\} \end{aligned}$$

(by (4.16))

$$\begin{aligned} &\quad = E\Big\{ \int \Big[\int \Big(\int f(x_{n+1}, Z_n, u_n, y_{n+1}) q_{n+1}(y_{n+1}; x_{n+1}) P_n(dx_{n+1}; x_n, u_n) \Big) \\ &\qquad\qquad \times \alpha(dy_{n+1}) \Big] e^{g_n(x_n,u_n)} P_n^I(dx_n; Z_n) \Big\} \end{aligned}$$

(by inductive assumption)

$$\begin{aligned} &\quad = E\Big\{ \int \Big[\int f(x_{n+1}, Z_n, u_n, y_{n+1}) q_{n+1}(y_{n+1}; x_{n+1}) P_n(dx_{n+1}; X_n, u_n) \Big] \alpha(dy_{n+1}) \\ &\qquad\qquad \times \exp\Big[\sum_{k=0}^{n} g_k(X_k, u_k) \Big] \Big\} \\ &\quad = E\Big\{ E\Big\{ f(X_{n+1}, Z_{n+1}) \Big| \mathcal{F}_n \vee \mathcal{G}_n \Big\} \exp\Big[\sum_{k=0}^{n} g_k(X_k, u_k) \Big] \Big\} \\ &\quad = E\Big\{ f(X_{n+1}, Z_{n+1}) \exp\Big[\sum_{k=0}^{n} g_k(X_k, u_k) \Big] \Big\}, \end{aligned}$$

where we have used elementary properties of conditional expectation. □

By using the recursive definition (4.16) it is easily checked that the information measures are indeed finite measures. A bound on $P_n^I(X; z_n)$ which is uniform in z_n will be useful later, and is given in the following lemma.

LEMMA 4.2. *For every $n = 0, \dots, N$ and every $z_n \in \mathbb{Z}_n$ we have*

$$|\log P_n^I(X; z_n)| \leq \sum_{k=0}^{n-1} \|g_k\|_\infty. \tag{4.19}$$

Proof. Since the filtering measures are probability measures, it is enough to show that for any $n = 0, \dots, N$ and every positive measurable function f

$$\begin{aligned} \int f(x_n) P_n^f(dx_n; z_n) e^{-\sum_{k=0}^{n-1} \|g_k\|_\infty} &\leq \int f(x_n) P_n^I(dx_n; z_n) \\ \leq \int f(x_n) P_n^f(dx_n; z_n) e^{\sum_{k=0}^{n-1} \|g_k\|_\infty}. \end{aligned} \tag{4.20}$$

The proof of (4.20) comes from an easy induction and is omitted. □

By using Lemma 4.1, we can rewrite the cost functional $J(\mathbf{u})$ as follows:

$$J(\mathbf{u}) = \log E\Big\{ \exp \Big[G_N(Z_N) \Big] \Big\}, \tag{4.21}$$

where

$$G_N(Z_N) = \log \int e^{g_N(x_N)} P_N^I(dx_N; Z_N). \tag{4.22}$$

The partially observed stochastic control problem described in section 4.1 has now been transformed into a totally observed one, with state variable Z_n and cost functional (4.21). For this system the variables Y_n should be thought of as noise variables.

The *value function* associated with (4.22) is defined by

$$V_n(z_n) = \inf_{\mathbf{u} \in ad(\mathcal{U})} \log E\Big\{ \exp \Big[G_N(Z_N) \Big] \Big| Z_n = z_n \Big\}. \tag{4.23}$$

It can be shown (see, e.g., [2]) that V_n satisfies the following recursion:

$$\begin{aligned} V_N(z_N) &= G_N(z_N), \\ V_n(z_n) &= \inf_{u \in \mathcal{U}} \log \int \exp[V_{n+1}(y_{n+1}, z_n, u)] P_n^O(dy_{n+1}; z_n, u). \end{aligned} \tag{4.24}$$

Remark 4.2. By using (4.24) and Lemma 4.2 it is easily seen that the functions V_n are bounded.

Remark 4.3. The stochastic control problem (4.21)–(4.22) is somewhat implicitly stated, since the cost function is given in terms of the solution of the recursion (4.16). Indeed, the cost functional $J(\mathbf{u})$ can be written only in terms of the measures (P_n^f, P_n^I). To see this, consider the stochastic dynamics on $\mathcal{M}_1(X) \times \mathcal{M}(\mathcal{X})$ given by

$$P_{n+1}^f(B) = \frac{\int_{\mathcal{X}} \Big(\int_B q_{n+1}(Y_{n+1}, x_{n+1}) P_n(dx_{n+1}; x_n, u_n) \Big) P_n^f(dx_n)}{\int_{\mathcal{X}} \Big(\int_{\mathcal{X}} q_{n+1}(Y_{n+1}, x_{n+1}) P_n(dx_{n+1}; x_n, u_n) \Big) P_n^f(dx_n)}, \tag{4.25}$$

$$P^I_{n+1}(A) = \frac{\int_{\mathcal{X}} \Big(\int_A e^{g_n(x_n,u_n)} q_{n+1}(Y_{n+1}, x_{n+1}) P_n(dx_{n+1}; x_n, u_n)\Big) P^I_n(dx_n)}{\int_{\mathcal{X}} \Big(\int_{\mathcal{X}} q_{n+1}(Y_{n+1}, x_{n+1}) P_n(dx_{n+1}; x_n, u_n)\Big) P^f_n(dx_n)}, \tag{4.26}$$

or, in short,

$$(P^f_{n+1}, P^I_{n+1}) = F(P^f_n, P^I_n, u_n, Y_{n+1}). \tag{4.27}$$

In (4.27) the Y_n's play the role of disturbances, whose distribution is determined by

$$\begin{aligned} &P(Y_{n+1} \in A | u_0, \dots, u_n, Y_0, \dots, Y_n) \\ &\quad = \int \Big(\int Q_{n+1}(A; x_{n+1}) P_n(dx_{n+1}; x_n, u_n)\Big) P^f_n(dx_n). \end{aligned} \tag{4.28}$$

If we define the cost functional

$$K(\mathbf{u}) = \log E\left\{ e^{g_N(x)} P^I_N(dx)\right\} \tag{4.29}$$

then we have that $K(\mathbf{u}) = J(\mathbf{u})$ for every $\mathbf{u} \in ad(\mathcal{U})$. This shows that, in a very precise sense, the pair (P^f_n, P^I_n) is a sufficient statistic for the risk-sensitive control problem or, in terms more commonly used in control theory, it is an *information state.*

As a consequence, it follows that the value function V_n can be thought of as a function of the information state. In [12, Theorem 3.2], the $\epsilon \to 0$ limit of the value function is studied by looking at the value function as a function of the information state (which is not the same as here, see Remark 4.4 below). In the generality of our model a statement of the type of Theorem 3.2 in [12] does not seem to make sense, and so we prefer to analyze the value function as a function of the information vector. In our construction the information state is only an auxiliary object that allows us to express the cost functional $J(\mathbf{u})$ in terms of the information vector.

Remark 4.4. The information state for risk-sensitive control problems, which is a rather recent achievement in stochastic control theory, was first introduced in [1] and is usually defined (see [12]) through a measure transformation that decouples the observation from the state. The notion of information state in [12] has the advantage, among others, that in the $\epsilon \to 0$ limit it induces a quite natural notion of information state for the limit dynamic game. However, when the partially observed control problem is transformed into a totally observed one by means of the information state in [12] one gets a value function which is, in general, unbounded. In order to use large deviation techniques to control the $\epsilon \to 0$ limit of the value function, some growth bounds are needed, and these bounds come from assumptions on the dynamics of the model. The assumptions that will be given in section 4.3 would not imply any growth bound. Our construction guarantees boundedness of the value function, and appears to be more robust in terms of assumptions on the model.

4.3. Small parameter limit. In this section we investigate the limit of the value function in (4.23) as $\epsilon \to 0$. We first introduce the basic assumptions on the model that are needed to study the small parameter limit.

Assumption A.

1. For $n = 0, \dots, N-1$ the families of probability measures $\{P^\epsilon_n(dx_{n+1}; x_n, u_n) : \epsilon > 0, (x_n, u_n) \in \mathcal{X} \times \mathcal{U}\}$ are WULDF's with rate functions $H^P_n(x_{n+1}; x_n, u_n)$, and they are exponentially tight. In addition, the map $(x_n, u_n) \to P^\epsilon_n(dx_{n+1}; x_n, u_n)$ is weakly continuous.

2. Let $A_n = \{(x, \xi, u) \in \mathcal{X} \times \mathcal{X} \times \mathcal{U} : H_n^P(x, \xi, u) < +\infty\}$. Then for every sequence $(\xi_n, u_n) \to (\xi, u)$ there exists a corresponding sequence $x_n \to x$ such that $H_n^P(x_n, \xi_n, u_n) \to H_n^P(x, \xi, u)$.
3. H_n^P is jointly lower semicontinuous in (x_{n+1}, x_n, u_n).
4. For $n = 1, \ldots, N$ the families of probability measures $\{Q_n^\epsilon(dy_n; x_n) : \epsilon > 0, x_n \in \mathcal{X}\}$ are WULDFs with finite and continuous rate functions $H_n^Q(y_n; x_n)$, and they are exponentially tight.
5. The reference measure α on $\mathcal{Y}$ such that $Q_n^\epsilon(dy_n; x_n) = q_n^\epsilon(y_n; x_n)\alpha(dy_n)$ satisfies
$$\inf_{y \in K} \alpha\left(B(y, \gamma)\right) > 0,$$
for every $K \subset \mathcal{Y}$ compact and $\gamma > 0$, where $B(y, \gamma)$ is the ball centered at y with radius γ. Moreover the density $q_n^\epsilon(y_n; x_n)$ is jointly continuous in $(y_n; x_n)$.
6. For every $K \subset \mathcal{Y}$ compact, $C \subset \mathcal{X}$ compact and every $\delta > 0$ there exist $\delta' > 0$ and $\epsilon' > 0$ such that if $y, y' \in K$ and $d(y, y') < \delta'$ then $|\epsilon \log q_n^\epsilon(y, x) - \epsilon \log q_n^\epsilon(y', x)| < \delta$ for all $x \in C$ and $\epsilon < \epsilon'$.
7. For every $K \subset \mathcal{Y}$, $C \subset \mathcal{X}$ compact the functions $\epsilon \log q_n^\epsilon(y, x)$ are uniformly bounded from above on $K \times \mathcal{X}$ and uniformly bounded from below on $K \times C$ (uniformly means that the bound is independent of ϵ).
8. The functions g_n appearing in the cost functional $J(\mathbf{u})$ are continuous and bounded.

Note that conditions 2–3 and 5–7 correspond to the assumptions of Propositions 3.4 and 3.7 respectively.

A sufficient condition for Assumption A to hold is provided by the following.

Assumption B.

1. Let $\mathcal{W}$ be a metric space. For $n = 0, \ldots, N-1$ let $f_n : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \to \mathcal{X}$ be continuous functions and $\{\mu_n^\epsilon : \epsilon > 0\}$ be an exponentially tight family of probability measures on $\mathcal{W}$ satisfying an LDP with rate function $h_n(w)$. The probability measures $P_n^\epsilon(dx_{n+1}; x_n, u_n)$ are defined by
$$P_n^\epsilon(A; x_n, u_n) = \mu_n^\epsilon\{w : f_n(x_n, u_n, w_n) \in A\}.$$
2. Let $\mathcal{Y} = \mathbb{R}^d$, and, for $n = 1, \ldots, N$, let $\phi_n : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}^d$ be continuous functions. Moreover, let $\{\nu_n^\epsilon : \epsilon > 0\}$, $n = 1, \ldots, N$, be exponentially tight families of probability measures satisfying an LDP with rate function $k_n(v)$, that is finite and continuous. Suppose the following conditions are satisfied:
a) for every fixed $x \in \mathcal{X}$ the map $v \to \phi_n(x, v)$ is a diffeomorphism in $\mathbb{R}^d$. Moreover the inverse map $\phi_n^{-1}(x, y)$ and $D_y\phi_n^{-1}(x, y)$ are continuous on $\mathcal{X} \times \mathbb{R}^d$, where D_y denotes differentiation with respect to y.
b) For every $K \subset \mathbb{R}^d$ compact, the map $\det(D_y\phi_n^{-1})$ is bounded on $\mathcal{X} \times K$.
c) $\nu_n^\epsilon \ll dv$, and $\{\epsilon \log \frac{d\nu_n^\epsilon}{dv} : \epsilon > 0\}$ is a family of functions that, when restricted to any compact subset of $\mathbb{R}^d$, are equicontinuous and uniformly bounded from below, and are uniformly bounded from above on all $\mathbb{R}^d$.
The probability measures $Q_n^\epsilon(dy_n; x_n)$ are defined by
$$Q_n^\epsilon(B; x_n) = \nu_n^\epsilon\{v : \phi_n(x_n, v) \in B\}.$$
3. The functions g_n appearing in the cost functional $J(\mathbf{u})$ are continuous and bounded.

Note that, under Assumption B, the dynamics for X_n, Y_n have the form

$$\begin{aligned} X_{n+1} &= f_n(X_n, u_n, W_n), \\ Y_n &= \phi_n(X_n, V_n), \end{aligned}$$

where $\{W_0, \dots, W_{N-1}, V_1, \dots, V_N\}$ are independent random variables with $W_n \sim \mu_n^\epsilon$ and $V_n \sim \nu_n^\epsilon$.

The following fact will be proved in the Appendix.

PROPOSITION 4.3. *Assumption B implies Assumption A.*

Example 4.5.

1. We give first some examples where Assumption B holds. Assume $\mathcal{W} = \mathbb{R}^m$. Suppose also that, for $n = 0, 1, \dots, N-1$, we are given Borel measurable functions $\tilde{h}_n : \mathbb{R}^m \to \mathbb{R}$, $\tilde{k}_n : \mathbb{R}^d \to \mathbb{R}$ such that
   i) $\tilde{h}_n(w), \tilde{k}_n(v) \to +\infty$ as $\|w\| \to +\infty, \|v\| \to +\infty$;
   ii) $e^{-\tilde{h}_n}$, $e^{-\tilde{k}_n}$ are integrable with respect to the Lebesgue measure;
   iii) $\tilde{h}_n$ is almost everywhere (a.e.) bounded from below, and $\tilde{k}_n$ is continuous.
   Then we can define

   $$\mu_n^\epsilon(dw) = \frac{e^{-\epsilon^{-1}\tilde{h}_n(w)}dw}{\int e^{-\epsilon^{-1}\tilde{h}_n(w)}dw}, \quad \nu_n^\epsilon(dw) = \frac{e^{-\epsilon^{-1}\tilde{k}_n(v)}dv}{\int e^{-\epsilon^{-1}\tilde{k}_n(v)}dv}.$$

   Then $\mu_n^\epsilon, \nu_n^\epsilon$ satisfy the requirements given in Assumption B, with corresponding rate functions $h_n(w) = \tilde{h}_n(w) - \inf \tilde{h}_n$, $k_n(v) = \tilde{k}_n(v) - \inf \tilde{k}_n$. Note that this example includes the Gaussian noise considered in [12].

   To complete the description of the model we can assign any continuous functions $f_n : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \to \mathcal{X}$ for the state dynamic equations, while examples of output functions ϕ_n satisfying Assumption B are provided by functions of type

   $$\phi_n(x, v) = \beta_n(x) + \gamma_n(x)v,$$

   where $\beta_n : \mathcal{X} \to \mathbb{R}^d$, $\gamma_n : \mathcal{X} \to \mathcal{L}(\mathbb{R}^d, \mathbb{R}^d)$ are continuous functions and, for all $v \in \mathbb{R}^d$, the inequality $\|\gamma_n(x)v\|^2 \geq \delta\|v\|^2$ holds for a constant δ independent of $x \in \mathcal{X}$. Note that no boundedness or growth assumptions on β_n are required.
2. Assumption A has the advantage of being more general than Assumption B, and somewhat more directly usable in the proofs. Besides technical convenience, the description of the model in terms of transition probabilities, rather than difference equations with noise, may be more natural in some contexts, e.g., when $\mathcal{X}$ and/or $\mathcal{Y}$ are finite sets. For instance, in the case of $\mathcal{X}$ finite, one may consider transition probabilities of the form

   $$P_n^\epsilon(x_{n+1}; x_n, u_n) = \frac{e^{-\epsilon^{-1}H_n^P(x_{n+1}; x_n, u_n)}}{\sum_{z \in \mathcal{X}} e^{-\epsilon^{-1}H_n^P(z; x_n, u_n)}}. \tag{4.30}$$

   If H_n^P is finite and continuous in u_n, then (4.30) automatically satisfies 1–3 of Assumption A. A similar transition mechanism can be defined for the output, when $\mathcal{Y}$ is finite; if $\mathcal{X}$ and $\mathcal{Y}$ are both finite, conditions 4–7 of Assumption A are trivially satisfied.

   Dynamics of type (4.30) appear naturally in statistical mechanical models of particle systems; in that context ϵ is a temperature parameter, while the control u_n may be seen as an external field perturbing some "free" evolution.

In what follows we will use objects defined in section 4.2. We find convenient to give a list of all identities we are going to use, showing explicitly the dependence on ϵ:

$$P^{f,\epsilon}_{n+1}(B;z_{n+1}) = \frac{\int_{\mathcal{X}}\Big(\int_B q^\epsilon_{n+1}(y_{n+1},x_{n+1})P^\epsilon_n(dx_{n+1};x_n,u_n)\Big)P^{f,\epsilon}_n(dx_n;z_n)}{\int_{\mathcal{X}}\Big(\int_{\mathcal{X}} q^\epsilon_{n+1}(y_{n+1},x_{n+1})P^\epsilon_n(dx_{n+1};x_n,u_n)\Big)P^{f,\epsilon}_n(dx_n;z_n)}, \tag{4.31}$$

$$\rho^\epsilon_n(z_{n+1}) = \int\Big(\int q^\epsilon_{n+1}(y_{n+1},x_{n+1})P^\epsilon_n(dx_{n+1};x_n,u_n)\Big)P^{f,\epsilon}_n(dx_n;z_n), \tag{4.32}$$

$$\begin{aligned} &P^{I,\epsilon}_{n+1}(A;z_{n+1}) \\ &= \frac{\int_{\mathcal{X}}\Big(\int_A e^{\epsilon^{-1}g_n(x_n,u_n)}q^\epsilon_{n+1}(y_{n+1},x_{n+1})P^\epsilon_n(dx_{n+1};x_n,u_n)\Big)P^{I,\epsilon}_n(dx_n;z_n)}{\rho^\epsilon_n(z_{n+1})}, \end{aligned} \tag{4.33}$$

$$J^\epsilon(\mathbf{u}) = \epsilon\log E\Big\{\exp\epsilon^{-1}\Big[G^\epsilon_N(z_N)\Big]\Big\}, \tag{4.34}$$

$$G^\epsilon_N(z_N) = \epsilon\log\int e^{\epsilon^{-1}g_N(x_N)}P^{I,\epsilon}_N(dx_N;z_N), \tag{4.35}$$

$$V^\epsilon_n(z_n) = \inf_{u\in\mathcal{U}}\epsilon\log\int\exp\epsilon^{-1}[V^\epsilon_{n+1}(z_n,u,y_{n+1})]P^{O,\epsilon}_{n+1}(y_{n+1};z_n,u). \tag{4.36}$$

Remark 4.6. By using Assumption B one checks by rather standard arguments that the value functions $V^\epsilon_n(z_n)$ are continuous. By Remark 4.2 we know that they are also bounded, and it is clear that the bound does not depend on ϵ.

We now give the main result of this section.

THEOREM 4.4. *There are functions $V_n : \mathcal{Z}_n \to \mathbb{R}$, for $n = 0,\dots,N$, such that $V^\epsilon_n \to V_n$ uniformly on the compact subsets of $\mathcal{Z}_n$. Moreover the functions V_n satisfy the following recursion.*

$$\begin{aligned} V_N(z_N) &= \sup_x\Big[g_N(x) - H^I_N(x;z_N)\Big], \\ V_n(z_n) &= \inf_{u_n}\sup_{y_{n+1}}\Big[V_{n+1}(z_n,u_n,y_{n+1}) - H^O_n(y_{n+1};z_n,u_n)\Big], \end{aligned} \tag{4.37}$$

where H^I_N, H^O_n are determined by the following recursions:

$$H^I_0(x) = H^f_0(x)\begin{cases} 0 & \text{if } x = \xi, \\ +\infty & \text{otherwise,}\end{cases}$$

$$\begin{aligned} H^O_n(y_{n+1};z_n,u_n) = \inf_{x_{n+1},x_n}\Big[&H^Q_{n+1}(y_{n+1};x_{n+1}) + H^P_n(x_{n+1};x_n,u_n) \\ &+H^f_n(x_n;z_n)\Big], \end{aligned} \tag{4.38}$$

$$\begin{aligned}H^f_{n+1}(x_{n+1};z_{n+1}) = H^Q_{n+1}(y_{n+1};x_{n+1}) \\ + \inf_{x_n}\Big[H^P_n(x_{n+1};x_n,u_n) + H^f_n(x_n;z_n)\Big] \\ - H^O_n(y_{n+1};z_n,u_n),\end{aligned}\tag{4.39}$$

$$\begin{aligned}H^I_{n+1}(x;z_{n+1}) = H^Q_{n+1}(y_{n+1};x) \\ + \inf_{\eta\in\mathcal{X}}\Big[H^P_n(x,\eta,u_n) + H^I_n(\eta,z_n) - g_n(\eta,u_n)\Big] \\ - H^O_n(y_{n+1};z_n,u_n).\end{aligned}\tag{4.40}$$

Remark 4.7. We have observed above that the functions V^ϵ_n are continuous and bounded. It follows by Theorem 4.4 that also the functions V_n are continuous and bounded. Indeed, in any metric space, uniform convergence on compact subsets preserves continuity.

The proof of Theorem 4.4 is based on the following result, whose proof is given at the end of this section.

PROPOSITION 4.5. *The three families* $\{P^{O,\epsilon}_n(dy_{n+1};z_n,u_n)\}$, $n=0,\dots,N-1$, $\{P^{f,\epsilon}_n(dx_n;z_n)\}$, *and* $\{P^{I,\epsilon}_n(dx_n,z_n)\}$, $n=0,\dots,N$, *are WULDF with rate functions* H^O_n, H^f_n, H^I_n, *respectively.*

In the proof of Theorem 4.4 we also use the following technical result, which we prove in the appendix.

LEMMA 4.6. *Let E be a metric space, F be a compact metric space, and $f^\epsilon : E\times F\to\mathbb{R}$, $\epsilon\geq 0$, be a family of continuous functions such that $f^\epsilon\to f^0$ uniformly on the compact subsets of $E\times F$. Define $g^\epsilon : E\to\mathbb{R}$ by*

$$g^\epsilon(x) = \inf_{y\in F} f^\epsilon(x,y).$$

Then $g^\epsilon\to g^0$ uniformly on the compact subsets of E.

Proof of Theorem 4.4. We prove the convergence $V^\epsilon_n\to V_n$ by backward induction on n. For $n=N$ the claim is an immediate consequence of (4.35) and Proposition 4.5. We now prove the inductive step. Define

$$T^\epsilon_n(z_n,u) = \epsilon\log\int\exp\epsilon^{-1}[V^\epsilon_{n+1}(z_n,u,y_{n+1})]P^{O,\epsilon}_{n+1}(y_{n+1};z_n,u),\tag{4.41}$$

so that

$$V^\epsilon_n(z_n) = \inf_{u\in\mathcal{U}} T^\epsilon_n(z_n,u).\tag{4.42}$$

By inductive assumption $V^\epsilon_{n+1}(z_n,u,y_{n+1})\to V_{n+1}(z_{n+1})$ uniformly on the compact subsets of $\mathcal{Z}_{n+1}$. Thus, by using Lemma 2.9 and Proposition 4.5,

$$T^\epsilon_n(z_n,u)\to\sup_{y_{n+1}\in\mathcal{Y}}\Big[V_{n+1}(z_n,u,y_{n+1}) - H^O_n(y_{n+1};z_n,u_n)\Big].\tag{4.43}$$

By (4.42) and Lemma 4.6 the conclusion follows. □

Proof of Proposition 4.5. We prove by induction that H^I_n and H^f_n are the rate functions for $P^{I,\epsilon}_n$ and $P^{f,\epsilon}_n$, respectively. The $n=0$ case is clear, since the singleton $\{\delta_\xi\}$ is a WULDF with rate function $H^I_0 = H^f_0$. The inductive step, in both cases, is a simple application of (4.31), (4.33), and Propositions 3.4 and 3.7. The fact that H^O_n is the rate function for $P^{O,\epsilon}_n$ also follows for (4.12) and Proposition 3.4. □

4.4. Interpretation of the limit value function. In this section we show that the limit value function V_n can be interpreted as the value function for a partially observed dynamic game. Although this would not be necessary, for conceptual simplicity Assumption B will be assumed throughout this section. Thus, the stochastic dynamics for the risk-sensitive control problem are given, in short, by

$$\left\{ \begin{array}{rcl} x_{n+1} & = & f_n(x_n, u_n, w_n), \\ y_n & = & \phi_n(x_n, v_n) \end{array} \right. \tag{4.44}$$

with $w_n \sim \mu_n^\epsilon$ and $v_n \sim \nu_n^\epsilon$. Now, consider the deterministic, zero-sum dynamic game with dynamics given by (4.44) and cost functional

$$J(\mathbf{u}) = \sup_{\mathbf{v},\mathbf{w}} \left[\sum_{n=0}^{N-1} \Big(g_n(x_n, u_n) - h_n(w_n) - k_{n+1}(v_{n+1}) \Big) + g_N(x_N) \right] \tag{4.45}$$

defined for $\mathbf{u} \in ad(\mathcal{U})$. The supremum in (4.45) is over all sequences $\mathbf{w} = (w_0, \dots, w_{N-1}) \in \mathcal{W}^N$, $\mathbf{v} = (v_1, \dots, v_N) \in (\mathbb{R}^d)^N$. Note that, for $\mathbf{u} \in ad(\mathcal{U})$ fixed, the expression in (4.45) within square brackets is a function of $\mathbf{w}$, $\mathbf{v}$.

PROPOSITION 4.7. *The following identity holds for $n = 0, \dots, N-1$:*

$$\begin{aligned} \sup\Bigg[\sum_{l=0}^{n} \Big(g_l(x_l, u_l) - h_l(w_l) - k_{l+1}(v_{l+1}) \Big) &: (4.44) \text{ holds, and} \\ w_0, \dots, w_n, v_1, \dots, v_{n+1} &\textit{ are such that} \\ (y_0, \dots, y_{n+1}, u_0, \dots, u_n) &= z_{n+1}, x_{n+1} = x \Bigg] \\ = -H^I_{n+1}(x; z_{n+1}) &- \sum_{k=0}^{n} H^O_k(y_{k+1}; z_k, u_k). \end{aligned} \tag{4.46}$$

Proof. Under Assumption B, we can rewrite (4.40) as

$$\begin{aligned} H^I_{n+1}(x; z_{n+1}) = \inf_{\eta \in \mathcal{X}} \inf_{w \in \mathcal{W}} \inf_{v \in \mathbb{R}^d} \Big[& k_{n+1}(v) + h_n(w) + H^I_n(\eta, z_n) - g_n(\eta, u_n) \\ & : f_n(\eta, u_n, w) = x, \phi_{n+1}(x, v) = y_{n+1} \Big] - H^O_n(y_{n+1}; z_n, u_n). \end{aligned} \tag{4.47}$$

We prove (4.46) by induction on n. For $n = 0$ the claim follows using (4.47). Otherwise, by using the inductive assumption and (4.47), we get

$$\begin{aligned} & \sup\Bigg[\sum_{l=0}^{n} \Big(g_l(x_l, u_l) - h_l(v_l) - k_{l+1}(v_{l+1}) \Big) : \text{ (4.44) holds, and} \\ & w_0, \dots, w_n, v_1, \dots, v_{n+1} \text{ are such that } (y_0, \dots, y_{n+1}, u_0, \dots, u_n) = z_{n+1}, x_{n+1} = x \Bigg] \\ & = \sup_{\eta \in \mathcal{X}} \sup_{w \in \mathcal{W}} \sup_{v \in \mathbb{R}^d} \Big[g_n(\eta, u_n) - k_{n+1}(v) - h_n(w) - H^I_n(\eta, z_n) : f_n(\eta, u_n, w) = x, \\ & \qquad \phi_{n+1}(x, v) = y_{n+1} \Big] \\ & \qquad - \sum_{k=0}^{n-1} H^O_k(y_{k+1}; z_k, u_k) \\ & = -H^I_{n+1}(x; z_{n+1}) - \sum_{k=0}^{n} H^O_k(y_{k+1}; z_k, u_k). \qquad \square \end{aligned} \tag{4.48}$$

Proposition 4.7 allows us to transform the dynamic game (4.44), (4.45) into a totally observed one, in terms of the information vector

$$z_{n+1} = (z_n, u_n, y_{n+1}), \tag{4.49}$$

$$J(\mathbf{u}) = \sup_{\mathbf{y}} \left[\sum_{n=0}^{N-1} G_n(z_n, u_n, y_{n+1}) + G_N(z_N) \right] \tag{4.50}$$

with

$$G_N(z_N) = \sup_{x \in \mathcal{X}} \Big[g_N(x) - H_N^I(x; z_N) \Big], \tag{4.51}$$

$$G_n(z_n, u_n, y_{n+1}) = -H_n^O(y_{n+1}; z_n, u_n), \tag{4.52}$$

and where the supremum in (4.50) is over all sequences $\mathbf{y} = (y_1, \dots, y_N) \in \mathcal{Y}^N$. We recall that the standard definition of the upper value function $V_n(z_n)$ for the dynamic game (4.49), (4.50) is the infimum of

$$\sup_{y_{n+1}, \cdots, y_N} \sum_{k=n}^{N-1} G_k(z_k, u_k, y_{k+1}) + G_N(z_N)$$

over $\mathbf{u} \in ad(\mathcal{U})$ where the dynamics (4.49) start at time n from z_n. A simple dynamic programming argument yields the following.

Proposition 4.8. *The upper value function V_n for the zero-sum, two-player dynamic game* (4.49), (4.50) *is given by* (4.37).

Remark 4.8. We have seen that the pair $(P_n^{f,\epsilon}, P_n^{I,\epsilon})$ is an information state for the risk-sensitive control problem. The corresponding pair (H_n^f, H_n^I) can be interpreted as an information state for the limit dynamic game. In fact, the following totally observed dynamic game with state variables (H_n^f, H_n^I) is equivalent to (4.49), (4.50):

$$\begin{aligned} H_{n+1}^f(x) &= H_{n+1}^Q(y_{n+1}; x) + \inf_{\eta} \Big[H_n^P(x_{n+1}; \eta, u_n) + H_n^f(\eta) \Big] \\ &\quad - \inf_x \Big\{ H_{n+1}^Q(y_{n+1}; x) + \inf_{\xi} \Big[H_n^P(x_{n+1}; \xi, u_n) + H_n^f(\xi) \Big] \Big\}, \end{aligned} \tag{4.53}$$

$$\begin{aligned} H_{n+1}^I(x) &= H_{n+1}^Q(y_{n+1}; x) + \inf_{\eta \in \mathcal{X}} \Big[H_n^P(x, \eta, u_n) + H_n^I(\eta) - g_n(\eta, u_n) \Big] \\ &\quad - \inf_x \Big\{ H_{n+1}^Q(y_{n+1}; x) + \inf_{\xi} \Big[H_n^P(x_{n+1}; \xi, u_n) + H_n^f(\xi) \Big] \Big\}, \end{aligned} \tag{4.54}$$

$$\begin{aligned} J(\mathbf{u}) = \sup_{\mathbf{y}} \Bigg\{ - \sum_{n=0}^{N-1} \inf_x \Big\{ H_{n+1}^Q(y_{n+1}; x) + \inf_{\xi} \Big[H_n^P(x_{n+1}; \xi, u_n) + H_n^f(\xi) \Big] \Big\} \\ + \sup_{x \in \mathcal{X}} \Big[g_N(x) - H_N^I(x) \Big] \Bigg\}. \end{aligned} \tag{4.55}$$

It should be noticed that there is a simpler notion of information state for the dynamic game (4.44)–(4.45), given by the real valued function $K_n : \mathcal{X} \to \mathbb{R}$, evolving according to the equation

$$K_{n+1}(x) = H^Q_{n+1}(y_{n+1};x) + \inf_{\eta\in\mathcal{X}} \Big[H^P_n(x,\eta,u_n) + K_n(\eta) - g_n(\eta,u_n)\Big]. \tag{4.56}$$

It can be shown that

$$J(\mathbf{u}) = \sup_{\mathbf{y}} \sup_{x\in\mathcal{X}} \Big[g_N(x) - K_N(x)\Big]. \tag{4.57}$$

It can be proved that the function $K_n(x)$ is the rate function of a WULDF that is recursively defined as in (4.33) where the denominator ρ^ϵ_n is dropped. The measure obtained in this way is closely related to the information state in [12]; the use of this measure in place of P^I_n gives rise to an unbounded value function, posing serious difficulty to the small parameter analysis.

Example 4.9. In the case $\mathcal{X} = \mathbb{R}^p$,

$$\mu^\epsilon_n(dw) = \frac{1}{(2\pi\epsilon)^{p/2}} e^{-\frac{1}{2\epsilon}||w||^2} dw, \qquad \text{and} \qquad \nu^\epsilon_n(dv) = \frac{1}{(2\pi\epsilon)^{d/2}} e^{-\frac{1}{2\epsilon}||v||^2} dv,$$

we have $h_n(w) = \frac{1}{2}||w||^2$ and $k_n(w) = \frac{1}{2}||v||^2$, and we recover the model in [12], but with much more general equations for the dynamics.

5. The completely observed case. The risk-sensitive control problems satisfying Assumption A do not include the completely observed case ($Y_n = X_n$). It is clear, however, that the method used in this paper can easily be directly applied to the dynamic programming equation of a completely observed problem.

Consider a probability space $(\Omega,\mathcal{F},P)$ with a filtration $(\mathcal{F}_n)_{n=0}^N$. Define $ad(\mathcal{U})$, the set of admissible controls, to be the set of the $\mathcal{F}_n$-adapted $\mathcal{U}$-valued processes. For $\mathbf{u} \in ad(\mathcal{U})$, we let $X^{\epsilon,\mathbf{u}}_n$ be the $\mathcal{F}_n$-adapted $\mathcal{X}$-valued process such that

$$P\{X^{\epsilon,\mathbf{u}}_{n+1} \in \cdot\,|\mathcal{F}_n\} = P^\epsilon_n(\cdot\,; X^{\epsilon,\mathbf{u}}_n, u_n).$$

The cost functional is as in (4.7), but defined in this new set of admissible controls. Consider the value function

$$V^\epsilon_n(x) \;=\; \inf_{\mathbf{u}\in\, ad(\mathcal{U})} \epsilon \log E\bigg\{\exp\bigg[\epsilon^{-1}\bigg(\sum_{k=n}^{N-1} g_k(X_n,u_n) + g_N(X_N)\bigg)\bigg]\bigg\}.$$

The following result is proved by induction as in Theorem 4.4. Note that the assumptions needed are much weaker than those of a similar result given in [4].

THEOREM 5.1. *Assume that part* 1 *of either Assumption A or B holds. Then there are functions V_n such that $V^\epsilon_n \to V_n$ as $\epsilon \to 0$, uniformly on compact subsets of $\mathcal{X}$.*

Moreover, if part 1 *of Assumption B holds, then V_n is the upper value function for the deterministic dynamic game with dynamic given by*

$$x_{n+1} \;=\; f_n(x_n,u_n,w_n),$$

and with cost functionals given by

$$J(\mathbf{u}) = \sup_{\mathbf{w}} \Bigg[\sum_{n=0}^{N-1} \Big(g_n(x_n,u_n) - h_n(w_n)\Big) + g_N(x_N)\Bigg].$$

6. Appendix.

6.1. Proof of Proposition 4.3. Properties 1 and 4 of Assumption A follow from Proposition 2.5 and Lemma 2.7.

In the rest of the proof we drop the index n everywhere.

Proof of property 2. Suppose $(x, \xi, u) \in A$, i.e., $H^P(x; \xi, u) < \infty$. First note that since the set $\{w : f(\xi, u, w) = x\}$ is closed, then there is $w \in \mathcal{W}$ such that $f(\xi, u, w) = x$ and $h(w) = H^P(x; \xi, u)$.

Suppose now that we have a sequence $(\xi_k, u_k) \to (\xi, u)$. We construct a sequence $x_k \to x$ such that $H^P(x_k; \xi_k, u_k) \to H^P(x; \xi, u)$. Define $x_k = f(\xi_k, u_k, w)$. By continuity of f, we have that $x_k \to x$. Then let w_k be such that $x_k = f(\xi_k, u_k, w_k)$ and $h(w_k) = H^P(x_k; \xi_k, u_k)$. Clearly $h(w_k) \leq h(w)$, and therefore the sequence w_k has a convergent subsequence $w_{n_k} \to w'$. By lower semicontinuity of h we have

$$h(w') \leq \liminf h(w_{n_k}) \leq h(w). \tag{6.1}$$

But, again by continuity of f, we also have $x = f(\xi, x, w')$, and therefore

$$h(w') \geq h(w). \tag{6.2}$$

By (6.1) and (6.2) we have

$$\lim_k h(w_k) = h(w)$$

as desired.

Proof of property 3. Consider a sequence $(x_k, \xi_k, u_k) \to (x, \xi, u)$. We must show that

$$\liminf H^P(x_k; \xi_k, u_k) \geq H^P(x; \xi, u). \tag{6.3}$$

It is enough to prove this statement: Suppose there is a subsequence $(x_{n_k}, \xi_{n_k}, u_{n_k})$ such that

$$\lim_k H^P(x_{n_k}; \xi_{n_k}, u_{n_k}) = l < \infty; \tag{6.4}$$

then $l \geq H^P(x; \xi, u)$.

To prove this, let w_{n_k} be such that $x_{n_k} = f(\xi_{n_k}, u_{n_k}, w_{n_k})$ and $h(w_{n_k}) = H^P(x_{n_k}; \xi_{n_k}, u_{n_k})$. By (6.4), the sequence w_{n_k} has a limit point w. By continuity of f, $f(\xi, u, w) = x$, and therefore $h(w) \geq H^P(x; \xi, u)$. Finally, by lower semicontinuity of h,

$$l = \lim_k h(w_{n_k}) \geq h(w) \geq H^P(x; \xi, u)$$

which completes the proof of property 3.

Proof of properties 5 *and* 6. Letting

$$\rho^\epsilon = \frac{d\nu^\epsilon}{dv}$$

we easily get

$$\epsilon \log q^\epsilon(y; x) = \epsilon \log \rho^\epsilon(\phi^{-1}(x, y)) + \epsilon \log \left| \det\Big(D_y \phi^{-1}(x, y)\Big) \right|. \tag{6.5}$$

Property 5 follows from (6.5), equicontinuity of $\epsilon \log \rho^\epsilon$, continuity of ϕ^{-1}, and the fact that the function $\log|\det(D_y\phi^{-1})|$, being continuous, is bounded on the compact subsets of $\mathcal{X} \times \mathbb{R}^d$. Property 6 follows from (6.5) and boundedness of $\epsilon \log \rho^\epsilon$ and $\det(D_y\phi^{-1})$.

The only things left to prove are the finiteness and continuity of the rate function H^Q. This follows from finiteness and continuity of k and the identity

$$H^Q(y;x) = k(\phi^{-1}(x,y)).$$

6.2. Proof of Lemma 4.6. We first show that, for $\epsilon \geq 0$, g^ϵ is continuous. We omit the index ϵ in this part of the proof. Upper semicontinuity is obvious. To prove lower semicontinuity, observe that, due to the compactness of F, for every $x \in E$ there is a "minimizer" $y \in F$ such that $g(x) = f(x,y)$. So let $x_n \to x$, and y_n be the corresponding sequence of minimizers. We have to show that

$$\liminf g(x_n) \geq g(x).$$

To do so, it is not restrictive to assume that the sequence $g(x_n)$ has limit. Moreover, let y be a limit point of $\{y_n\}$. We have:

$$\lim g(x_n) = \lim f(x_n, y_n) = f(x,y) \geq g(x).$$

Thus the functions g^ϵ, $\epsilon \geq 0$, are continuous.

Now, let $K \subset E$ be compact. Let $\delta > 0$ be arbitrary, and let $\epsilon' > 0$ be such that for every $\epsilon < \epsilon'$

$$\left|f^\epsilon(x,y) - f^0(x,y)\right| < \delta \tag{6.6}$$

for any $x \in K, y \in F$. Given $x \in K$, let y^ϵ be such that $g^\epsilon(x) = f^\epsilon(x, y^\epsilon)$. By (6.6), for $\epsilon < \epsilon'$ and $x \in K$,

$$g^\epsilon(x) = f^\epsilon(x,y^\epsilon) \geq f^0(x,y^\epsilon) - \delta \geq g^0(x) - \delta.$$

To complete the proof we have to show that there exists ϵ'' such that for any $\epsilon < \epsilon''$

$$g^\epsilon(x) \leq g^0(x) + \delta \tag{6.7}$$

for all $x \in K$. Suppose, by contradiction, that there is no such ϵ''. Then there is a sequence $\epsilon_n \to 0$ and a corresponding sequence x_n in K such that

$$g^{\epsilon_n}(x_n) > g^0(x_n) + \delta \tag{6.8}$$

for all n. Denote by x a limit point of $\{x_n\}$ and by y a limit point of the sequence of minimizers $y_n^{\epsilon_n}$. By possibly passing to subsequences, we may assume that $\{x_n\} \to x$ and $y_n^{\epsilon_n} \to y$. Thus

$$\lim_n g^{\epsilon_n}(x_n) = \lim_n f^{\epsilon_n}(x_n, y_n^{\epsilon_n}) = f^0(x,y). \tag{6.9}$$

On the other hand, by continuity of g^0,

$$\lim_n g^0(x_n) = g^0(x). \tag{6.10}$$

Thus, by (6.8), (6.9), and (6.10) we have $f^0(x,y) \geq g^0(x) + \delta$. Therefore, there is a y' with $f^0(x,y) \geq f^0(x,y') + \delta$. This implies

$$\lim_n \left[f^{\epsilon_n}(x_n, y_n^{\epsilon_n}) - f^{\epsilon_n}(x_n, y')\right] \geq \delta,$$

which is impossible since $f^{\epsilon_n}(x_n, y_n^{\epsilon_n}) = \inf_z f^{\epsilon_n}(x_n, z)$.

REFERENCES

[1] A. Bensoussan and J. H. van Schuppen, *Optimal control of partially observable stochastic systems with an exponential-of-integral performance index*, SIAM J. Control Optim., 23 (1985), pp. 599–613.

[2] D. P. Bertsekas, *Dynamic Programming and Stochastic Control*, Academic Press, London, 1976.

[3] C. D. Charalambous, *The role of information state and adjoint in relating nonlinear output feedback risk-sensitive control and dynamic games*, IEEE-Trans. Automatic Control, 42 (1997), pp. 1163–1170.

[4] P. Dai Pra and C. Rudari, *Large deviation limit for discrete-time, totally observed stochastic control problems with multiplicative cost*, Appl. Math. Optim., 35 (1997), pp. 221–235.

[5] P. Dai Pra and C. Rudari, *Risk-sensitive control and dynamic games: The discrete-time case*, Proc. IFIP '95, System Modelling and Optimization, Dolezal et al., eds., Chapman & Hall, Prague, 1996.

[6] P. Dai Pra, L. Meneghini, and W. J. Runggaldier, *Some connections between stochastic control and dynamic games*, Math. Control Signal Systems, 9 (1996), pp. 303–326.

[7] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Jones and Bartlett, Boston, 1993.

[8] W. H. Fleming and W. M. McEneaney, *Risk sensitive control and differential games*, in Stochastic Theory and Adaptive Control, Lecture Notes in Control and Inform. Sci. 184, Springer-Verlag, Berlin, 1992, pp. 185–197.

[9] W. H. Fleming and W. M. McEneaney, *Risk sensitive control on an infinite time horizon*, SIAM J. Control Optim., 33 (1995), pp. 1881–1915.

[10] D. H. Jacobson, *Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games*, IEEE Trans. Automat. Control 18 (1973), pp. 124–131.

[11] M. James, *Asymptotic analysis of nonlinear stochastic risk-sensitive control and differential games*, Math. Control Signal Systems, 5 (1992), pp. 401–417.

[12] M. R. James, J. S. Baras, and R. J. Elliott, *Risk sensitive control and dynamic games for partially observed discrete-time nonlinear systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 780–792.

[13] M. R. James, J. S. Baras, and R. J. Elliott, *Output feedback risk sensitive control and differential games for continuous-time nonlinear systems*, in Proc. 32nd IEEE Conf. on Decision and Control, San Antonio, TX, 1993, pp. 3357–3360.

[14] T. Runolfsson, *The equivalence between infinite-horizon optimal control of stochastic systems with exponential-of-integral performance index and stochastic differential games*, IEEE Trans. Automat. Control, 39 (1994), pp. 1551–1563.

[15] S. R. S. Varadhan, *Large Deviations and Applications*, SIAM, Philadelphia, PA, 1984.

[16] P. Whittle, *A risk-sensitive maximum principle*, Systems and Control Lett. 15 (1990), pp. 183–192.

[17] P. Whittle, *Risk-sensitive linear quadratic Gaussian control*, Adv. in Appl. Probab., 13 (1981), pp. 764–777.

[18] P. Whittle, *A risk-sensitive maximum principle: The case of imperfect state observation*, IEEE Trans. Automat. Control, 36 (1991), pp. 793–801.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 1, pp. 33–53

HIGH-ORDER APPROXIMATIONS AND GENERALIZED NECESSARY CONDITIONS FOR OPTIMALITY[*]

URSZULA LEDZEWICZ[†] AND HEINZ SCHÄTTLER[‡]

Abstract. In this paper we derive generalized necessary conditions for optimality for an optimization problem with equality and inequality constraints in a Banach space. The equality constraints are given in operator form as $Q = \{x \in X : F(x) = 0\}$ where $F : X \to Y$ is an operator between Banach spaces; the inequality constraints are given by smooth functionals or by closed convex sets. Models of this type are common in the optimal control problem. The paper addresses the case when the Fréchet-derivative $F'(x_*)$ is not onto and hence the classical Lyusternik theorem does not apply to describe the tangent space to Q. In this case the classical Euler–Lagrange type necessary conditions are trivially satisfied, generating abnormal cases. A high-order generalization of the Lyusternik theorem derived earlier [U. Ledzewicz and H. Schättler, *Nonlinear Anal.,* 34 (1998), pp. 793–815] is used to calculate high-order tangent cones to the equality constraint at points $x_* \in Q$ where $F'(x_*)$ is not onto. Combining these with high-order approximating cones related to the other constraints of the problem (feasible cones respectively cones of decrease) a high-order generalization of the Dubovitskii–Milyutin theorem is given and then applied to derive *generalized* necessary conditions for optimality. These conditions reduce to classical conditions for normal cases, but they give new and nontrivial conditions for abnormal cases.

Key words. Lyusternik theorem, high-order necessary conditions, high-order tangent sets, high-order necessary conditions for optimality, abnormal processes

AMS subject classifications. Primary, 49K27, 46N10; Secondary, 41A10, 47N10

PII. S0363012997317748

1. Introduction.

We consider the problem of minimizing a functional $I : X \to \mathbb{R}$ in a Banach space X under both equality and inequality constraints. The inequality constraints are of two types, described by smooth functionals $f : X \to \mathbb{R}$ as $P = \{x \in X : f(x) \leq 0\}$ or described by closed convex sets C. The equality constraints are given in operator form as $Q = \{x \in X : F(x) = 0\}$ where $F : X \to Y$ is an operator between Banach spaces. Lagrange multiplier type necessary conditions for optimality at the point x_* in the form of an Euler–Lagrange equation,

$$\lambda_0 I'(x_*) + \sum_{j=1}^{m} \lambda_j f_j'(x_*) + F'^*(x_*)y^* = 0, \tag{1.1}$$

while not all of the multipliers $\lambda_0, \dots, \lambda_m, y^*$ are zero, can be derived (see, for instance, [8, 11]) from approximations to the equality constraints (tangent sets, respectively, tangent cones), the inequality constraints (feasible sets/cones) and the directions of decrease for the functional to be minimized (sets/cones of decrease). However this necessary condition can be satisfied in a trivial way for any point $x_* \in Q$, where the equality constraints are not regular in the sense that $F'(x_*)$ is not onto. Assuming that $\operatorname{Im} F'(x_*)$ is closed, then it is possible simply to choose a nontrivial multiplier

[*] Received by the editors March 7, 1997; accepted for publication (in revised form) October 10, 1997; published electronically September 25, 1998.
http://www.siam.org/journals/sicon/37-1/31774.html

[†] Department of Mathematics and Statistics, Southern Illinois University at Edwardsville, Edwardsville, IL 62026-1653 (uledzew@siue.edu). The research of this author was supported in part by NSF grant DMS-9622967, SIUE Research Scholar Award, and 1996 Summer Fellowship.

[‡] Department of Systems Science and Mathematics, Washington University, St. Louis, MI 63130-4899 (hms@cec.wustl.edu). The research of this author was supported in part by NSF grant DMS-9503356.

which annihilates $\operatorname{Im} F'(x_*)$ and set all other multipliers to zero, and this choice will satisfy the Euler–Lagrange equation. But clearly this describes only the equality constraint degeneration without any relation to optimality.

One method to overcome the difficulty of abnormality known as "weakening equality constraints" was introduced by Milyutin in [20] and was investigated further by Dmitruk in [7]. In this method, which applies to extremum problems where the degenerate part of the operator F has a finite-dimensional image, a penalty-type approach is pursued where the degenerate equality constraints are replaced by inequality constraints. Here we pursue a different approach based on a p-order generalization of the Lyusternik theorem derived earlier [16] which determines the precise structure of polynomial approximations to Q at x_* when the surjectivity condition on $F'(x_*)$ is not satisfied but when instead a certain operator G_p which takes into account all derivatives up to and including order p is onto. The order p can be arbitrary, and it will be chosen precisely as the minimum number for which the operator G_p becomes onto. If G_p is onto, then the precise structure of q-order polynomial approximations to Q at x_* for any $q \geq p$ can be determined. Thus our result can be used in connection with appropriate approximations to inequality constraints and directions of decrease for the functional I to derive generalized necessary conditions for optimality based on expansions of increasing orders q. In this paper we describe this framework and the relevant structures for $q = p$. Extensions to $q \geq p$ are rather straightforward and will not be pursued here.

First results of this type have been obtained in the work of Avakov [3, 4, 5] and in our own work [13, 14] for the case of $p = 2$. Some of these conditions have been analyzed further by Arutyunov [1, 2]. The paper by Izmailov [12] also analyzes the case $p = 2$, but for inequality constraints. In our own previous work we have embedded Avakov's results into a second order Dubovitskii–Milyutin theory within the framework of second-order approximating cones [14]. This was achieved by introducing a reparametrization variable. However, this approach does not directly generalize to an arbitrary order p. In this paper we present a different approach which for any order p generates approximating cones in the extended state-space $X \times \mathbb{R}$ and is equivalent to the reparametrization for $p = 2$. Based on these p-order approximating cones we then formulate a p-order version of the Dubovitskii–Milyutin theorem, which we use to derive necessary conditions for optimality. These conditions for optimality are nontrivial also for nonregular constraints or abnormal processes due to the use of the p-order generalization of the Lyusternik theorem proved in [16]. Although derived by means of general p-order approximations, the conditions generalize the classical Euler–Lagrange equation and provide a nonnegativity condition to distinguish between minimizing and maximizing extremals. We therefore call them *generalized* necessary conditions for optimality rather than high-order conditions. They reduce to classical conditions for normal cases, but they are generalized in the sense that they give new and nontrivial conditions for abnormal cases.

The paper is organized as follows: In section 2 we define p-order approximating sets and cones. The definitions of p-order approximations are by standard polynomial expansions and are similar to the variational sets of order p defined in [9] or to its nonsmooth extensions considered in [21]. Our idea is to embed them into a conical structure which has the advantage that classical and well-known results can be used in the calculation of dual and polar cones thus simplifying the analysis. In this section we also give a p-order version of the Dubovitskii–Milyutin theorem [8] which we use to derive the necessary conditions for optimality. In section 3 we describe p-order tangent cones for nonregular equality constraints while we analyze p-order sets and cones of decrease in section 4. In section 5 we include a brief derivation of the generalized necessary conditions for optimality for a minimization problem in a Banach space with smooth operator equality constraints. For this problem there is an equivalence between the occurrence of abnormal points and points at which the equality constraint is nonregular. Also this derivation does not use the Dubovitskii–Milyutin framework, but is based on the geometry of the approximations and thus gives insight into the geometric meaning of the necessary conditions. In section 6 we then consider p-order feasible sets and cones to inequality constraints. We include the cases of inequality constraints described both by smooth functionals as well as inequality constraints given by closed convex sets to create a model which can be applied to the optimal control problem.

Then in section 7 we implement all these results on p-order approximating cones into the Dubovitskii–Milyutin framework and formulate generalized necessary conditions for optimality for our problem. Our results differ from existing first- and second-order necessary conditions for optimality like, for instance, those in [6, 9, 10, 19, 21, 22, 23] by giving conditions which are nontrivial for both regular and nonregular cases. Probably the closest results to ours among the long list of publications which derive necessary conditions for extremum problems assuming regularity of the equality constraints are the maximum principles of pth order by Hoffmann and Kornstaedt in [9, Theorems 5.1 and 5.2]. These results are also derived using p-order approximations, but they become trivial in the case of nonregular constraints. Our results developed here give a general (i.e., p-order) version of our results in [14] which were based on second-order approximations. Like this earlier version, the new results directly apply to the optimal control problem [15], but this formulation will be pursued elsewhere [17].

2. A high-order formulation of the Dubovitskii–Milyutin theorem.

In this paper we derive generalized necessary conditions for optimality based on general p-order approximations for smooth extremum problems in Banach spaces with operator equality constraints and inequality constraints given both by functionals and closed convex sets. We recall that a constraint is called an inequality constraint if the set of admissible points has nonempty interior, while it is called an equality constraint otherwise. In this section we formulate a version of the Dubovitskii–Milyutin theorem which gives general abstract necessary conditions for optimality based on high-order approximations and corresponding high-order approximating cones to the constraints and to the objective. The motivation is to consider p-order polynomial approximations along a given $(p-1)$-order polynomial approximating curve for which the necessary conditions were still inconclusive or even trivially satisfied as for abnormal points. The order p itself can be arbitrary.

Let X and Y be Banach spaces, let $I : X \to \mathbb{R}$ be a functional, $F : X \to Y$ an operator, $f_j : X \to \mathbb{R}$, $j = 1, \dots, m$, functionals and let $C \subset X$ be a closed convex set with nonempty interior. We assume that I, the functionals f_j, and the operator F are sufficiently often continuously Fréchet-differentiable and consider the following problem.

Problem (M).

Minimize I over all $x \in X$ which satisfy

· the inequality constraints $x \in P_j = \{x \in X : \ f_j(x) \leq 0\}$ for $j = 1, \dots, m$,

· the equality constraint $x \in Q = \{x \in X : \ F(x) = 0\}$, and

· the convex inequality constraint $x \in C$.

The set $A = \cap_{j=1}^{m} P_j \cap Q \cap C$ is called the admissible domain. We investigate the structure of high-order polynomial approximations to A. We denote sequences $(h_1, \dots, h_k) \in X^k$ by H_k with the subscript giving the length of the sequence. The following definition is standard (see, for instance, [9]).

Definition 2.1. *Let $H_{p-1} = (h_1, \dots, h_{p-1}) \in X^{p-1}$ and set $x(\varepsilon) \doteq x_* + \sum_{i=1}^{p-1} \varepsilon^i h_i$. We call H_{p-1} a $(p-1)$-order approximating sequence to a set $S \subseteq X$ at $x_* \in Clos\, S$; respectively, we call $x : \varepsilon \longmapsto x(\varepsilon)$ a $(p-1)$-order approximating curve if there exist an $\varepsilon_0 > 0$ and a function r defined on $[0, \varepsilon_0]$ with values in X, $r : [0, \varepsilon_0] \to X$, with the property that for $\varepsilon > 0$*

$$x(\varepsilon) + r(\varepsilon) = x_* + \sum_{i=1}^{p-1} \varepsilon^i h_i + r(\varepsilon) \in S \tag{2.1}$$

and

$$\lim_{\varepsilon \to 0} \frac{||r(\varepsilon)||}{\varepsilon^{p-1}} = 0. \tag{2.2}$$

Using the standard Landau notation we say that the function r is of order $o(\varepsilon^{p-1})$ as $\varepsilon \to 0$. Also, we call a $(p-1)$-order approximating sequence/curve $(p-1)$-*order feasible* if S is an inequality constraint (respectively, $(p-1)$-*order tangent* if S is an equality constraint).

Let $x_* \in F$ and assume as given a $(p-1)$-order approximating sequence $H_{p-1} = (h_1, \dots, h_{p-1}) \in X^{p-1}$ with corresponding $(p-1)$-order approximation $x(\varepsilon) \doteq x_* + \sum_{i=1}^{p-1} \varepsilon^i h_i$. It is implicitly assumed that x_* has not been ruled out for optimality. (For instance, since the functional I was decreasing along the corresponding $(p-1)$-order approximating curve or simply by using other approximating sequences.) Then we would like to extend the existing $(p-1)$-order approximations to p-order approximations and derive the corresponding necessary conditions for optimality. In this section we give the general definitions and formulate these ideas precisely. In the following sections we will describe how these extensions are done. The following definitions are direct generalizations of standard existing definitions as can be found in [8], for instance.

Definition 2.2. *We call v_0 a p-order vector of decrease for a functional $I : X \to \mathbb{R}$ at $x_* \in X$ in the direction of the sequence $H_{p-1} = (h_1, \dots, h_{p-1}) \in X^{p-1}$ if there exist a neighborhood V of v_0 and a number $\alpha < 0$ so that for all $v \in V$ we have*

$$I\left(x_* + \sum_{i=1}^{p-1} \varepsilon^i h_i + \varepsilon^p v\right) = I(x(\varepsilon) + \varepsilon^p v) \leq I(x_*) + \alpha \varepsilon^p. \tag{2.3}$$

The collection of all p-order vectors of decrease for I at x_* in the direction of the sequence H_{p-1} will be called the p-order set of decrease to I at x_* in the direction of the sequence H_{p-1} and will be denoted by $DS^{(p)}(I; x_*, H_{p-1})$.

Definition 2.3. *We call v_0 a p-order feasible vector for an inequality constraint P at $x_* \in X$ in direction of H_{p-1} if there exist an $\varepsilon_0 > 0$ and a neighborhood V of v_0 so that for all $0 < \varepsilon \leq \varepsilon_0$*

$$x_* + \sum_{i=1}^{p-1} \varepsilon^i h_i + \varepsilon^p V = x(\varepsilon) + \varepsilon^p V \subset P. \tag{2.4}$$

The collection of all p-order feasible vectors v_0 for P at x_* in the direction of the sequence H_{p-1} will be called the p-order feasible set to P at x_* in the direction of the sequence H_{p-1} and will be denoted by $FS^{(p)}(P; x_*, H_{p-1})$.

Note that by definition the p-order set of decrease to I and the p-order feasible set to P, both at x_* in direction of the sequence H_{p-1}, are open.

DEFINITION 2.4. *We call h_p a p-order tangent vector to an equality constraint Q at x_* in the direction of the sequence H_{p-1} if $H_p = (h_1, \dots, h_{p-1}, h_p) \in X^p$ is a p-order approximating sequence to the set Q at $x_* \in Q$. The collection of all p-order tangent vectors to Q at x_* in the direction of the sequence H_{p-1} will be called the p-order tangent set to Q at x_* in the direction of the sequence H_{p-1} and will be denoted by $TS^{(p)}(Q; x_*, H_{p-1})$.*

Rather than working with approximating sets as is done for instance by Ben-Tal and Zowe in [6], we prefer to embed these approximating sets into cones in the extended state-space $X \times \mathbb{R}$. This has the advantage that many classical results like the Minkowski–Farkas lemma or the annihilator lemma can be directly applied in calculating dual cones (see also [14]). Let us generally refer to p-order sets of decrease, feasible sets and tangent sets as p-order approximating sets and denote them by $AS^{(p)}(Z; x_*, H_{p-1})$. Then we define the corresponding approximating cones as follows.

DEFINITION 2.5. *Given a p-order approximating set $AS^{(p)}(Z; x_*, H_{p-1})$ to a set $Z \subset X$ at x_* in direction of the sequence H_{p-1}, the p-order approximating cone to Z at x_* in direction of H_{p-1}, $AC^{(p)}(Z; x_*, H_{p-1})$, is the cone in $X \times \mathbb{R}$ generated by the vectors $(v, 1) \in AS^{(p)}(Z; x_*, H_{p-1}) \times \mathbb{R}$, i.e.,*

$$AC^{(p)}(Z; x_*, H_{p-1}) = \bigcup_{\gamma > 0} \gamma \left(AS^{(p)}(Z; x_*, H_{p-1}) \times \{1\} \right). \tag{2.5}$$

Thus we talk of the p-order cone of decrease for the functional I, p-order feasible cones for inequality constraints and p-order tangent cones for equality constraints, all at x_* in direction of the sequence H_{p-1}.

DEFINITION 2.6 (see [8]). *Let $C \subseteq Z$ be a cone in a Banach space Z with apex at 0. The dual (or polar) cone to C consists of all continuous linear functionals $\lambda \in Z^*$ which are nonnegative on C, i.e.,*

$$C^* = \{\lambda \in Z^* : \ \langle \lambda, v \rangle \geq 0 \text{ for all } v \in C\}. \tag{2.6}$$

We are now ready to state and prove a general p-order formulation of the classical Dubovitskii–Milyutin theorem [8, Lemma 5.11].

THEOREM 2.7. *Suppose the functional I attains a local minimum for problem (M) at $x_* \in A$. Let $H_{p-1} = (h_1, \dots, h_{p-1}) \in X^{p-1}$ be a $(p-1)$-order approximating sequence such that the p-order cone of decrease for the functional I, the p-order feasible cones for the inequality constraints P_j, $j = 1, \dots, m$, and C, and the p-order tangent cone to the equality constraint Q, all at x_* in direction of the sequence H_{p-1}, are nonempty and convex. Then there exist continuous linear functionals*

$$\Psi_0 = (\lambda_0, \mu_0) \in \left(DC^{(p)}(I; x_*, H_{p-1}) \right)^*, \tag{2.7}$$

$$\Psi_j = (\lambda_j, \mu_j) \in \left(FC^{(p)}(f_j; x_*, H_{p-1}) \right)^*, \qquad j = 1, \dots, m, \tag{2.8}$$

$$\Omega = (\lambda_{m+1}, \mu_{m+1}) \in \left(FC^{(p)}(C; x_*, H_{p-1}) \right)^* \tag{2.9}$$

and

$$\Phi = (\lambda_{m+2}, \mu_{m+2}) \in \left(TC^{(p)}(Q; x_*, H_{p-1})\right)^*, \tag{2.10}$$

all depending on H_{p-1}, such that the generalized Euler–Lagrange equation

$$\sum_{j=0}^{m+2} \lambda_j \equiv 0 \tag{2.11}$$

and the condition

$$\sum_{j=0}^{m+2} \mu_j \equiv 0 \tag{2.12}$$

hold. Furthermore, not all the λ_j, $j = 0, 1, \dots, m+2$, vanish identically.

Proof. By assumption all p-order approximating cones at x_* in direction of the sequence H_{p-1} are nonempty and convex. Since x_* is optimal, the intersection of these cones must be empty. For, if (w, γ) lies in this intersection, then setting $v = \frac{w}{\gamma}$ we have in particular that $H = (h_1, \dots, h_{p-1,}v)$ is a p-order approximating sequence to the equality constraint Q at $x_* \in Q$. Hence there exists an $\varepsilon_0 > 0$ and a function r defined on $[0, \varepsilon_0]$ with values in X, $r : [0, \varepsilon_0] \to X$, which is of order $o(\varepsilon^p)$ as $\varepsilon \to 0$, with the property that $\gamma(\varepsilon) \doteq x_* + \sum_{i=1}^{p-1} \varepsilon^i h_i + \varepsilon^p v + r(\varepsilon) \in Q$. But this curve is admissible for problem (P) since it also satisfies the inequality constraints and the functional I decreases along the curve $\gamma(\varepsilon)$. This contradicts the optimality of x_*.

Since by definition the p-order cones of decrease for I, $DC^{(p)}(I; x_*, H_{p-1})$, the p-order feasible cones $FC^{(p)}(P_j; x_*, H_{p-1})$ for $j = 1, \dots, p$, and the p-order feasible cone $FC^{(p)}(C; x_*, H_{p-1})$ are open, it therefore follows from the classical Dubovitskii–Milyutin lemma [8, Lemma 5.11] that there exist linear functionals in the respective dual cones $\Psi_0 \in \left(DC^{(p)}(I; x_*, H_{p-1})\right)^*$, $\Psi_j \in \left(FC^{(p)}(f_j; x_*, H_{p-1})\right)^*$, $j = 1, \dots, m$, $\Omega \in \left(FS^{(p)}(C; x_*, H_{p-1})\right)^*$ and $\Phi \in \left(TC^{(p)}(Q; x_*, H_{p-1})\right)^*$, which are not all identically zero, so that

$$\Psi_0 + \sum_{j=1}^{m} \Psi_j + \Omega + \Phi \equiv 0. \tag{2.13}$$

Writing $\Psi_j = (\lambda_j, \mu_j)$, $j = 0, 1, \dots, m$, $\Omega = (\lambda_{m+1}, \mu_{m+1})$, and $\Phi = (\lambda_{m+2}, \mu_{m+2})$, (2.11) and (2.12) follow. Furthermore, the nontriviality of the multipliers Ψ_j, $j = 0, \dots, m$, Ω, and Φ is actually equivalent to the nontriviality of the multipliers λ_j, $j = 0, 1, \dots, m+2$. For, if λ_j vanishes identically, then $\mu_j \geq 0$ since (λ_j, μ_j) lies in the dual to a cone in $X \times \mathbb{R}_+$. Thus, if all the λ_j, $j = 0, \dots, m+2$, vanish identically, then $\mu_j \geq 0$ for all $j = 0, \dots, m+2$. But by (2.12) also $\sum_{j=0}^{m+2} \mu_j \equiv 0$, and thus all μ_j are zero as well contradicting the nontriviality of the multipliers Ψ_j, $j = 0, \dots, m$, Ω and Φ. □

This theorem provides the general mechanism to derive generalized necessary conditions for optimality. In the next sections we derive the structure of general p-order approximating cones and their duals, starting in section 3 with the most difficult construction, p-order tangent cones.

3. High-order approximations for equality constraints.

3.1. A high-order generalization of the Lyusternik theorem. We briefly recall the p-order Lyusternik theorem [16] and the required notation: X and Y are Banach spaces, $F : X \to Y$ is an operator, and $Q = \{x \in X : F(x) = F(x_*)\}$. Assuming that $F : X \to Y$ is sufficiently often continuously Fréchet differentiable in a neighborhood of x_*, we consider the Taylor expansion of F along a curve $\gamma(\varepsilon) = x_* + \sum_{i=1}^m \varepsilon^i h_i$. We have

$$F(x_* + \sum_{i=1}^{m} \varepsilon^i h_i) = F\,(x_*) + \sum_{i=1}^{m} \varepsilon^i \nabla^i F(x_*)(h_1, \dots, h_i) + \widetilde{r}(\varepsilon), \tag{3.1}$$

where

$$\nabla^i F(x_*)(h_1, \dots, h_i) \doteq \sum_{r=1}^{i} \frac{1}{r!} \left(\sum_{j_1 + \ldots + j_r = i} F^{(r)}(x_*)(h_{j_1}, \dots, h_{j_r}) \right) \tag{3.2}$$

and $\widetilde{r}(\varepsilon)$ is a function of order $o(\varepsilon^m)$ as $\varepsilon \to 0$. We call the quantities the ith-*order* directional derivatives along the sequence $H_i = (h_1 \dots, h_i)$, $1 \le i \le m$. The higher-order directional derivative no longer acts linearly, but $\nabla^i F(x_*)$ is homogeneous of degree i in the sense that

$$\nabla^i F(x_*)(\varepsilon h_1, \dots, \varepsilon^i h_i) = \varepsilon^i \nabla^i F(x_*)(h_1, \dots, h_i).$$

In particular, no indices $j_1$and j_2 with $j_1 + j_2 > i$ can occur together as arguments in any of the terms in $\nabla^i F(x_*)$. Thus all the vectors h_j whose index satisfies $2j > i$ appear linearly and in separate terms. There are linear operators $G_k = G_k[F](x_*; H_{k-1})$, $k \in \mathbb{N}$, depending on the derivatives up to order k of F in the point x_* (i.e., the k-jet of F in x_*) and the vectors $H_{k-1} = (h_1, \dots, h_{k-1})$, which describe the contributions of these components. In fact, for $k \in \mathbb{N}$ we have $G_k = G_k[F](x_*; H_{k-1}) : X \to Y,\ v \longmapsto G_k(v)$, defined by

$$G_k[F](x_*; H_{k-1})(v) = \sum_{r=0}^{k-1} \frac{1}{r!} \left(\sum_{j_1 + \ldots + j_r = k-1} F^{(r+1)}(x_*)(h_{j_1}, \dots, h_{j_r}, v) \right). \tag{3.3}$$

For simplicity of notation we often suppress the arguments. For example, we write

$$G_1\,(v) = F'(x_*)v, \qquad G_2\,(v) = F''(x_*)(h_1, v),$$
$$G_3(v) = F''(x_*)(h_2, v) + \frac{1}{2} F'''(x_*)(h_1, h_1, v).$$

Given an order p we can therefore separate the linear contributions of the vectors $h_p, \dots, h_{2p-1}$ in derivatives of orders p through $2p-1$ and have for $i = 1, \dots, p$ that

$$\begin{aligned} \nabla^{p-1+i} F(x_*)(h_1, \dots, h_{p-1+i}) = &\left(\sum_{k=1}^{i} G_k[F](x_*; H_{k-1})(h_{p+i-k}) \right) \\ &+ R_{p-1,i}[F](x_*; H_{p-1}). \end{aligned} \tag{3.4}$$

Here the sum gives all the terms which contain a vector $h_p, \dots, h_{p-1+i}$, and the remainder R combines all the remaining terms which only include vectors of index

$\leq p-1$. Similar to the operators G_k the remainders R depend on the $(p-1+i)$-jet of the operator F at x_*, but if the map is clear we omit it in the notation. The general structure of these remainders is given by

$$R_{\ell,i}[F](x_*;H_\ell) = \sum_{r=2}^{\ell+i} \frac{1}{r!} \left(\sum_{\substack{j_1+\cdots+j_r=\ell+i \\ 1\leq j_k\leq \ell,\ 1\leq k\leq r}} F^{(r)}(x_*)(h_{j_1},\dots,h_{j_r}) \right). \tag{3.5}$$

Thus $R_{\ell,i}$ consists of the terms which are homogeneous of degree $\ell+i$, but only involve vectors from H_ℓ. In particular, the remainders only have contributions from derivatives of at least order two. In our applications $\ell = p-1$ and $i = 1,\dots,p$, but the structure above is general.

Definition 3.1. *We say the operator F is p-regular at x_* in direction of the sequence $H_{p-1} \in X^{p-1}$if the following conditions are satisfied:*

(A1) *$F: X \to Y$ is $(2p-1)$-times continuously Fréchet differentiable in a neighborhood of x_*.*

(A2) *The subspaces Y_i, $i = 1,\dots,p$,*

$$Y_i = \sum_{k=1}^{i} \operatorname{Im} G_k = \sum_{k=1}^{i} \operatorname{Im} G_k[F](x_*;H_{k-1})$$

are closed. Also let $Y_0 = \{0\}$.

(A3) *The map $\mathcal{G}_p = \mathcal{G}_p[F](x_*;H_{p-1})$*

$$\begin{aligned} \mathcal{G}_p \ &: \ X \to Z \doteq Y_1 \times Y_2/Y_1 \times \cdots \times Y/Y_{p-1}, \\ v &\mapsto \mathcal{G}_p(v) = (G_1(v), \pi_1 G_2(v), \dots, \pi_{p-1} G_p(v)), \end{aligned} \tag{3.6}$$

where the $\pi_i : Y_{i+1} \to Y_{i+1}/Y_i$ denote the canonical projections into the quotient space, is onto.

In the sense of this definition 1-regularity then corresponds to the classical Lyusternik condition, while 2-regularity is similar to Avakov's definition [3].

The p-order Lyusternik theorem proven in [16] gives a precise description of the p-order approximating sequences for a p-regular operator.

Theorem 3.2 (p-order Lyusternik theorem). *Let $H_{p-1} = (h_1,\dots,h_{p-1}) \in X^{p-1}$ be given such that the ith directional derivatives of F vanish along H_i for $i = 1,\dots,p-1$,*

$$\nabla^i F(x_*)(H_i) = 0 \text{ for } i = 1,\dots,p-1, \tag{3.7}$$

and suppose the operator F is p-regular at x_* in direction of the sequence H_{p-1}. Then $H_p = (h_1,\dots,h_p) \in X^p$ is a p-order approximating sequence to the set Q at $x_* \in Q$ if and only if the following conditions hold for $i = 1,\dots,p$:

$$G_i[F](x_*;H_{i-1})h_p + R_{p-1,i}[F](x_*;H_{p-1}) \in Y_{i-1}. \tag{3.8}$$

We include a brief motivation of Theorem 3.2 and an outline of its proof. Full details are given in [16]. The conditions can be understood by looking at the Taylor expansion of F along a curve $x(\varepsilon) = x_* + \varepsilon h_1 + \varepsilon^2 h_2 + \cdots$. For reasons of simplicity, we consider the case $p = 2$:

$$\begin{aligned} F(x(\varepsilon)) = F(x_*) + \varepsilon F'(x_*)h_1 + \varepsilon^2(F'(x_*)h_2 + \frac{1}{2}F''(x_*)(h_1, h_1)) \\ +\varepsilon^3(F'(x_*)h_3 + F''(x_*)(h_1, h_2) + \frac{1}{6}F'''(x_*)(h_1, h_1, h_1)) + o(\varepsilon^3). \end{aligned}$$

Clearly we must have $F'(x_*)h_1 = 0$. This term cannot be affected by choosing higher-order directions and therefore must vanish generating (3.7). But if F is not regular at x_*, then $F'(x_*)h_1 = 0$ is no longer sufficient for h_1 to be a tangent vector to Q at x_*. For, if $F''(x_*)(h_1, h_1) \notin \operatorname{Im} F'(x_*)$, then it is not possible to choose h_2 to make the term at ε^2 vanish. Even stronger, we need to be able to choose h_2 so that it simultaneously cancels the quadratic term and reduces the cubic term to a vector which lies in $\operatorname{Im} F'(x_*)$. Only then it is possible to choose h_3 to make the term at ε^3 zero. This generates the two conditions in (3.8). The necessity of the conditions in Theorem 3.2 follows along these lines in general.

In order to prove sufficiency if F is p-regular at x_* in direction of the sequence H_{p-1}, we construct the remainder term r in the definition of p-order approximating sequences using Newton's method. This is the essential step of the argument and we briefly outline how it is done.

By (A2) the spaces Y_i, $i = 1, \dots, p-1$, are closed and therefore are Banach spaces. So then are the quotient spaces Y_{i+1}/Y_i, $i = 1, \dots, p-1$, and their product $Z \doteq Y_1 \times Y_2/Y_1 \times \cdots \times Y/Y_{p-1}$. By (A3) the linear and continuous operator $\mathcal{G}_p : X \to Z$ has a bounded right inverse $\mathcal{H}_p : Z \to X$, i.e., $\mathcal{G}_p(x_*, H_{p-1}) \circ \mathcal{H}_p = id_Z$, and there exists a constant C such that $||\mathcal{H}_p(z)|| \le C||z||$ for all $z \in Z$. Let $U = \{x \in X : ||x - x_*|| < \delta\}$ be a sufficiently small neighborhood of x_* so that F is p-times continuously differentiable on U. By choosing ε_0 sufficiently small the entire curve $x(\varepsilon) = x_* + \sum_{i=1}^{p} \varepsilon^i h_i$, $0 \le \varepsilon \le \varepsilon_0$ lies in U. Henceforth we consider one point $x(\varepsilon)$ on this curve; i.e., ε will be fixed. The desired function r is found by constructing a solution to the equation $F(x) = F(x_*)$ using Newton's method. To this end set $x_0 = x(\varepsilon)$, $z_0^{(i)} = \varepsilon^{p+i} y_i$, for $i = 1, \dots, p-1$, and then define

$$\varsigma_0 = z_0^{(p)} = F(x_0) - F(x_*) - \sum_{i=1}^{p-1} z_0^{(i)}. \tag{3.9}$$

It is convenient to have a separate label ς for the pth component and we use $z_n^{(p)}$ and ς_n interchangeably. Then inductively define sequences $\{x_n\}_{n\in\mathbb{N}} \subset X$, and $\{z_n^{(i)}\}_{n\in\mathbb{N}} \subset Y_i$, $i = 1, \dots, p$, as

$$x_n \doteq x_{n-1} - \mathcal{H}_p\left(z_{n-1}^{(1)}, \frac{\pi_1\left(z_{n-1}^{(2)}\right)}{\varepsilon}, \frac{\pi_2\left(z_{n-1}^{(3)}\right)}{\varepsilon^2}, \cdots, \frac{\pi_p\left(z_{n-1}^{(p)}\right)}{\varepsilon^{p-1}}\right), \tag{3.10}$$

and for $i = 1, \dots, p-1$ define

$$z_n^{(i)} \doteq z_{n-1}^{(i+1)} + \varepsilon^i G_{i+1}(x_n - x_{n-1}) \tag{3.11}$$

while

$$\varsigma_n = z_n^{(p)} \doteq F(x_n) - F(x_*) - \sum_{i=1}^{p-1} z_n^{(i)}. \tag{3.12}$$

Note that these sequences are well defined: applying the operator $\mathcal{G}_p$ to (3.10) we get that

$$G_1(x_n - x_{n-1}) = F'(x_*)(x_n - x_{n-1}) = -z_{n-1}^{(1)} \tag{3.13}$$

and for $i = 1, \dots, p-1$ that

$$\varepsilon^i \pi_i \left(G_{i+1}(x_n - x_{n-1})\right) = -\pi_i z_{n-1}^{(i+1)}. \tag{3.14}$$

Thus we have $\pi_i\left(z_n^{(i)}\right) = 0$, so that $z_n^{(i)} \in Y_i$ for all $i = 1, \dots, p-1$.

In the proof [16] the following inequalities are shown by induction for all $n \geq 0$ and $i = 1, \dots, p-1$:

$$||z_n^{(i)}|| \leq M \left(\frac{1}{2}\right)^n \varepsilon^{p+i} \tag{3.15}$$

and

$$||\varsigma_n|| \leq M \left(\frac{1}{2}\right)^n \varepsilon^{2p-1}\alpha(\varepsilon), \tag{3.16}$$

$$||x_{n+1} - x_n|| \leq pMC \left(\frac{1}{2}\right)^n \varepsilon^p \alpha(\varepsilon), \tag{3.17}$$

where $M = \max(1, ||y_i||,\ i = 1, \dots, p)$ and α is a fixed function of order $o(1)$ as $\varepsilon \to 0$. Therefore the sequences $\{z_n^{(i)}\}_{n\in\mathbb{N}},\ i = 1, \dots, p$, all converge to 0 and $\{x_n\}_{n\in\mathbb{N}}$ is Cauchy and hence convergent to some limit, which we call $l(\varepsilon)$. By the continuity of F in x_* and (3.12) it follows that

$$F(l(\varepsilon)) = \lim_{n\to\infty} F(x_n) = F(x_*).$$

Defining

$$r(\varepsilon) = l(\varepsilon) - x(\varepsilon) = \left(\lim_{n\to\infty} x_n\right) - x_0,$$

we therefore have

$$F(x_*) = F\left(x(\varepsilon) + r(\varepsilon)\right)$$

and

$$||r(\varepsilon)|| = \left\|\sum_{n=1}^{\infty}(x_n - x_{n-1})\right\| \leq pMC\sum_{n=1}^{\infty}\left(\frac{1}{2}\right)^{n-1}\varepsilon^p\alpha(\varepsilon) = pMC\varepsilon^p\alpha(\varepsilon)$$

is of order $o(\varepsilon^p)$ as $\varepsilon \to 0$. This verifies the theorem.

3.2. High-order tangent cones. Note that (3.8) implies as necessary condition for the existence of a p-order approximating sequence along H_{p-1} that for $i = 1, \dots, p-1$, we have the following *compatability conditions*:

$$R_{p-1,i}[F](x_*; H_{p-1}) \in Y_i. \tag{3.18}$$

For $i = p$ this condition is satisfied under our assumption that $\mathcal{G}_p$ is onto. We can therefore restate Theorem 3.2 as follows.

COROLLARY 3.3. *Let H_{p-1} be a sequence so that $\nabla^i F(x_*)(H_i) = 0$ for $i = 1, \dots, p-1$, and suppose the operator F is p-regular at x_* in direction of H_{p-1}. Then $TS^{(p)}(Q; x_*, H_{p-1})$ is nonempty if and only if for $i = 1, \dots, p-1$ we have*

$$R_{p-1,i}[F](x_*; H_{p-1}) \in Y_i. \tag{3.19}$$

In this case $TS^{(p)}(Q; x_*, H_{p-1})$ is the closed affine subspace of X given by the solutions to the linear equation

$$\mathcal{G}_p[F](x_*; H_{p-1})(v) + \mathcal{R}_{p-1}[F](x_*, H_{p-1}) = 0, \tag{3.20}$$

where $\mathcal{R}_{p-1}[F](x_*, H_{p-1}) \in Z$ is the point with components

$$(R_{p-1,1}[F](x_*; H_{p-1}), \pi_1 R_{p-1,2}[F](x_*; H_{p-1}), \dots, \pi_{p-1} R_{p-1,p}[F](x_*; H_{p-1})) . \tag{3.21}$$

This formulation of the result clearly brings out the geometric structure of the p-order tangent sets as closed affine linear subspaces of X generated by the kernel of $\mathcal{G}_p$, $\ker \mathcal{G}_p$.

COROLLARY 3.4. *Let H_{p-1} be a sequence such that the operator F is p-regular at x_* in direction of H_{p-1}, that the first $(p-1)$ directional derivatives $\nabla^i F(x_*)(H_i)$ vanish for $i = 1, \dots, p-1$, and that the compatability conditions $R_{p-1,i}[F](x_*; H_{p-1}) \in Y_i$ are satisfied for $i = 1, \dots, p$. Then the p-order tangent cone to $Q = \{x \in X : F(x) = F(x_*)\}$ at x_* in direction of H_{p-1}, $TC^{(p)}(Q; x_*, H_{p-1})$, consists of all solutions $(w, \gamma) \in X \times \mathbb{R}_+$ (i.e., $\gamma > 0$) of the linear equation*

$$\mathcal{G}_p[F](w) + \gamma \mathcal{R}_{p-1}[F](x_*, H_{p-1}) = 0. \tag{3.22}$$

For applications to optimization problems we need the subspace of continuous linear functionals which annihilate $\ker \mathcal{G}_p$. Since the operator $\mathcal{G}_p$ is onto by assumption (A3), it follows by the annihilator lemma or the closed-range theorem [11] that

$$(\ker \mathcal{G}_p)^{\perp} = \operatorname{Im}\left(\mathcal{G}_p^*\right), \tag{3.23}$$

where

$$\mathcal{G}_p^* : Z^* = Y_1^* \times (Y_2/Y_1)^* \times \cdots \times (Y/Y_{p-1})^* \to X^*$$

denotes the adjoint map. Let

$$\tau_i : (Y_{i+1}/Y_i)^* \to Y_i^{\perp_{i+1}} \tag{3.24}$$

denote the canonical isomorphism. Here $\perp_{i+1}$ denotes the annihilator in Y_{i+1}, i.e.,

$$Y_i^{\perp_{i+1}} = \{z^* \in Y_{i+1}^* : \langle z^*, v\rangle = 0 \ \forall \ v \in Y_i\},$$

and we formally set $Y_0 = \{0\}$, so that $Y_0^{\perp_1} \cong Y_1^*$. Then, more specifically, we get the following.

PROPOSITION 3.5. *A functional $\lambda \in X^*$ lies in $(\ker \mathcal{G}_p)^{\perp}$ if and only if it can be represented in the form*

$$\lambda = \sum_{i=1}^{p} G_i^*[F](x_*; H_{i-1}) y_i^* \tag{3.25}$$

for some functionals $y_i^* \in Y_{i-1}^{\perp_i}$, $i = 1, \dots, p$.

Proof. By the annihilator lemma, $\lambda \in (\ker \mathcal{G}_p)^{\perp}$ if and only if there exists a continuous linear functional $\widetilde{y}^* = (\widetilde{y}_1^*, \dots .\widetilde{y}_p^*) \in Z^*$, i.e., $\widetilde{y}_i^* \in (Y_i/Y_{i-1})^*$, so that

$$\lambda = \mathcal{G}_p^* \widetilde{y}^* = \mathcal{G}_p^* \left(\widetilde{y}_1^*, \widetilde{y}_2^*, \dots , \widetilde{y}_p^*\right). \tag{3.26}$$

Thus we have for every $v \in \ker \mathcal{G}_p$

$$\begin{aligned} 0 = \langle \lambda, v\rangle &= \left\langle \mathcal{G}_p^* \widetilde{y}^*, v\right\rangle = \langle \widetilde{y}^*, \mathcal{G}_p v\rangle \\ &= \left\langle \left(\widetilde{y}_1^*, \widetilde{y}_2^*, \dots , \widetilde{y}_p^*\right), \left(G_1(v), \pi_1 G_2(v), \dots , \pi_{p-1} G_p(v)\right)\right\rangle \\ &= \sum_{i=1}^{p} \langle \widetilde{y}_i^*, \pi_{i-1} G_i(v)\rangle . \end{aligned}$$

Define $y_1^* \doteq \widetilde{y}_1^*$ and $y_i^* \doteq \tau_{i-1} \circ \widetilde{y}_i^*$ for $i = 2, \dots , p$. Since $\tau_{i-1} : (Y_i/Y_{i-1})^* \to Y_{i-1}^{\perp_i}$ is the canonical isomorphism, we have that $\tau_{i-1} \circ\, z^* = z^* \circ \pi_{i-1}$ for every $z^* \in Y_i$. Hence

$$\sum_{i=1}^{p} \langle \widetilde{y}_i^*, \pi_{i-1} G_i(v)\rangle = \sum_{i=1}^{p} \langle y_i^*, G_i(v)\rangle = \sum_{i=1}^{p} \langle G_i^* y_i^*, v\rangle .$$

Since this holds for all $v \in \ker \mathcal{G}_p$, (3.25) follows. □

PROPOSITION 3.6. *The dual or polar p-order tangent cone consists of all linear functionals $(\lambda, \mu) \in X^* \times \mathbb{R}$ which can be represented in the following form: There exist functionals $y_i^* \in Y_{i-1}^{\perp_i}$, $i = 1, \dots , p$, and a number $r \geq 0$ such that*

$$\lambda = \sum_{i=1}^{p} G_i^*[F](x_*; H_{i-1}) y_i^* \ , \tag{3.27}$$

$$\mu = \sum_{i=1}^{p} \langle y_i^*, R_{p-1,i}[F](x_*; H_{p-1})\rangle + r. \tag{3.28}$$

Proof. This proof is analogous to the proof of [14, Theorem 2.2] and thus we indicate only the main steps. Define the operator

$$\begin{aligned} \mathcal{K} \ &: \ X \times \mathbb{R} \to Z = Y_1 \times Y_2/Y_1 \times \cdots \times Y/Y_{p-1}, \\ (w, \gamma) &\mapsto \mathcal{K}(w, \gamma) = \mathcal{G}_p[F](w) + \gamma \mathcal{R}_{p-1}[F](x_*, H_{p-1}). \end{aligned}$$

Then

$$TC^{(p)}(Q; x_*, H_{p-1}) = \ker \mathcal{K}(w, \gamma) \cap \{(w, \gamma) \in X \times \mathbb{R} : \ \gamma > 0\}$$

and thus

$$\left(TC^{(p)}(Q; x_*, H_{p-1})\right)^* = (\ker \mathcal{K})^* + \{0\} \times \{r \in \mathbb{R} : \ r \geq 0\}.$$

Since $\ker \mathcal{K}$ is a linear space, the continuous functionals which are nonnegative on $\ker \mathcal{K}$ are given by $(\ker \mathcal{K})^{\perp}$ and by the annihilator lemma

$$(\ker \mathcal{K})^{\perp} = \operatorname{Im} \mathcal{K}^*.$$

Now the result follows analogously as in the proof of Proposition 3.5. □

4. High-order cones of decrease. In this section we determine the p-order sets of decrease of a functional $I : X \to \mathbb{R}$. These results also apply to p-order feasible sets to inequality constraints defined by smooth functionals. We assume as given a $(p-1)$-order sequence H_{p-1} and we calculate the p-order set of decrease of I at x_* along H_{p-1}. Trivial cases arise if there exists a first non-zero directional derivative $\nabla^i I(x_*)(H_i)$ of I with $i \le p-1$. In this case we have either $DS^{(p)}(I; x_*, H_{p-1}) = \emptyset$ if $\nabla^i I(x_*)(H_i) > 0$ or $DS^{(p)}(I; x_*, H_{p-1}) = X$ if $\nabla^i I(x_*)(H_i) < 0$. In the first case the sequence H_{p-1} cannot be used to exclude optimality of x_* since indeed x_* is a local minimum along the approximating curve generated by H_{p-1}. In the second case h_i is an ith-order direction of decrease along H_{i-1} and thus every vector $v \in X$ is admissible as a pth-order component. The only nontrivial case arises if $\nabla^i I(x_*)(H_i) = 0$ for all i with $i \le p-1$ and if $I'(x_*) \ne 0$. Then replacing the operator F with the functional I in (3.4) for $i = 1$, it follows that $DS^{(p)}(I; x_*, H_{p-1})$ consists of all vectors $v \in X$ which satisfy

$$\nabla^p I(x_*)(H_{p-1}, v) = I'(x_*)v + R_{p-1,1}[I](x_*; H_{p-1}) < 0.$$

Here we indicate in the notation that the operator $R_{p-1,1}$ is taken for the functional I. If $I'(x_*) = 0$, then $DS^{(p)}(I; x_*, H_{p-1})$ is still the full space or empty set depending on whether $R_{p-1,1}[I](x_*; H_{p-1})$ is negative or positive. Thus, if $I'(x_*) \ne 0$, then $DS^{(p)}(I; x_*, H_{p-1})$ is an open half-space with normal vector $I'(x_*)$. Summarizing we have (see also [14, Proposition 3.1])

PROPOSITION 4.1. *Suppose $I'(x_*) \ne 0$ and $\nabla^i I(x_*)(H_i) = 0$ for all i with $i \le p-1$. Then the p-order cone of decrease for the functional I at x_* in direction of H_{p-1}, $DC^{(p)}(I; x_*, H_{p-1})$, is given by*

$$\begin{aligned} DC^{(p)}(I; x_*, H_{p-1}) = \{(w, \gamma) \in X \times \mathbb{R} : \gamma > 0, \\ I'(x_*)w + \gamma R_{p-1,1}[I](x_*; H_{p-1}) < 0\}. \end{aligned} \tag{4.1}$$

Thus $DC^{(p)}(I; x_*, H_{p-1})$ is nonempty, open, and convex. The dual or polar cone to $DC^{(p)}(I; x_*, H_{p-1})$ can easily be calculated using the Minkowski–Farkas lemma [8].

PROPOSITION 4.2. *Suppose $I'(x_*) \ne 0$ and $\nabla^i I(x_*)(H_i) = 0$ for all i with $i \le p-1$. Then*

$$\begin{aligned} \left(DC^{(p)}(I; x_*, H_{p-1})\right)^* = \Bigg\{ &(\lambda, \mu) \in X^* \times \mathbb{R} : \ \exists\ \alpha_1 \le 0,\ \alpha_2 \ge 0 \text{ such that} \\ &\begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \begin{pmatrix} I'(x_*) & 0 \\ R_{p-1,1}[I](x_*; H_{p-1}) & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \Bigg\}. \end{aligned} \tag{4.2}$$

Proof. Define a linear operator $H : X \times \mathbb{R} \to \mathbb{R}^2$ by

$$H(w, \gamma) = \begin{pmatrix} I'(x_*) & R_{p-1,1}[I](x_*; H_{p-1}) \\ 0 & 1 \end{pmatrix} \begin{pmatrix} w \\ \gamma \end{pmatrix}$$

and let K denote the open cone $K = \{(\alpha_1, \alpha_2) \in \mathbb{R}^2 : \ \alpha_1 < 0,\ \alpha_2 > 0\}$. The polar cone K^* can be identified with the closure of K. Then

$$DC^{(p)}(I; x_*, H_{p-1}) = \{(v, \gamma) \in X \times \mathbb{R} : \ H(w, \gamma) \in K\}.$$

This cone is nonempty and open. Hence it follows by the Minkowski–Farkas lemma [8, Thm. 10.4(a)] that

$$\left(DC^{(p)}(I; x_*, H_{p-1})\right)^* = H^* K^* \cong H^* \ Clos\, K.$$

This proves the proposition. □

Remark. The formulas give the general formulations of well-known formulas for the cone of decrease [8] or the expressions for second-order cones of decrease derived in [14].

5. Generalized necessary conditions for optimality for a problem with equality constraints. We include a brief derivation of the generalized necessary conditions for optimality for the minimization problem under equality constraints which does not use the Dubovitskii–Milyutin framework of Theorem 2.7. Instead this argument is based on the geometry of p-order tangent sets/cones and clearly brings out the geometric meaning of the necessary conditions. Also for this problem nonregularity of the equality constraint is equivalent to the occurrence of abnormal points.

Consider the problem (E) to minimize a functional $I : X \to \mathbb{R}$ over the set $Q = \{x \in X : \ F(x) = 0\}$ where $F : X \to Y$ is an operator between Banach spaces. We assume the following:

(B0) $x_* \in Q$ is a local minimum for problem (E).

(B1) The functional I and the operator F are sufficiently often continuously Fréchet differentiable in a neighborhood of x_*.

Naturally, a p-order approximation along $H_{p-1} = (h_1, \dots , h_{p-1}) \in X^{p-1}$ would be considered only if a $(p-1)$-order approximation remained inconclusive. Thus we also assume as given a sequence $H_{p-1} = (h_1, \dots , h_{p-1}) \in X^{p-1}$ with the following properties:

(B2) The first $p-1$ directional derivatives of F along H_{p-1} vanish,

$$\nabla^i F(x_*)(H_i) = 0 \text{ for all } \ i = 1, \dots , p-1.$$

(B3) The compatability conditions are satisfied for $i = 1, \dots , p$

$$R_{p-1,i}(x_*; H_{p-1}) \in Y_i.$$

(B4) The operator F is p-regular at x_* in direction of the sequence H_{p-1}.

(B5) $\nabla^i I(x_*)(H_i) = 0$ for all $i = 1, \dots , p-1$.

Remark. By Theorem 3.2 and assumptions (B2)–(B4) there exist vectors $v \in X$ such that $H = (h_1, \dots , h_{p-1}, v)$ is a p-order approximating sequence to the set Q at $x_* \in Q$. Hence there exists an $\varepsilon_0 > 0$ and a function r defined on $[0, \varepsilon_0]$ with values in X, $r : [0, \varepsilon_0] \to X$, which is of order $o(\varepsilon^p)$ as $\varepsilon \to 0$ with the property that $x(\varepsilon) \doteq x_* + \sum_{i=1}^{p-1} \varepsilon^i h_i + \varepsilon^p v + r(\varepsilon) \in Q$. By assumption (B0) the function φ: $[0, \varepsilon_0] \to X$, $\varepsilon \to \varphi(\varepsilon) = I(x(\varepsilon))$, has a local minimum at $\varepsilon = 0$ and by assumption (B1) φ is p-times differentiable at $\varepsilon = 0$ from the right. Modulo a factorial multiplicative constant these derivatives are precisely the ith-order directional derivatives $\nabla^i I(x_*)(H_i)$ of I along the sequence H_i, $1 \le i \le p$. The first $(p-1)$ derivatives only depend on the sequence $H_{p-1} = (h_1, \dots , h_{p-1})$ and do not involve the p-order approximating vector v. Since x_* is optimal for problem (E), the first nonvanishing derivative must be positive. If this happens for a derivative $i < p$, then the necessary condition for minimality along $H = (h_1, \dots , h_{p-1}, v)$ is satisfied and indeed x_* is the minimum of I along $x(\varepsilon)$ for $\varepsilon > 0$ small enough. No further information can then be gained by considering a p-order approximating sequence along H_{p-1}. The only nontrivial situation therefore arises under assumption (B5).

Also suppose that $I'(x_*) \neq 0$. It then follows from Proposition 4.1 that the p-order set of decrease $DS^{(p)}(I; x_*, H_{p-1})$ consists of all vectors $v \in X$ which satisfy

$$I'(x_*)v + R_{p-1,1}[I](x_*; H_{p-1}) < 0 \tag{5.1}$$

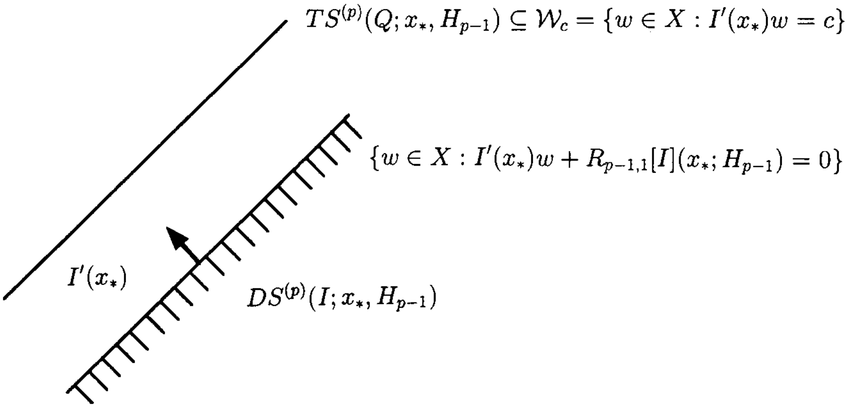

$TS^{(p)}(Q;x_*,H_{p-1}) \subseteq \mathcal{W}_c = \{w \in X : I'(x_*)w = c\}$
$\{w \in X : I'(x_*)w + R_{p-1,1}[I](x_*;H_{p-1}) = 0\}$
$I'(x_*)$
$DS^{(p)}(I;x_*,H_{p-1})$

FIG. 5.1. *Geometry of the p-order tangent set and p-order set of decrease.*

and thus is a half-space with normal vector $I'(x_*)$. As in the proof of Theorem 2.7 the optimality of x_* implies that

$$DS^{(p)}(I;x_*,H_{p-1}) \cap TS^{(p)}(Q;x_*,H_{p-1}) = \emptyset, \tag{5.2}$$

where $TS^{(p)}(Q;x_*,H_{p-1})$denotes the p-order tangent set to Q at x_* in direction of the sequence H_{p-1}.

Let $\mathcal{W}_c = \{w \in X : I'(x_*)w = c\}$, $c \in \mathbb{R}$, denote the family of hyperplanes in X parallel to $\ker I'(x_*)$. The empty intersection property (5.2) is then geometrically equivalent to the fact that there exists a $c \in \mathbb{R}$ such that

$$TS^{(p)}(Q;x_*,H_{p-1}) \subseteq \mathcal{W}_c \tag{5.3}$$

and

$$c + R_{p-1,1}[I](H_{p-1}) \geq 0. \tag{5.4}$$

Condition (5.3) states that $TS^{(p)}(Q;x_*,H_{p-1})$ is contained in a parallel translation of the boundary hyperplane of $DS^{(p)}(I;x_*,H_{p-1})$, while (5.4) enforces that $TS^{(p)}(Q;x_*,H_{p-1})$ lies to the right side of $DS^{(p)}(I;x_*,H_{p-1})$ (see Fig. 5.1).

Analytically, (5.3) is equivalent to

$$\ker \mathcal{G}_p \subseteq \ker I'(x_*) = \mathcal{W}_0 \tag{5.5}$$

and thus

$$I'(x_*) \in (\ker \mathcal{G}_p)^{\perp}. \tag{5.6}$$

Hence there exist multipliers $y_i^* \in Y_{i-1}^{\perp_i}$, $i = 1, \ldots, p$, so that

$$I'(x_*) + \sum_{i=1}^{p} G_i^* y_i^* \equiv 0. \tag{5.7}$$

Furthermore, we can evaluate the constant c by substituting an arbitrary vector $v \in TS^{(p)}(Q;x_*,H_{p-1})$. Using the same notation as in the proof of Proposition 3.5, we obtain

$$\begin{aligned} c &= \langle I'(x_*), v\rangle = -\left\langle \mathcal{G}_p^* \widetilde{y}^*, v\right\rangle = -\left\langle \widetilde{y}^*, \mathcal{G}_p v\right\rangle \\ &= \langle \widetilde{y}^*, \mathcal{R}_{p-1}[F](x_*; H_{p-1})\rangle \\ &= \sum_{i=1}^{p} \langle \widetilde{y}_i^*, \pi_{i-1} R_{p-1,i}[F](x_*; H_{p-1})\rangle \\ &= \sum_{i=1}^{p} \langle y_i^*, R_{p-1,i}[F](x_*; H_{p-1})\rangle \,. \end{aligned} \tag{5.8}$$

Thus we have proven the following result.

THEOREM 5.1. *Suppose assumptions* (B0)–(B5) *are satisfied and* $I'(x_*) \neq 0$. *Then there exist multipliers* $y_i^* \in Y_{i-1}^{\perp_i}$, $i = 1, \dots, p$, *so that the following* generalized Euler–Lagrange *equation and inequality condition are satisfied*:

$$0 \equiv I'(x_*) + \sum_{i=1}^{p} G_i^* y_i^*, \tag{5.9}$$

$$0 \leq R_{p-1,1}[I](x_*; H_{p-1}) + \sum_{i=1}^{p} \langle y_i^*, R_{p-1,i}[F](x_*; H_{p-1})\rangle \,. \tag{5.10}$$

The case when $I'(x_*) = 0$ is trivial in the sense that the necessary conditions are either trivially satisfied or violated.

6. High-order feasible cones.

6.1. High-order feasible cones to inequality constraints given by smooth functionals. In this section we calculate the form of the p-order feasible cones, $FC^{(p)}(P; x_*, H_{p-1})$, introduced in section 2, for inequality constraints P described by smooth functionals

$$P = \{x \in X : f(x) \leq 0\}.$$

As with sets of decrease, if there exists a first index $i \leq p-1$ such that $\nabla^i f(x_*)(H_i) \neq 0$, then the constraint will either be satisfied for any p-order vector $v \in X$ if $\nabla^i f(x_*)(H_i) < 0$ or it will be violated if $\nabla^i f(x_*)(H_i) > 0$. This leads to the definition of p-order active constraints.

DEFINITION 6.1. *The inequality constraint P is said to be* p-order active *along the sequence H_{p-1} if for all i, $i = 1, \dots, p-1$, we have $\nabla^i f(x_*)(H_i) = 0$.*

Only p-order active constraints enter the necessary conditions for optimality derived via p-order approximations along an admissible sequence H_{p-1}; p-order inactive constraints generate zero multipliers since $DS^{(p)}(P; x_*, H_{p-1}) = X$ (p-order complementary slackness conditions). If the constraint $P = \{x \in X : f(x) \leq 0\}$ is p-order active along the sequence H_{p-1}, then v is a p-order feasible vector for P at x_* in the direction of H_{p-1}, $v \in FS^{(p)}(P; x_*, H_{p-1})$, if and only if

$$\nabla^p f(x_*)(h_1, \dots, h_{p-1}, v) = f'(x_*)v + R_{p-1,1}[f](x_*; H_{p-1}) < 0. \tag{6.1}$$

Correspondingly the p-order feasible cone, $FC^{(p)}(P; x_*, H_{p-1})$, consists of all vectors $(w, \gamma) \in X \times \mathbb{R}_+$ which satisfy

$$f'(x_*)w + \gamma R_{p-1,1}[f](x_*; H_{p-1}) < 0. \tag{6.2}$$

Thus, if $f'(x_*) \neq 0$, then

$$FC^{(p)}(P; x_*, H_{p-1}) = \{(w, \gamma) \in X \times \mathbb{R} : \gamma > 0, \quad f'(x_*)w + \gamma R_{p-1,1}[f](x_*; H_{p-1}) < 0\} \tag{6.3}$$

and thus $FC^{(p)}(P; x_*, H_{p-1})$ is nonempty, open, and convex.

6.2. High-order feasible cones to closed convex inequality constraints. We also incorporate nonoperator inequality constraints described by closed convex sets with nonempty interior into the problem formulation. This is needed for instance for the application of the results to optimal control problems.

Let $C \subset X$ be a closed convex set with nonempty interior. Again we assume that H_{p-1} is a $(p-1)$-order feasible sequence. Note that it follows from Definition 2.3 that $FS^{(p)}(C; x_*, H_{p-1})$ is open (since any vector in the neighborhood V of v also lies in $FS^{(p)}(C; x_*, H_{p-1})$). It is also clear that $FS^{(p)}(C; x_*, H_{p-1})$ is convex, since C is. Thus $FC^{(p)}(C; x_*, H_{p-1})$ is an open, convex cone. Furthermore, if there exists an integer $j < p$ so that $h_j \in FS^{(j)}(C; x_*, H_{j-1})$, then any vector v is allowed as a p-order feasible direction and thus trivially $FS^{(p)}(C; x_*, H_{p-1}) = X$, i.e., the convex constraint $x \in C$ is not p-order active. In this case the necessary conditions for optimality along H_{p-1} are exactly the same as without C.

The dual or polar cone $FC^{(p)}(C; x_*, H_{p-1})^*$ can be identified with all supporting hyperplanes to $FS^{(p)}(C; x_*, H_{p-1})$ at x_*. More precisely, we have

$$FC^{(p)}(C; x_*, H_{p-1})^* = \{(\lambda, \mu) \in X^* \times \mathbb{R} : \langle \lambda, v\rangle + \mu \geq 0 \ \forall\, v \in FS^{(p)}(C; x_*, H_{p-1})\}.$$

An important property of p-order feasible sets to convex sets relates them to the classical feasible cones as they are defined in [8]: Given a set $S \subset X$ with nonempty interior, a vector $w \in X$ is called a feasible direction to S at x_* if there exist an $\varepsilon_0 > 0$ and a neighborhood W of w so that for all $0 < \varepsilon \leq \varepsilon_0$

$$x_* + \varepsilon W \subset S. \tag{6.4}$$

It is clear that the set of feasible directions to S at x_* is a cone, called the feasible cone to S at x_* and denoted by $FC(S; x_*)$.

Proposition 6.2. *Let $C \subset X$ be a closed convex set with nonempty interior, and let H_{p-1}be a $(p-1)$-order feasible sequence. If $FS^{(p)}(C; x_*, H_{p-1})$ is nonempty, then we have that*

$$FC(C; x_*) + FS^{(p)}(C; x_*, H_{p-1}) \subseteq FS^{(p)}(C; x_*, H_{p-1}) \tag{6.5}$$

and

$$FC(C; x_*) \times \{0\} \subseteq \mathit{Clos}\ FC^{(p)}(C; x_*, H_{p-1}). \tag{6.6}$$

Proof. Both statements hold trivially if $FC(C; x_*)$ is empty. Otherwise, let $w \in FC(C; x_*)$ and take any $v \in FS^{(p)}(C; x_*, H_{p-1})$. Since $FC(C; x_*)$ is a cone, given λ, $0 \leq \lambda < 1$, also $\frac{1}{1-\lambda} w \in FC(C; x_*)$. Thus it is possible to choose $\varepsilon_0 > 0$ and a neighborhood V of v so that for all $0 < \varepsilon \leq \varepsilon_0$ we have

$$x_* + \sum_{i=1}^{p-1} \varepsilon^i h_i + \varepsilon^p V \subset \operatorname{int} C$$

and

$$x_* + \varepsilon^p \frac{1}{1-\lambda} w \in C. \tag{6.7}$$

Here we use the fact that without loss of generality we can rescale ε as ε^p in the definition of feasible directions. Since C is convex, we therefore get that

$$x_* + \lambda \sum_{i=1}^{p-1} \varepsilon^i h_i + \varepsilon^p(\lambda V + w) \subset \operatorname{int} C. \tag{6.8}$$

This holds for all λ, $0 \le \lambda < 1$, and thus taking the limit as $\lambda \to 1$ we obtain $v + w \in FS^{(p)}(C; x_*, H_{p-1})$.

Since $FC(C; x_*)$ is a cone, we thus have for all $\rho > 0$ that

$$(\rho w + v, 1) \in FC^{(p)}(C; x_*, H_{p-1}). \tag{6.9}$$

Equivalently,

$$\left(w + \frac{1}{\rho} v, \frac{1}{\rho} \right) \in FC^{(p)}(C; x_*, H_{p-1})$$

and taking the limit $\rho \to \infty$

$$(w, 0) \in Clos\ FC^{(p)}(C; x_*, H_{p-1}). \qquad \Box$$

COROLLARY 6.3. *Let $C \subset X$ be a closed convex set with nonempty interior and suppose the p-order feasible set $FS^{(p)}(C; x_*, H_{p-1})$ is nonempty. If $(\lambda, \mu) \in FC^{(p)}(C; x_*, H_{p-1})^*$, then $\lambda \in FC(C; x_*)^*$ and thus λ is a supporting hyperplane to C at x_*.*

Proof. If $(\lambda, \mu) \in FC^{(p)}(C; x_*, H_{p-1})^*$, then for all $w \in FC(C; x_*)$

$$0 \le \left\langle (\lambda, \mu), \begin{pmatrix} w \\ 0 \end{pmatrix} \right\rangle = \langle \lambda, w \rangle$$

and thus $\lambda \in FC(C; x_*)^*$. $\Box$

7. Generalized necessary conditions for optimality. In this section we give generalized first- and second-order necessary conditions for optimality for problem (P) based on general p-order approximations. We assume as given a sequence $H_{p-1} = (h_1, \dots, h_{p-1}) \in X^{p-1}$ with the following properties:

(P1) The first $p-1$ directional derivatives of F along H_{p-1} vanish,

$$\nabla^i F(x_*)(H_i) = 0 \text{ for all } i = 1, \dots, p-1,$$

the compatability conditions

$$R_{p-1,i}[F](x_*; H_{p-1}) \in Y_i$$

are satisfied for $i = 1, \dots, p-1$, and the operator F is p-regular at x_* in direction of the sequence H_{p-1}.

(P2) Either the first nonvanishing derivative $\nabla^i I(x_*)(H_i)$ is negative or

$$\nabla^i I(x_*)(H_i) = 0 \quad \text{for} \quad i = 1, \dots, p-1.$$

(P3) If the jth constraint is not p-order active, then the first nonzero derivative $\nabla^i f(x_*)(H_i)$ is negative.

(P4) $FS^{(p)}(C; x_*, H_{p-1})$ is nonempty.

These conditions guarantee respectively that the corresponding p-order approximating cones to the constraints or the functional I are nonempty and convex. The next theorem is a generalization of results in [14] for $p = 2$ to a general order p. It gives a generalized version of the classical first-order necessary conditions for optimality for a mathematical programming problem with convex inequality constraints [8, Thm. 11.4].

THEOREM 7.1. *If x_* is optimal for problem* (P), *then given any sequence $H_{p-1} = (h_1, \dots, h_{p-1}) \in X^{p-1}$ for which conditions* (P1)–(P4) *are satisfied, there exist Lagrange multipliers $\nu_i \geq 0$, $i = 0, 1, \dots, m$, functionals $y_i^* \in Y_{i-1}^{\perp_i}$, $i = 1, \dots, p$, and a supporting hyperplane $\langle \lambda, v \rangle + \mu \geq 0$ for all $v \in FS^{(p)}(C; x_*, H_{p-1})$, all depending on the sequence H_{p-1}, such that the multipliers ν_i, $i = 0, 1, \dots, m$, and λ do not all vanish, and*

$$\lambda \equiv \nu_0 I'(x_*) + \sum_{j=1}^{m} \nu_j f_j'(x_*) + \sum_{i=1}^{p} G_i^* y_i^*, \tag{7.1}$$

$$\begin{aligned} \mu \leq \nu_0 R_{p-1,1}[I](x_*; H_{p-1}) &+ \sum_{j=1}^{m} \nu_j R_{p-1,1}[f_j](x_*; H_{p-1}) \\ &+ \sum_{i=1}^{p} \langle y_i^*, R_{p-1,i}[F](H_{p-1}) \rangle \,. \end{aligned} \tag{7.2}$$

Furthermore, the following p-order complementary slackness conditions hold:

· $\nu_0 = 0$ *if* $DS^{(p)}(I; x_*, H_{p-1}) = X$,

· $\nu_j = 0$ *if* $FS^{(p)}(P_j; x_*, H_{p-1}) = X$ *for* $j = 1, \dots, m$,

· $\lambda = 0$ *if* $FS^{(p)}(C; x_*, H_{p-1}) = X$.

Proof. Let $H_{p-1} = (h_1, \dots, h_{p-1}) \in X^{p-1}$ be a sequence for which conditions (P1)-(P4) are satisfied. We first take care of some special cases. For simplicity of notation set $I = f_0$ and suppose that $f_j'(x_*) = 0$ for some index $j \in \{0, 1, \dots, m\}$. If also $R_{p-1,1}[f_j](x_*; H_{p-1}) \geq 0$, then the conditions of the theorem can trivially be satisfied by taking $\nu_j = 1$ and setting all other multipliers equal to zero. On the other hand, if $R_{p-1,1}[f_j](x_*; H_{p-1}) < 0$, then the corresponding p-order feasible cone (respectively cone of decrease) will be the full space (with dual cone given by $C^* = \{0\}$). Since this will not put any restrictions on admissible p-order directions we may drop the corresponding constraint along with all the inactive constraints. The corresponding multipliers must be zero and this already implies the p-order complementary slackness conditions.

Without loss of generality (or after dropping the inactive constraints) we may therefore assume that all the constraints are p-order active and that none of the gradients $f_j'(x_*)$, $j = 0, 1, \dots, m$, vanishes. In this case the p-order cone of decrease is open, nonempty and convex given by Proposition 4.1 with dual given by Proposition 4.2. These formulas also describe the p-order feasible cones and their duals for the constraints $j = 1, \dots, m$. Furthermore, under assumption (P1) the p-order tangent cone is a closed nonempty subspace described in Corollary 3.4 with dual given in Proposition 3.6.

The assertions of the theorem therefore follow from Theorem 2.7 using the specific forms of the functionals given in these results: Using Propositions 3.6 and 4.2 it follows that there exist nonnegative constants ν_j and ζ_j, $j = 0, 1, \dots, m$, functionals $y_i^* \in Y_{i-1}^{\perp_i}$, $i = 1, \dots, p$, a constant $r \geq 0$, and a functional $(\lambda, \mu) \in FC^{(p)}(C; x_*, H_{p-1})^*$ such that

$$0 = -\nu_0 I'(x_*) - \sum_{j=1}^{m} \nu_j f_j'(x_*) - \sum_{i=1}^{p} G_i^* y_i^* + \lambda \tag{7.3}$$

and

$$\begin{aligned} 0 = &-\nu_0 R_{p-1,1}[I](x_*; H_{p-1}) + \zeta_0 - \sum_{j=1}^{m} (\nu_j R_{p-1,1}[f_j](x_*; H_{p-1}) + \zeta_j) \\ &- \sum_{i=1}^{p} \langle y_i^*, R_{p-1,i}[F](H_{p-1}) \rangle + r + \mu. \end{aligned} \tag{7.4}$$

The first equation gives the generalized Euler–Lagrange equation (7.1) and the inequality condition (7.2) is obtained by dropping the nonnegative terms arising in the $\mathbb{R}$-components of the functionals in the duals to the feasible cones and the tangent cone.

It remains to verify the nontriviality of the multipliers ν_j, $j = 0, 1, \dots, m$, and λ. If all of these vanish, then by (7.1) $\sum_{i=1}^{p} G_i^* y_i^* \equiv 0$. It follows from the proof of Proposition 3.5 that there exists a $\widetilde{y}^* = (\widetilde{y}_1^*, \dots .\widetilde{y}_p^*) \in Z^*$ such that with $y_1^* \doteq \widetilde{y}_1^*$ and $y_i^* \doteq \tau_{i-1} \circ \widetilde{y}_i^*$ for $i = 2, \dots, p$, we have for all $v \in X$ that

$$\langle \widetilde{y}^*, \mathcal{G}_p v \rangle = \sum_{i=1}^{p} \langle G_i^* y_i^*, v \rangle = 0.$$

Under assumption (P1) the operator $\mathcal{G}_p$ is onto and thus $\widetilde{y}^* = 0$ and consequently also $y_i^* = 0$ for $i = 1, \dots, p$. But this contradicts the statement about the nontriviality of the multipliers λ_j in Theorem 2.7. □

Remark. Theorem 2.7 gives the formulation for the case which is *nondegenerate* in the sense that the operator $\mathcal{G}_p$ is onto. If $\mathcal{G}_p$ is not onto, but still closed, while all the other conditions of Theorem 7.1 remain in effect, then a degenerate version of this theorem can easily be obtained by choosing a nontrivial multiplier $\widetilde{y}^* \in (\operatorname{Im} \mathcal{G}_p)^\perp$ which then gives rise to nontrivial multipliers $y_i^* \in Y_{i-1}^{\perp_i}$ which have the property that $\sum_{i=1}^{p} G_i^* y_i^* \equiv 0$. Thus (7.1) still holds if we set $\nu_j = 0$ for $j = 0, 1, \dots, m$, and $\lambda = 0$. Thus the difference is that it can only be asserted that not all of the multipliers ν_j, $j = 0, 1, \dots, m$, $y_i^* \in Y_{i-1}^{\perp_i}$, $i = 1, \dots, p$, and λ do vanish.

Remark. It follows from Corollary 6.3 that λ is also a supporting functional to C at x_*. However, this is weaker than the statement that (λ, μ) defines a supporting hyperplane to $FS^{(p)}(C; x_*, H_{p-1})$.

8. Conclusion. In this paper we derived necessary conditions for optimality for extremum problems in the presence of nonregular equality constraints. Our results are based on a generalized version of the Lyusternik Theorem which describes the structure of p-order approximating sequences if the operator $\mathcal{G}_p$ is onto. Coupled with p-order approximations to inequality constraints, the Dubovitskii–Milyutin framework was then used to derive the results. In this paper we present only the result for the general mathematical programming problem with inequality constraints described by smooth functionals, but we also include inequality constraints described by convex sets. This allows us to apply these results to the optimal control problem. For the case of second-order approximations this is done in [15]; for the general case this is outlined in [17, 18].

REFERENCES

[1] A. V. Arutyunov, *Higher-order conditions in abnormal extremal problems with constraints of equality type*, Soviet Math. Dokl., 42 (1991), pp. 799–804.

[2] A. V. Arutyunov, *Second-order conditions in extremal problems with a finite-dimensional image.* 2*-normal mappings*, Izv. Ross. Akad. Nauk Ser. Mat., 60 (1996), pp. 37–62 (in Russian).

[3] E. R. Avakov, *Extremum conditions for smooth problems with equality-type constraints*, USSR Comput. Math. and Math. Phys., 25 (1985), pp. 24–32 (translated from Zh. Vychisl. Mat. i Fiz., 25 (1985)).

[4] E. R. Avakov, *Necessary conditions for a minimum for nonregular problems in Banach spaces. Maximum principle for abnormal problems in optimal control*, Trudy Mat. Inst. Akad. Nauk. SSSR, 185 (1988), pp. 3–29 (in Russian).

[5] E. R. Avakov, *Necessary extremum conditions for smooth abnormal problems with equality- and inequality constraints*, Mat. Zametki, 45 (1989), pp. 3–11 (in Russian). Translated in J. Math. Notes, 45 (1989), pp. 431–437.

[6] A. Ben-Tal and J. Zowe, *A unified theory of first and second order conditions for extremum problems in topological vector spaces*, Mathematical Programming Study, 19 (1982), pp. 39–76.

[7] A. V. Dmitruk, *Quadratic conditions for a Pontryagin minimum in an optimal control problem linear in the control.* II. *Theorems on weakening equality constraints*, Math. USSR Izvestiya, 31 (1986), pp. 121–141.

[8] I. V. Girsanov, *Lectures on Mathematical Theory of Extremum Problems*, Lecture Notes in Econom. and Math. Systems 67, Springer-Verlag, Berlin, 1972.

[9] K. H. Hoffmann and H. J. Kornstaedt, *Higher-order necessary conditions in abstract mathematical programming*, J. Optim. Theory Appl., 26 (1978), pp. 533–568.

[10] A. D. Ioffe, *Necessary and sufficient conditions for a local minimum.* 3*: Second-order conditions and augmented duality*, SIAM J. Control Optim., 17 (1979), pp. 266–288.

[11] A. D. Ioffe and V. M. Tikhomirov, *Theory of Extremal Problems*, North–Holland, Amsterdam, The Netherlands, 1979.

[12] A. F. Izmailov, *Optimality conditions for degenerate extremum problems with inequality-type constraints*, Comput. Math. Math. Phys., 34 (1994), pp. 723–736.

[13] U. Ledzewicz, *An extension of the local maximum principle to abnormal optimal control problems*, J. Optim. Theory Appl., 77 (1993), pp. 661–681.

[14] U. Ledzewicz and H. Schättler, *Second-order conditions for extremum problems with nonregular equality constraints*, J. Optim. Theory Appl., 86 (1995), pp. 113–144.

[15] U. Ledzewicz and H. Schättler, *An extended maximum principle*, Nonlinear Anal., 29 (1997), pp. 159–183.

[16] U. Ledzewicz and H. Schättler, *A high-order generalization of the Lyusternik theorem*, Nonlinear Anal., 34 (1998), pp. 793–815.

[17] U. Ledzewicz and H. Schättler, *A high-order generalization of the Lyusternik theorem and its application to optimal control problems*, Dynamical Systems and Differential Equations, II, W. Chen and S. Hu, eds., Missouri State University, 1998, pp. 45–59.

[18] U. Ledzewicz and H. Schättler, *High-order extended maximum principles for optimal control problems with non-regular constraints*, in Optimal Control: Theory, Algorithms and Applications, W. W. Hager and P. M. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 298–325.

[19] E. S. Levitin, A. A. Milyutin, and N. P. Osmolovskii, *Conditions of higher order for a local minimum in problems with constraints*, Russian Math. Surveys, 33 (1978), pp. 97–168.

[20] A. A. Milyutin, *On quadratic conditions for an extremum in smooth problems with a finite-dimensional image*, in Methods of the Theory of Extremal Problems in Economics, V. L. Levin, ed., Nauka, Moscow, 1981, pp. 138–177 (in Russian).

[21] Zs. Páles and V. Zeidan, *Nonsmooth optimum problems with constraints*, SIAM J. Control Optim., 32 (1994), pp. 1476–1502.

[22] Zs. Páles and V. Zeidan, *First- and second-order necessary conditions for control problems with constraints*, Trans. Amer. Math. Soc., 346 (1994), pp. 421–453.

[23] A. A. Tretyakov, *Necessary and sufficient conditions for optimality of p-th order*, USSR Comput. Math. Math. Phys., 24 (1984), pp. 123–127.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 1, pp. 54–91

ON REPRESENTATIONS AND INTEGRABILITY OF MATHEMATICAL STRUCTURES IN ENERGY-CONSERVING PHYSICAL SYSTEMS[*]

MORTEN DALSMO[†] AND ARJAN VAN DER SCHAFT[‡]

Abstract. In the present paper we elaborate on the underlying Hamiltonian structure of interconnected energy-conserving physical systems. It is shown that a power-conserving interconnection of port-controlled generalized Hamiltonian systems leads to an implicit generalized Hamiltonian system, and a power-conserving partial interconnection to an implicit port-controlled Hamiltonian system. The crucial concept is the notion of a (generalized) Dirac structure, defined on the space of energy-variables or on the product of the space of energy-variables and the space of flow-variables in the port-controlled case. Three natural representations of generalized Dirac structures are treated. Necessary and sufficient conditions for closedness (or integrability) of Dirac structures in all three representations are obtained. The theory is applied to implicit port-controlled generalized Hamiltonian systems, and it is shown that the closedness condition for the Dirac structure leads to strong conditions on the input vector fields.

Key words. Hamiltonian systems, Dirac structures, implicit systems, external variables, integrability, actuated mechanical systems, kinematic constraints, interconnections

AMS subject classifications. 93C10, 93A30, 70F25, 58F05

PII. S0363012996312039

1. Introduction.

Most of the current modelling and simulation approaches to (complex) physical systems (e.g., multibody systems) are based on some sort of *network representation*, where the physical system under consideration is seen as the interconnection of a (possible large) number of simple subsystems. This way of modelling has several advantages. From a physical point of view it is usually natural to regard the system as composed of subsystems, possibly from different domains (mechanical, electrical, and so on). The knowledge about subsystems can be stored in libraries, and is reusable for later occasions. Because of the modularity the modelling process can be performed in an "iterative" manner, gradually refining—if necessary—the model by adding other subsystems. Further, the approach is suited to general control design where the overall behavior of the system is sought to be improved by the addition of other subsystems or controlling devices. From a system-theoretic point of view this modular approach naturally emphasizes the need for models of systems *with* external variables, e.g., inputs and outputs.

In this paper we concentrate on the mathematical description of network representations of (lumped-parameter) *energy-conserving* physical systems. In our previous work we have shown how energy-conserving physical systems with *independent* energy variables can be naturally described as generalized Hamiltonian systems (with external variables). However, a general power-conserving interconnection of such systems will lead to a system described by differential *and* algebraic equations, that is, an *implicit* dynamical system, which can no longer be directly described as an explicit

[*] Received by the editors November 13, 1996; accepted for publication (in revised form) January 22, 1998; published electronically September 25, 1998.
http://www.siam.org/journals/sicon/37-1/31203.html

[†] ABB Corporate Research, Department of Information Technology and Control Systems, P.O. Box 90, N-1361 Billingstad, Norway (morten.dalsmo@nocrc.abb.no).

[‡] University of Twente, Systems and Control Group, Department of Applied Mathematics, 7500 AE Enschede, The Netherlands (A.J.vanderSchaft@math.utwente.nl).

generalized Hamiltonian system. This motivates the definition of *implicit* generalized Hamiltonian systems, as introduced in [SM2, SM3]. The main ingredient in this definition is that of a (generalized) *Dirac structure*. The relevance of Dirac structures in the Hamiltonian modelling of electrical LC-circuits with dependent storage elements (a clear example of interconnected energy-conserving systems) was already recognized in [C2].

The notion of Dirac structures was introduced by Courant and Weinstein [CW] and further investigated by Courant in [C1] as a generalization of Poisson *and* (pre)-symplectic structures. Dorfman [D1, D2] developed an algebraic theory of Dirac structures in the context of the study of completely integrable systems of partial differential equations, with the aim of describing within a Hamiltonian framework certain sets of PDEs which do not admit an easy Hamiltonian formulation in terms of Poisson or symplectic structures, due to nonlocality of the involved operators. The conceptual novelty in the approach initiated in [C2, SM2, SM3] is to use Dirac structures for the direct Hamiltonian description of differential-algebraic equations resulting from the interconnection of energy-conserving systems, including constrained systems. Although the terminology Dirac structure is derived from the "Dirac bracket" introduced by Dirac in his study of *constrained* Hamiltonian systems arising from degenerate Lagrangians [D3], our use of Dirac structures determining, together with the stored energy (Hamiltonian), the algebraic constraints as well as the dynamical equations of motion seems to be new. Furthermore, we stress the "physical" relevance of Dirac structures as naturally capturing the geometric structure of the system as arising from the *interconnection* of subsystems (see e.g., Proposition 2.2).

In Courant and Dorfman [C1, D2] the definition of a Dirac structure includes a *closedness* (or integrability) condition generalizing the Jacobi-identity for Poisson brackets or the closedness of two-forms defining symplectic structures. This condition is naturally satisfied for *constant* Dirac structures (as in the case of LC-circuits) and for Dirac structures arising from holonomic kinematic constraints in mechanical systems, but *not* for the *generalized* Dirac structures arising from nonholonomic kinematic constraints [SM1, SM3] or from general kinematic pairs in multibody systems [M2].

The structure of this paper is as follows. In section 2 we will recall the definitions of a (generalized) Dirac structure and of an implicit Hamiltonian system, and we will show how the power-conserving interconnection of port-controlled (explicit) Hamiltonian systems leads to such an implicit Hamiltonian system. In section 3 we will investigate various useful ways of representing generalized Dirac structures and consequently of representing implicit Hamiltonian systems, and we will study their relationship. Then in section 4 the closedness (or integrability) condition for Dirac structures will be worked out for the three different representations obtained. Both sections 3 and 4 use extensively techniques and results from the work of Courant and Dorfman, although the emphasis is rather different. The results of sections 3 and 4 are applied in section 5 to Dirac structures as arising in implicit generalized Hamiltonian systems with external variables. In particular it is shown that the closedness condition translates into strong conditions on the input vector fields.

A main motivation for the Hamiltonian modelling of interconnected energy-conserving physical systems is, apart from the clear motivation from a general modelling and simulation point of view, the generalization of the theory of "passivity-based control" to complex interconnected physical systems. Key concepts in this theory (see, e.g., [TA, OS, S]) are the use of the internal energy as candidate Lyapunov function, the shaping of the internal energy via state feedback, and the injection of "damping" in order to achieve asymptotic stability. This approach has shown to be very powerful in the robust and/or adaptive control of physical systems described by Euler–Lagrange or Hamiltonian equations of motion (such as robot manipulators, mobile robots, and electrical machines) and can be expected to be equally powerful for interconnected physical systems. Although it is not the topic of the present paper to demonstrate this, we indicate at the end of section 4 how the usual stability theory of Hamiltonian systems based on the Hessian matrix of the Hamiltonian can be naturally extended to implicit Hamiltonian systems. Moreover, at the end of section 5 we show the link between results in this paper and "passivity-based control" of actuated mechanical systems with kinematic constraints.

In the control design of interconnected physical systems also the system-theoretic properties (such as controllability and observability) of implicit port-controlled Hamiltonian systems will prove to be instrumental (e.g., in the analysis how much damping injection is needed for asymptotic stabilization). For explicit port-controlled generalized Hamiltonian systems some of these topics already have been studied in our previous work [SM2, MS1, MS2]. Section 5 provides only a basic framework for a study of these issues. Apart from "passivity-based control", the further exploitation of the structure of symmetries and conservation laws also has a great potential (see, e.g., [BKMM] for related developments). All this is a large area for further research.

2. Generalized Hamiltonian modelling of interconnected systems. In our previous work [MS1, MS2, MBS, MSB1, MSB2, SM1, SM2, SM3] we have argued that the basic dynamic building blocks in the network representation of energy-conserving physical systems are systems of the form

$$\begin{aligned} \dot{x} &= J(x)\tfrac{\partial H}{\partial x}(x) + g(x)f, \\ e &= g^T(x)\tfrac{\partial H}{\partial x}(x). \end{aligned} \tag{2.1}$$

Here $x = (x_1, \dots, x_n)$ denotes the vector of (independent) energy variables, defining local coordinates for the state space manifold $\mathcal{X}$, $H(x_1, \dots, x_n)$ is the total stored energy in the system, with $\frac{\partial H}{\partial x}(x)$ denoting the column-vector of partial derivatives of H, and the $n \times n$ *skew-symmetric* structure matrix $J(x)$ is associated with the network topology of the system. The columns $g_j(x)$, $j = 1, \dots, m$, of the matrix $g(x)$ define the (state modulated) transformers describing the influence of the external *flow* sources (or *inputs*) f_j, $j = 1, \dots, m$. The components e_j of e are the corresponding conjugated (with respect to the power) *efforts* (or *outputs*). Since the matrix $J(x)$ is skew-symmetric we immediately obtain the energy balance

$$\frac{\mathrm{d}}{\mathrm{dt}} H = e^T f \tag{2.2}$$

expressing that the increase in energy equals the externally supplied *power* ($e_j f_j$ is the power of the jth source). Thus (2.1) describes an energy-conserving physical system with *internal* variables $x_1, \dots, x_n$ (associated with energy storage) and *external* (or *port*) variables $f_1, \dots, f_m, e_1, \dots, e_m$ (associated with power), which can be regarded, respectively, as input and output variables.

The system (2.1) is called a port-controlled generalized *Hamiltonian* system because of the following. We may define a generalized *Poisson bracket* operation on the real functions on $\mathcal{X}$ as

$$\{F, G\}(x) = \left[\frac{\partial F}{\partial x}(x)\right]^T J(x)\, \frac{\partial G}{\partial x}(x), \quad F, G : \mathcal{X} \to \mathbb{R}, \tag{2.3}$$

Clearly, this bracket is skew-symmetric and satisfies the Leibniz identity

$$\{F, G_1G_2\}(x) = \{F, G_1\}(x)G_2(x) + G_1(x)\{F, G_2\}(x) \text{ for all } F, G_1, G_2 : \mathcal{X} \to \mathbb{R} \tag{2.4}$$

and thus $\dot{x} = J(x)\frac{\partial H}{\partial x}(x)$ can be seen as the *generalized* Hamiltonian vector field corresponding to H and the generalized Poisson bracket $\{\, ,\, \}$. This generalized Poisson bracket is a true Poisson bracket if additionally the *Jacobi-identity* is satisfied, that is,

$$\{F, \{G, K\}\} + \{G, \{K, F\}\} + \{K, \{F, G\}\} = 0 \quad \text{for all } F, G, K : \mathcal{X} \to \mathbb{R}. \tag{2.5}$$

If (and only if) the Jacobi-identity holds, there exist in a neighborhood of every point $x_0 \in \mathcal{X}$ where $J(x)$ has constant rank local *canonical* coordinates $(q, p, r) = (q_1, \dots, q_k, p_1, \dots, p_k, r_1, \dots, r_l)$ for $\mathcal{X}$ in which $J(x)$ takes the form (see e.g., [O])

$$J(q, p, r) = \begin{bmatrix} 0 & I_k & 0 \\ -I_k & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{2.6}$$

implying that the Hamiltonian vector field $\dot{x} = J(x)\frac{\partial H}{\partial x}(x)$ takes the form

$$\begin{array}{rcl} \dot{q} & = & \frac{\partial H}{\partial p}(q, p, r), \\ \dot{p} & = & -\frac{\partial H}{\partial q}(q, p, r), \\ \dot{r} & = & 0 \end{array} \tag{2.7}$$

which are almost the standard Hamiltonian equations of motion except for the appearance of the conserved quantities $r_1, \dots, r_l$. Although in many cases of interest the Jacobi-identity is satisfied, there are clear examples where it is not satisfied (e.g., mechanical systems with nonholonomic kinematic constraints; see [SM1]).

The overall energy-conserving physical system is now obtained by interconnecting the various port-controlled generalized Hamiltonian subsystems as above in a power-continuous fashion (e.g., by using Kirchhoff's laws). In general this will result in a *mixed* set of differential and algebraic equations, which nevertheless is expected to be again Hamiltonian in some sense. Indeed, it can be seen that it is an *implicit* generalized Hamiltonian system, as defined in [SM2, SM3]. The key concept in the definition of an implicit generalized Hamiltonian system is the notion of a generalized Dirac structure, as introduced (in a rather different context) in [C1, D2].

First we concentrate on interconnected energy-conserving physical systems without any remaining external sources; see section 5 for the general case. In this case the Dirac structure for the interconnected system is defined solely on the space of energy-variables. Let $\mathcal{X}$ be an n-dimensional manifold with tangent bundle $T\mathcal{X}$ and cotangent bundle $T^*\mathcal{X}$. We define $T\mathcal{X} \oplus T^*\mathcal{X}$ as the smooth vector bundle over $\mathcal{X}$ with fiber at each $x \in \mathcal{X}$ given by $T_x\mathcal{X} \times T_x^*\mathcal{X}$. Let X be a smooth vector field and α a smooth one-form on $\mathcal{X}$ respectively. Then we say that the pair (X, α) belongs to a smooth vector subbundle $\mathcal{D} \subset T\mathcal{X} \oplus T^*\mathcal{X}$ (denoted $(X, \alpha) \in \mathcal{D}$) if $(X(x), \alpha(x)) \in \mathcal{D}(x)$ for every $x \in \mathcal{X}$. Furthermore for a smooth vector subbundle $\mathcal{D} \subset T\mathcal{X} \oplus T^*\mathcal{X}$ we define the smooth vector subbundle $D^\perp \subset T\mathcal{X} \oplus T^*\mathcal{X}$ as

$$\mathcal{D}^\perp = \{(X, \alpha) \in T\mathcal{X} \oplus T^*\mathcal{X} \mid \langle \alpha \mid \hat{X} \rangle + \langle \hat{\alpha} \mid X \rangle = 0, \text{ for all } (\hat{X}, \hat{\alpha}) \in \mathcal{D}\} \tag{2.8}$$

with $\langle \mid \rangle$ denoting the natural pairing between a one-form and a vector field. In (2.8) and throughout in the sequel the pairs $(X, \alpha), (\hat{X}, \hat{\alpha})$ are assumed to be pairs of smooth vector fields and smooth one-forms.

DEFINITION 2.1 (see [C1, D2]). *A generalized Dirac structure on an n-dimensional manifold $\mathcal{X}$ is a smooth vector subbundle $\mathcal{D} \subset T\mathcal{X} \oplus T^*\mathcal{X}$ such that $\mathcal{D}^\perp = \mathcal{D}$.*

If $\mathcal{D}$ satisfies an additional *closedness* (or integrability) condition, then $\mathcal{D}$ defines a Dirac structure; see section 4. Later on we will see that the *dimension* of the fibers of a generalized Dirac structure on an n-dimensional manifold is equal to n. By taking $\hat{\alpha} = \alpha$, $\hat{X} = X$ in (2.8) we obtain

$$\langle \alpha \,|\, X \rangle = 0 \quad \text{for all } (X, \alpha) \in \mathcal{D}. \tag{2.9}$$

Conversely, if (2.9) holds, then for every $(X, \alpha), (\hat{X}, \hat{\alpha}) \in \mathcal{D}$

$$\begin{aligned} 0 = \langle \alpha + \hat{\alpha} \,|\, X + \hat{X} \rangle &= \langle \alpha \,|\, X \rangle + \langle \alpha \,|\, \hat{X} \rangle + \langle \hat{\alpha} \,|\, X \rangle + \langle \hat{\alpha} \,|\, \hat{X} \rangle \\ &= \langle \alpha \,|\, \hat{X} \rangle + \langle \hat{\alpha} \,|\, X \rangle, \end{aligned} \tag{2.10}$$

and thus $\mathcal{D} \subset \mathcal{D}^\perp$. Hence a Dirac structure is a smooth vector subbundle of $T\mathcal{X} \oplus T^*\mathcal{X}$ which is *maximal* with respect to property (2.10) or (2.9).

Let now $\mathcal{X}$ be an n-dimensional manifold with a generalized Dirac structure $\mathcal{D}$, and let $H : \mathcal{X} \to \mathbb{R}$ be a Hamiltonian (energy function). Then the *implicit* generalized Hamiltonian system on $\mathcal{X}$ corresponding to $\mathcal{D}$ and H is given by the specification (see [SM2])

$$\left(\dot{x}, \frac{\partial H}{\partial x}(x) \right) \in \mathcal{D}(x). \tag{2.11}$$

By (2.9) we immediately obtain the energy conservation property $\frac{dH}{dt} = \langle \frac{\partial H}{\partial x}(x) | \dot{x} \rangle = 0$. Note that in general the specification (2.11) puts algebraic constraints on $\mathcal{X}$, since in general there will not exist for every $x \in \mathcal{X}$ a tangent vector $\dot{x} \in T_x\mathcal{X}$ such that (2.11) is satisfied. Thus (2.11) is in general a set of differential algebraic equations (DAEs). It can be seen that (2.11) generalizes the notion of an (explicit) generalized Hamiltonian system

$$\dot{x} = J(x) \frac{\partial H}{\partial x}(x), \quad J(x) = -J^T(x), \tag{2.12}$$

by noting that

$$\mathcal{D} = \{ (X, \alpha) \in T\mathcal{X} \oplus T^*\mathcal{X} \,|\, X(x) = J(x)\alpha(x),\ x \in \mathcal{X} \}$$

defines a generalized Dirac structure. (If $\alpha^T(x)J(x)\hat{\alpha}(x) + \hat{\alpha}^T(x)X(x) = 0$ for all $\hat{\alpha}$, then $X(x) = J(x)\alpha(x)$.)

A special case of a Dirac structure is that of a *constant* Dirac structure on a *linear* space.

DEFINITION 2.2. *A constant Dirac structure on a linear n-dimensional space $\mathcal{V}$ is a linear subspace $\mathcal{D} \subset \mathcal{V} \times \mathcal{V}^*$ with the property that $\mathcal{D}^\perp = \mathcal{D}$, where*

$$\mathcal{D}^\perp = \{ (v, v^*) \in \mathcal{V} \times \mathcal{V}^* \,|\, \langle v^* \,|\, \hat{v} \rangle + \langle \hat{v}^* \,|\, v \rangle = 0 \textit{ for all } (\hat{v}, \hat{v}^*) \in \mathcal{D} \}$$

where $\langle \mid \rangle$ denotes the natural pairing between $\mathcal{V}$ and $\mathcal{V}^*$.

The following proposition is derived straightforwardly.

PROPOSITION 2.1. *Let $\mathcal{V}$ be an n-dimensional linear space. A linear subspace $\mathcal{D} \subset \mathcal{V} \times \mathcal{V}^*$ defines a constant Dirac structure if and only if* $\dim \mathcal{D} = n$ *and*

$$\langle v^* \mid v \rangle = 0 \quad \textit{for all } (v, v^*) \in \mathcal{D}. \tag{2.13}$$

Proof. (Sketch; see [SM3] for details.) As in (2.9) and (2.10) we see that if $\mathcal{D}$ defines a constant Dirac structure, then (2.13) holds, while if (2.13) holds, then equivalently

$$\langle v^* \mid \hat{v} \rangle + \langle \hat{v}^* \mid v \rangle = 0 \quad \text{for all } (\hat{v}, \hat{v}^*) \in \mathcal{D}. \tag{2.14}$$

Furthermore, a subspace $\mathcal{D}$ of $\mathcal{V} \times \mathcal{V}^*$ defines a Dirac structure if it is *maximal* with respect to property (2.14), which is equivalent (see [C1]) to the property $\dim \mathcal{D} = n$. □

Now let us consider k port-controlled generalized Hamiltonian systems as in (2.1), i.e., for $i = 1, \ldots, k$

$$\begin{aligned} \dot{x}_i &= J_i(x_i)\tfrac{\partial H_i}{\partial x_i}(x_i) + g_i(x_i) f_i, \\ e_i &= g_i^T(x_i)\tfrac{\partial H_i}{\partial x_i}(x_i), \\ x_i &\in \mathcal{X}_i,\ f_i \in \mathcal{F}_i := \mathbb{R}^{m_i},\ e_i \in \mathcal{E}_i := \mathcal{F}_i^* = \mathbb{R}^{m_i}, \end{aligned} \tag{2.15}$$

with $\mathcal{X}_i$ an n_i-dimensional state space. Consider a general *power-conserving interconnection* of these systems given by an $(m_1 + \ldots + m_k)$-dimensional subspace (possibly parametrized by $x_1, \ldots, x_k$)

$$I(x_1, \ldots, x_k) \subset \mathcal{F}_1 \times \cdots \times \mathcal{F}_k \times \mathcal{E}_1 \times \cdots \times \mathcal{E}_k \tag{2.16}$$

with the property

$$(f_1, \ldots, f_k, e_1, \ldots, e_k) \in I(x_1, \ldots, x_k) \ \Rightarrow \sum_{i=1}^{k} e_i^T f_i = 0. \tag{2.17}$$

REMARK 2.1. *By Proposition* 2.1 *it follows that $I(x_1, \ldots, x_k)$ defines a constant Dirac structure on $\mathcal{F}_1 \times \ldots \times \mathcal{F}_k$, parameterized by $(x_1, \ldots, x_k)$.*

PROPOSITION 2.2. *Consider k port-controlled generalized Hamiltonian systems* (2.15) *subject to an interconnection* (2.16) *satisfying* (2.17)*. Then the resulting interconnected system is an implicit generalized Hamiltonian system with state space $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$, Hamiltonian $H(x_1, \ldots, x_k) := H_1(x_1) + \cdots + H_k(x_k)$, and generalized Dirac structure $\mathcal{D}$ on $\mathcal{X}$ given as*

$$\begin{aligned} &(X, \alpha) = (X_1, \ldots, X_k, \alpha_1, \ldots, \alpha_k) \in \mathcal{D} \iff \\ &\textit{for all } x_i \in \mathcal{X}_i,\ \ i = 1, \ldots, k,\ \ \exists (f_1, \ldots, f_k, e_1, \ldots, e_k) \in I(x_1, \ldots, x_k) \textit{ such that} \\ &X_i(x_i) = J_i(x_i)\alpha_i(x_i) + g_i(x_i) f_i, \\ &e_i = g_i^T(x_i)\alpha_i(x_i). \end{aligned} \tag{2.18}$$

Proof. The main point is in proving that $\mathcal{D}$ given by (2.18) defines a generalized Dirac structure. Let $(X,\alpha)=(X_1,\dots,X_k,\alpha_1,\dots,\alpha_k)$ be in $\mathcal{D}^\perp$, that is, $\langle\hat\alpha\,|\,X\rangle+\langle\alpha\,|\,\hat X\rangle=0$ for all $(\hat X,\hat\alpha)=(\hat X_1,\dots,\hat X_k,\hat\alpha_1,\dots,\hat\alpha_k)$ satisfying (2.18). This means

$$\begin{aligned}
0&=\sum_{i=1}^k\Big[\hat\alpha_i^T(x_i)X_i(x_i)+\alpha_i^T(x_i)\hat X_i(x_i)\Big]\\
&=\sum_{i=1}^k\Big[\hat\alpha_i^T(x_i)X_i(x_i)+\alpha_i^T(x_i)J_i(x_i)\hat\alpha_i(x_i)+\alpha_i^T(x_i)g_i(x_i)\hat f_i\Big]\\
&=\sum_{i=1}^k\Big(\hat\alpha_i^T(x_i)\left[X_i(x_i)-J_i(x_i)\alpha_i(x_i)\right]+\alpha_i^T(x_i)g_i(x_i)\hat f_i\Big) \qquad (2.19)
\end{aligned}$$

for all $\hat\alpha_i,\hat f_i$ such that $\hat e_i=g_i^T(x_i)\hat\alpha_i(x_i)$ satisfies $(\hat f_1,\dots,\hat f_k,\hat e_1,\dots,\hat e_k)\in I(x_1,\dots,x_k)$. Letting first $\hat f_i=0$ and $\hat e_i=0$, we obtain

$$\sum_{i=1}^k\hat\alpha_i^T(x_i)\left[X_i(x_i)-J_i(x_i)\alpha_i(x_i)\right]=0 \qquad (2.20)$$

for all $\hat\alpha_i(x_i)$ such that $g_i^T(x_i)\hat\alpha_i(x_i)=0$. This means that there exist vectors $f_1,\dots,f_k$ such that

$$X_i(x_i)=J_i(x_i)\alpha_i(x_i)+g_i(x_i)f_i. \qquad (2.21)$$

Substitution into (2.19) yields

$$\begin{aligned}
0&=\sum_{i=1}^k\Big(\hat\alpha_i^T(x_i)g_i(x_i)f_i+\alpha_i^T(x_i)g_i(x_i)\hat f_i\Big)\\
&=\sum_{i=1}^k\Big(\hat e_i^Tf_i+e_i^T\hat f_i\Big) \qquad (2.22)
\end{aligned}$$

for all $\hat f_i$ and $\hat e_i=g_i^T(x_i)\hat\alpha_i(x_i)$ satisfying $(\hat f_1,\dots,\hat f_k,\hat e_1,\dots,\hat e_k)\in I(x_1,\dots,x_k)$. If $g_i^T(x_i)$ is surjective for all $i=1,\dots,k$, this means that (2.22) is satisfied for all $(\hat f_1,\dots,\hat f_k,\hat e_1,\dots,\hat e_k)\in I(x_1,\dots,x_k)$, and by Proposition 2.1 and Remark 2.1 this implies that $(f_1,\dots,f_k,e_1,\dots,e_k)\in I(x_1,\dots,x_k)$ and thus $(X,\alpha)\in\mathcal{D}$. In general we proceed as follows. Define the space of achievable flows and efforts

$$C(x_1,\dots,x_k):=\{(\hat f_1,\dots,\hat f_k,\hat e_1,\dots,\hat e_k)\,|\,\hat f_i\in\mathcal{F}_i,\ \hat e_i\in\operatorname{Im}g_i^T(x_i),\ i=1,\dots,k\}.$$

Then (2.22) implies that

$$\begin{aligned}
(f_1,\dots,f_k,e_1,\dots,e_k)&\in\left(I(x_1,\dots,x_k)\cap C(x_1,\dots,x_k)\right)^\perp\\
&=I^\perp(x_1,\dots,x_k)+C^\perp(x_1,\dots,x_k)
\end{aligned}$$

where $\perp$ denotes orthogonal complement with respect to property (2.22). By Proposition 2.1 it follows that $I^\perp(x_1,\dots,x_k)=I(x_1,\dots,x_k)$, while $C^\perp(x_1,\dots,x_k)$ is seen to be given as

$$C^\perp(x_1,\dots,x_k)=\{(f_1,\dots,f_k,e_1,\dots,e_k)\,|\,f_i\in\ker\,g_i(x_i),\ e_i=0,\ i=1,\dots,k\}.$$

Thus there exist flow vectors $f'_1, \dots, f'_k$ such that $(f'_1, \dots, f'_k, e_1, \dots, e_k) \in I(x_1, \dots, x_k)$, with $X_i(x_i) = J_i(x_i)\alpha_i(x_i) + g_i(x_i)f'_i$, $e_i = g_i^T(x_i)\alpha_i(x_i)$, showing that $(X, \alpha) \in \mathcal{D}$. Hence $\mathcal{D}^\perp \subset \mathcal{D}$. Since it is easily seen that $\mathcal{D} \subset \mathcal{D}^\perp$, this shows that $\mathcal{D}$ defines a Dirac structure. $\square$

We note that the definition of a power-conserving interconnection is very general and for example includes Kirchhoff's laws for electrical systems, the interconnection relations for generalized velocities and forces for interconnected mechanical systems (Newton's third law), as well as transformers in electrical circuits and kinematic pairs in multibody systems.

From a classical control point of view an important example of a power-conserving interconnection is the standard *feedback interconnection.*

EXAMPLE 2.1. *Consider two input-state-output systems* ("*plant*" *and* "*controller*")

$$\begin{aligned} \dot{x}_i &= g_i(x_i, u_i), \\ y_i &= h_i(x_i), \quad u_i, y_i \in \mathbb{R}^m, \quad i = 1, 2, \end{aligned} \tag{2.23}$$

and impose the (*negative*) *feedback interconnection*

$$\begin{aligned} u_2 &= y_1, \\ u_1 &= -y_2, \end{aligned} \tag{2.24}$$

leading to the explicit system

$$\begin{aligned} \dot{x}_1 &= g_1(x_1, -h_2(x_2)), \\ \dot{x}_2 &= g_2(x_2, h_1(x_1)). \end{aligned} \tag{2.25}$$

If we equate the input vectors u_i with flow vectors, and the output vectors y_i with effort vectors, then (2.24) *is a power-conserving interconnection. Proposition* 2.2 *applied to this particular case says that if both systems in* (2.23) *are Hamiltonian, then also* (2.25) *is Hamiltonian. This can be regarded as a special instance of the* passivity theorem *in input-output stability theory.*

3. Representations of generalized Dirac structures and implicit generalized Hamiltonian systems. There are different ways of representing generalized Dirac structures, and consequently of writing the equations of an implicit generalized Hamiltonian system. These representations each have their own advantages and are connected to different but equivalent ways of mathematically modelling the energy-conserving physical systems.

Before going into these representations we first note that a generalized Dirac structure $\mathcal{D}$ on an n-dimensional manifold $\mathcal{X}$ defines the smooth distributions

$$\begin{aligned} G_0 &= \{X \in T\mathcal{X} \mid (X, 0) \in \mathcal{D}\}, \\ G_1 &= \{X \in T\mathcal{X} \mid \exists \alpha \in T^*\mathcal{X} \text{ s.t. } (X, \alpha) \in \mathcal{D}\} \end{aligned} \tag{3.1}$$

and the smooth codistributions

$$\begin{aligned} P_0 &= \{\alpha \in T^*\mathcal{X} \mid (0, \alpha) \in \mathcal{D}\}, \\ P_1 &= \{\alpha \in T^*\mathcal{X} \mid \exists X \in T\mathcal{X} \text{ s.t. } (X, \alpha) \in \mathcal{D}\}. \end{aligned} \tag{3.2}$$

Define for any smooth distribution G the smooth codistribution $\operatorname{ann} G$ as

$$\operatorname{ann} G = \{\alpha \in T^*\mathcal{X} \mid \langle \alpha \mid X \rangle = 0 \text{ for all } X \in G\} \tag{3.3}$$

and for any smooth codistribution P the smooth distribution ker P as

$$\ker P = \{X \in T\mathcal{X} \mid \langle \alpha \mid X \rangle = 0 \text{ for all } \alpha \in P\}. \tag{3.4}$$

The smooth (co)distributions G_0, G_1 and P_0, P_1 are related as follows.

Proposition 3.1. *Let $\mathcal{D}$ be a generalized Dirac structure on $\mathcal{X}$ and define G_0, G_1, P_0, P_1 as in* (3.1), (3.2). *Then*

1. $G_0 = \ker P_1$, $P_0 = \operatorname{ann} G_1$;
2. $P_1 \subset \operatorname{ann} G_0$, $G_1 \subset \ker P_0$, *with equality if G_1, respectively, P_1, is constant-dimensional.*

Proof.

1. $Z \in G_0$ if and only if $(Z, 0) \in \mathcal{D}$, if and only if

$$\langle 0 \mid X \rangle + \langle \alpha \mid Z \rangle = 0 \text{ for all } (X, \alpha) \in \mathcal{D}$$

or equivalently $\langle \alpha \mid Z \rangle = 0$ for all $\alpha \in P_1$. Thus $G_0 = \ker P_1$. Similarly $\beta \in P_0$ if and only if $(0, \beta) \in \mathcal{D}$, if and only if $\langle \beta \mid X \rangle = 0$ for all $X \in G_1$, which implies $P_0 = \operatorname{ann} G_1$.

2. This follows from property 1 and the inequalities $P \subset \operatorname{ann} \ker P$, $G \subset \ker \operatorname{ann} G$, for any smooth (co)distribution P and G, with equality if P and G are constant-dimensional [NS]. □

Remark 3.1. *The distribution G_1 and the co-distribution P_1 have the following interpretation. Consider the implicit generalized Hamiltonian system* (2.11) *corresponding to a generalized Dirac structure $\mathcal{D}$ and a Hamiltonian H. Then the distribution G_1 describes the set of admissible flows $\dot{x}$. In particular, if G_1 is constant-dimensional and involutive then there are $(n - \dim G_1)$ independent conserved quantities for* (2.11). *Dually the codistribution P_1 describes the set of algebraic constraints of* (2.11), *i.e.*,

$$\frac{\partial H}{\partial x}(x) \in P_1(x). \tag{3.5}$$

Definition 3.1. *A point $x \in \mathcal{X}$ is a regular point for the Dirac structure $\mathcal{D}$ on $\mathcal{X}$ if the dimension of G_1 and P_1 (and hence, see Proposition* 3.1, *of G_0, P_0) is constant in a neighborhood of x.*

At every regular point $x \in X$ we have

$$\mathcal{D}^{\perp}(x) = \{(v, v^*) \in T_x\mathcal{X} \times T_x^*\mathcal{X} \mid \langle v^* \mid \hat{v} \rangle + \langle \hat{v}^* \mid v \rangle = 0 \text{ for all } (\hat{v}, \hat{v}^*) \in \mathcal{D}(x)\}, \tag{3.6}$$

and since $\mathcal{D}^{\perp}(x) = \mathcal{D}(x)$, we may regard $\mathcal{D}(x) \subset T_x\mathcal{X} \times T_x^*\mathcal{X}$ as a *constant* Dirac structure on $T_x\mathcal{X}$ (see Definition 2.2). Invoking Proposition 2.1 we deduce that $\dim \mathcal{D}(x) = n$ for every regular point $x \in \mathcal{X}$. Since the set of regular points is open and dense in $\mathcal{X}$, and $\mathcal{D}$ is a vector subbundle, it thus follows that

$$\dim \mathcal{D}(x) = n \quad \text{for all } x \in \mathcal{X}, \tag{3.7}$$

and therefore we may regard $\mathcal{D}(x) \subset T_x\mathcal{X} \times T_x^*\mathcal{X}$ as a constant Dirac structure on $T_x\mathcal{X}$ for *every* $x \in \mathcal{X}$. In particular it follows, since $\mathcal{D}$ is a smooth vector subbundle, that locally about every point in $\mathcal{X}$ we may find $n \times n$ matrices $E(x)$ and $F(x)$, depending smoothly on x, such that locally

$$\begin{aligned} &\mathcal{D}(x) = \{(v, v^*) \in T_x\mathcal{X} \times T_x^*\mathcal{X} \mid F(x)v = E(x)v^*\}, \\ &\operatorname{rank}[F(x) \ : \ -E(x)] = n. \end{aligned} \tag{3.8}$$

Furthermore, because $\mathcal{D} = \mathcal{D}^{\perp}$ necessarily (see [SM2])

$$E(x)F^T(x) + F(x)E^T(x) = 0. \tag{3.9}$$

We will refer to this *local* representation (3.8), (3.9) of a Dirac structure as *representation* I. Given a Hamiltonian $H : \mathcal{X} \to \mathbb{R}$ the corresponding implicit generalized Hamiltonian system in representation I is locally given as

$$F(x)\dot{x} = E(x)\frac{\partial H}{\partial x}(x). \tag{3.10}$$

Example 3.1 ([SM2]; see also [MSB2]). *An LC-circuit is composed of a set of* (*multiport*) *inductors and capacitors interconnected through their ports by the network graph. An n-port inductor is defined by flux linkage variables* $\phi \in \mathbb{R}^n$ (*the energy variables*) *and an energy function* $H_L(\phi)$. *The port variables are the voltages* $v_L \in \mathbb{R}^n$ *and the currents* $i_L \in \mathbb{R}^n$ *defined as*

$$v_L = \dot{\phi}, \;\; i_L = \frac{\partial H_L}{\partial \phi}. \tag{3.11}$$

Similarly, an n-port capacitor is defined by charge variables $q \in \mathbb{R}^n$ *and energy function* $H_C(q)$, *with port variables the currents* $i_C \in \mathbb{R}^n$ *and voltages* $v_C \in \mathbb{R}^n$ *defined as*

$$i_C = \dot{q}, \;\; v_C = \frac{\partial H_C}{\partial q}. \tag{3.12}$$

By Kirchhoff's laws we obtain $n_L + n_C$ *independent equations*

$$F_C i_C + E_C i_L = 0, \;\; F_L v_L + E_L v_C = 0 \tag{3.13}$$

for certain matrices F_C, F_L, E_C, *and* E_L *satisfying (Tellegen's theorem)*

$$E_C F_L^T + F_C E_L^T = 0. \tag{3.14}$$

Using (3.11), (3.12) *and defining the total energy* $H(q, \phi) = H_L(\phi) + H_C(q)$, *we may rewrite* (3.13) *as the implicit generalized Hamiltonian system*

$$\underbrace{\begin{bmatrix} F_C & 0 \\ 0 & F_L \end{bmatrix}}_{F} \begin{bmatrix} \dot{q} \\ \dot{\phi} \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & -E_C \\ -E_L & 0 \end{bmatrix}}_{E} \begin{bmatrix} \frac{\partial H}{\partial q} \\ \frac{\partial H}{\partial \phi} \end{bmatrix}, \tag{3.15}$$

where $EF^T + FE^T = 0$ *by* (3.14).

Two other useful types of representations of generalized Dirac structures, which admit a *global* and *coordinate-free* definition, can be given provided an extra regularity condition is satisfied. We will denote them as representation II and representation III, respectively.

Theorem 3.1 (representation II). *Let* $\mathcal{X}$ *be an n-dimensional manifold. Let* G *be a constant-dimensional distribution on* $\mathcal{X}$, *and* $J(x) : T_x^*\mathcal{X} \to T_x\mathcal{X}$, $x \in \mathcal{X}$, *a skew-symmetric vector bundle map. Then*

$$\mathcal{D} = \{(X, \alpha) \in T\mathcal{X} \oplus T^*\mathcal{X} \mid X(x) - J(x)\alpha(x) \in G(x), x \in \mathcal{X}, \alpha \in \operatorname{ann} G\} \tag{3.16}$$

defines a generalized Dirac structure. Conversely, let $\mathcal{D}$ *be any generalized Dirac structure having the property that the codistribution* P_1 (*see* (3.2)) *is constant-dimensional.*

Then there exists a skew-symmetric vector bundle map $J(x) : P_1(x) \to (P_1(x))^*$, $x \in \mathcal{X}$, *which locally can be extended to a skew-symmetric vector bundle map* $J(x) : T_x^*\mathcal{X} \to T_x\mathcal{X}$, $x \in \mathcal{X}$, *such that* $\mathcal{D}$ *is given by* (3.16) *with* $G := \ker P_1$.

Proof (see also [C1] for the constant case). Let $\mathcal{D}$ be given by (3.16). We have to show that $\mathcal{D}^\perp = \mathcal{D}$.

1. Take $(X, \alpha) = (J\alpha + Z, \alpha) \in \mathcal{D}$, with $Z \in G$. Then for all $(\hat{X}, \hat{\alpha}) = (J\hat{\alpha} + \hat{Z}, \hat{\alpha}) \in \mathcal{D}$, $\hat{Z} \in G$

$$\langle \alpha \,|\, \hat{X} \rangle + \langle \hat{\alpha} \,|\, X \rangle = \langle \alpha \,|\, J\hat{\alpha} \rangle + \langle \hat{\alpha} \,|\, J\alpha \rangle + \langle \alpha \,|\, \hat{Z} \rangle + \langle \hat{\alpha} \,|\, Z \rangle = 0$$

because $J(x)$ is skew-symmetric, and $\alpha, \hat{\alpha} \in \operatorname{ann} G$.

2. Take $(X, \alpha) \in \mathcal{D}^\perp$, that is for all $(\hat{X}, \hat{\alpha}) = (J\hat{\alpha} + \hat{Z}, \hat{\alpha}) \in \mathcal{D}$, $\hat{Z} \in G$, $\hat{\alpha} \in \operatorname{ann} G$

$$0 = \langle \alpha \,|\, \hat{X} \rangle + \langle \hat{\alpha} \,|\, X \rangle = \langle \alpha \,|\, J\hat{\alpha} \rangle + \langle \alpha \,|\, \hat{Z} \rangle + \langle \hat{\alpha} \,|\, X \rangle$$

First let $\hat{Z} = 0$. Then

$$0 = \langle \alpha \,|\, J\hat{\alpha} \rangle + \langle \hat{\alpha} \,|\, X \rangle = \langle \hat{\alpha} \,|\, X - J\alpha \rangle$$

for all $\hat{\alpha} \in \operatorname{ann} G$, implying that $X - J\alpha \in \ker \operatorname{ann} G = G$, since G is constant-dimensional. Now let $\hat{\alpha} = 0$. Then

$$0 = \langle \alpha \,|\, \hat{Z} \rangle$$

for all $\hat{Z} \in G$, implying that $\alpha \in \operatorname{ann} G$.

Conversely, let $\mathcal{D}$ be a generalized Dirac structure on $\mathcal{X}$, with P_1 constant-dimensional. Then we define for every $x \in \mathcal{X}$ a linear map

$$J(x) : P_1(x) \subset T_x^*\mathcal{X} \to (P_1(x))^* \subset T_x\mathcal{X}$$

as follows. Let $v^* \in P_1(x)$, that is, there exists $v \in T_x\mathcal{X}$ such that $(v, v^*) \in \mathcal{D}(x)$. Then define

$$J(x)v^* = v \in (P_1(x))^*. \tag{3.17}$$

To see that $J(x)$ is well-defined, let also $(\hat{v}, v^*) \in \mathcal{D}(x)$. Then $(v - \hat{v}, 0) \in \mathcal{D}(x)$, which means $v - \hat{v} \in G_0(x) = \ker P_1(x)$, and thus v and $\hat{v}$ define the same linear function on $P_1(x)$. Skew-symmetry of the map $J(x) : P_1(x) \to (P_1(x))^*$ follows from

$$\langle \hat{v}^* \,|\, v \rangle + \langle v^* \,|\, \hat{v} \rangle = 0.$$

for all $(v, v^*), (\hat{v}, \hat{v}^*) \in \mathcal{D}(x)$. Finally we may locally *extend* $J(x)$ to a skew-symmetric map from $T_x\mathcal{X}$ to $T_x^*\mathcal{X}$. Now, let $(v, v^*) \in \mathcal{D}(x)$. Then by (3.17) $v = J(x)v^*$ modulo $G(x) := \ker P_1(x)$, while $v^* \in P_1(x)$, and thus $\mathcal{D}$ is indeed given by (3.16). □

REMARK 3.2. *Note* (*see* (3.17)) *that the kernel of* $J(x) : P_1(x) \to (P_1(x))^*$ *is given by* $P_0(x)$.

Given a Hamiltonian $H : \mathcal{X} \to \mathbb{R}$ the equations of the implicit generalized Hamiltonian system corresponding to representation II now take the form

$$\begin{array}{rcl} \dot{x} &=& J(x)\frac{\partial H}{\partial x}(x) + g(x)\lambda, \\ 0 &=& g^T(x)\frac{\partial H}{\partial x}(x), \end{array} \tag{3.18}$$

where $g(x)$ is any full rank matrix such that $\operatorname{Im} g(x) = G(x)$. The variables λ can be seen as Lagrange multipliers, required to keep the constraint equations $g^T(x)\frac{\partial H}{\partial x}(x) = 0$ to be satisfied for all time. Note that (3.18) can be also interpreted as a port-controlled generalized Hamiltonian system (see section 2) with the efforts (or outputs) e set equal to zero.

"Dualizing" representation II we obtain the following.

THEOREM 3.2 (representation III). *Let $\mathcal{X}$ be an n-dimensional manifold. Let P be a constant-dimensional codistribution on $\mathcal{X}$, and $\omega(x) : T_x\mathcal{X} \to T_x^*\mathcal{X}$, $x \in \mathcal{X}$, a skew-symmetric vector bundle map. Then*

$$\mathcal{D} = \{(X, \alpha) \in T\mathcal{X} \oplus T^*\mathcal{X} \mid \alpha(x) - \omega(x)X(x) \in P(x), x \in \mathcal{X}, X \in \ker P\} \tag{3.19}$$

defines a generalized Dirac structure. Conversely, let $\mathcal{D}$ be any generalized Dirac structure having the property that the distribution G_1 (*see* (3.1)) *is constant-dimensional. Then there exists a skew-symmetric vector bundle map $\omega(x) : G_1(x) \to (G_1(x))^*$, $x \in \mathcal{X}$, which locally can be extended to a skew-symmetric vector bundle map $\omega(x) : T_x\mathcal{X} \to T_x^*\mathcal{X}$, $x \in \mathcal{X}$, such that $\mathcal{D}$ is given by* (3.19) *with $P := \operatorname{ann} G_1$.*

Proof. The proof is completely dual to the proof of Theorem 3.1 □

REMARK 3.3 (see Remark 3.2). *The kernel of $\omega(x) : G_1(x) \to (G_1(x))^*$ is given by $G_0(x)$.*

The equations of an implicit generalized Hamiltonian system corresponding to Representation III and a Hamiltonian H take the form

$$\begin{aligned} \frac{\partial H}{\partial x}(x) &= \omega(x)\dot{x} + p^T(x)\lambda, \\ 0 &= p(x)\dot{x}, \end{aligned} \tag{3.20}$$

where $p(x)$ is any full rank matrix such that $\operatorname{Im} p(x) = P(x)$. A main feature of (3.20) in comparison with (3.18) is that in (3.20) the *flow constraints* $p^T(x)\dot{x} = 0$ are made explicit, while in (3.18) the algebraic constraints $g^T(x)\frac{\partial H}{\partial x}(x) = 0$ are distinguished.

EXAMPLE 3.2. *Let Q be an n-dimensional configuration manifold of a mechanical system. Classical* (*kinematic*) *constraints are given in local coordinates $q = (q_1, \dots, q_n)$ for Q as*

$$A^T(q)\dot{q} = 0 \tag{3.21}$$

with $A(q)$ an $n \times k$ matrix, $k \leq n$, with entries depending smoothly on q. We will assume that $A(q)$ has rank equal to k everywhere. The constrained Hamiltonian equations on T^*Q are classically given as (*see, e.g.,* [SM1])

$$\begin{aligned} \begin{bmatrix} \dot{q} \\ \dot{p} \end{bmatrix} &= \underbrace{\begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}}_{J} \begin{bmatrix} \frac{\partial H}{\partial q}(q,p) \\ \frac{\partial H}{\partial p}(q,p) \end{bmatrix} + \begin{bmatrix} 0 \\ A(q) \end{bmatrix} \lambda, \\ 0 &= \begin{bmatrix} 0 & A^T(q) \end{bmatrix} \begin{bmatrix} \frac{\partial H}{\partial q}(q,p) \\ \frac{\partial H}{\partial p}(q,p) \end{bmatrix}. \end{aligned} \tag{3.22}$$

Here the constraint forces $A(q)\lambda$, with $\lambda \in \mathbb{R}^k$, are uniquely determined by the requirement that the constraints (3.21) *have to be satisfied for all time. It is straightforward* *to see that an equivalent description of the equations* (3.22) *is given as follows*

$$\begin{aligned}\begin{bmatrix}\frac{\partial H}{\partial q}(q,p)\\ \frac{\partial H}{\partial p}(q,p)\end{bmatrix} &= \underbrace{\begin{bmatrix}0 & -I_n\\ I_n & 0\end{bmatrix}}_{\omega}\begin{bmatrix}\dot q\\ \dot p\end{bmatrix} + \begin{bmatrix}A(q)\\ 0\end{bmatrix}\lambda,\\ 0 &= \begin{bmatrix}A^T(q) & 0\end{bmatrix}\begin{bmatrix}\dot q\\ \dot p\end{bmatrix}.\end{aligned}\tag{3.23}$$

Let G and P be the distribution and the codistribution, respectively, on T^*Q spanned by the columns of the matrix $\left[\begin{smallmatrix}0\\ A(q)\end{smallmatrix}\right]$ and the rows of the matrix $[A^T(q) \quad 0]$, respectively. Then, since both J and ω are skew-symmetric, it follows from Theorems 3.1 *and* 3.2 *that the pairs (J, G) and (ω, P) define representation* II *and representation* III, *respectively, of the same generalized Dirac structure. We will refer to this generalized Dirac structure as $\mathcal{D}_A$.*

As the last part of this section we will now briefly show how we can directly go from representation I to a local version of representation II or III, and vice versa. This is particularly useful in analysis, where some aspects may be more easily studied in one representation, while others are easier to address in a different representation. The transformation from representation II or III to I is direct and consists of eliminating the Lagrange multipliers λ. Indeed, consider the implicit generalized Hamiltonian system (3.18) corresponding to representation II. Since $\operatorname{rank} g(x) = k$ for all $x \in \mathcal{X}$, we can locally find an $(n-k)\times n$ matrix $s(x)$ of constant rank $n-k$ such that $s(x)g(x) = 0$. Premultiplying the first n equations of (3.18) by $s(x)$ then transforms (3.18) into the following n equations:

$$\begin{bmatrix}s(x)\\ 0\end{bmatrix}\dot x = \begin{bmatrix}s(x)J(x)\\ g^T(x)\end{bmatrix}\frac{\partial H}{\partial x}(x),\tag{3.24}$$

which is easily seen to be of the form (3.10) with $F(x) = \left[\begin{smallmatrix}s(x)\\ 0\end{smallmatrix}\right]$ and $E(x) = \left[\begin{smallmatrix}s(x)J(x)\\ g^T(x)\end{smallmatrix}\right]$ satisfying (3.8), (3.9). The transformation from Representation III to I is completely similar.

Example 3.3. *Consider again the mechanical system with kinematic constraints in Example* 3.2. *Since* $\operatorname{rank} A(q) = k$ *for all $q \in Q$, we can locally find an $(n-k)\times n$ matrix $S(q)$ of constant* rank $n-k$ *such that $S(q)A(q) = 0$. Premultiplying the first $2n$ equations of* (3.22) *by the $(2n-k)\times 2n$ matrix*

$$\begin{bmatrix}I_n & 0\\ 0 & S(q)\end{bmatrix}\tag{3.25}$$

of constant rank $2n-k$ *then transforms* (3.22) *into the following $2n$ equations*:

$$\begin{bmatrix}I_n & 0\\ 0 & S(q)\\ 0 & 0\end{bmatrix}\begin{bmatrix}\dot q\\ \dot p\end{bmatrix} = \begin{bmatrix}0 & I_n\\ -S(q) & 0\\ 0 & A^T(q)\end{bmatrix}\begin{bmatrix}\frac{\partial H}{\partial q}(q,p)\\ \frac{\partial H}{\partial p}(q,p)\end{bmatrix}.\tag{3.26}$$

The transformation from representation I to II or III is more substantial. Consider representation I as given by (3.8), (3.9). Since

$$\ker\,[F(x) \ : \ -E(x)] = \operatorname{Im}\begin{bmatrix}E^T(x)\\ -F^T(x)\end{bmatrix}\tag{3.27}$$

we deduce that locally

$$G_1(x) = \operatorname{Im} E^T(x), \quad P_1(x) = \operatorname{Im} F^T(x) \tag{3.28}$$

(while $G_0(x) = \ker F(x)$, $P_0(x) = \ker E(x)$ if $F(x)$ (respectively, $E(x)$) has constant rank). In order to obtain representation II we need to assume that P_1 has constant dimension (see Theorem 3.1), or equivalently by (3.28), $F(x)$ *has constant rank.* Then we may always locally transform the equations $F(x)v = E(x)v^*$ into the form

$$\begin{bmatrix} F_1(x) \\ 0 \end{bmatrix} v = \begin{bmatrix} E_1(x) \\ E_2(x) \end{bmatrix} v^*, \tag{3.29}$$

where $F_1(x)$ has full row rank for every x in this neighborhood. Since

$$0 = E(x)F^T(x) + F(x)E^T(x) = \begin{bmatrix} E_1(x)F_1^T(x) + F_1(x)E_1^T(x) & F_1(x)E_2^T(x) \\ E_2(x)F_1^T(x) & 0 \end{bmatrix} \tag{3.30}$$

it follows that

$$E_1(x)F_1^T(x) + F_1(x)E_1^T(x) = 0 \tag{3.31}$$

and $E_2^T(x)F_1(x) = 0$, or actually since rank $[F(x) \ : \ -E(x)] = n$

$$\ker F_1(x) = \operatorname{Im} E_2^T(x). \tag{3.32}$$

By injectivity of $F_1^T(x)$ it follows that there exists an $n \times n$ matrix $J(x)$ satisfying $J(x)F_1^T(x) = -E_1^T(x)$, which is by (3.31) skew-symmetric on $\operatorname{Im} F_1^T(x)$, and extendable to a skew-symmetric matrix on $\mathbb{R}^n$. Thus the equations (3.29) can be written as

$$\begin{aligned} &v - J(x)v^* \in \ker F_1(x) = \operatorname{Im} E_2^T(x), \\ &0 = E_2(x)v^* \end{aligned} \tag{3.33}$$

or equivalently, defining the constant rank matrix $g(x) := E_2^T(x)$,

$$\begin{aligned} v &= J(x)v^* + g(x)\lambda, \\ 0 &= g^T(x)v^* \end{aligned} \tag{3.34}$$

which is representation II. Representation III can be obtained similarly by manipulating instead of $F(x)$ the constant rank matrix $E(x)$.

4. Closedness of generalized Dirac structures. The Dirac structures $\mathcal{D}$ of Definition 2.1 are called *generalized* because they do not necessarily satisfy the following *closedness* (or *integrability*) condition.

Definition 4.1 (see [D2]). *A generalized Dirac structure $\mathcal{D}$ on $\mathcal{X}$ is called closed* (*or simply a Dirac structure*) *if for arbitrary* (X_1, α_1), (X_2, α_2), *and* $(X_3, \alpha_3) \in \mathcal{D}$ *there holds*

$$\langle L_{X_1}\alpha_2 \,|\, X_3\rangle + \langle L_{X_2}\alpha_3 \,|\, X_1\rangle + \langle L_{X_3}\alpha_1 \,|\, X_2\rangle = 0. \tag{4.1}$$

The following theorem gives a very useful characterization of closedness of a generalized Dirac structure.

Theorem 4.1 (cf. [D2, Theorem 2.1]; see also [C1]). *$\mathcal{D}$ is closed if and only if*

$$([X_1, X_2], \mathbf{i}_{X_1}\mathbf{d}\alpha_2 - \mathbf{i}_{X_2}\mathbf{d}\alpha_1 + \mathbf{d}\langle \alpha_2 \,|\, X_1\rangle) \in \mathcal{D} \quad \text{for all } (X_1, \alpha_1),\ (X_2, \alpha_2) \in \mathcal{D}. \tag{4.2}$$

Proof. First note that the identities (see, e.g., [AMR])

$$L_X\alpha = \mathbf{d}\mathbf{i}_X\alpha + \mathbf{i}_X\mathbf{d}\alpha, \tag{4.3}$$

$$\mathbf{i}_{[X,Y]}\alpha = L_X\mathbf{i}_Y\alpha - \mathbf{i}_Y L_X\alpha \tag{4.4}$$

are satisfied for all vector fields X, Y and k-forms α on $\mathcal{X}$. (The formula (4.3) is also known as *Cartan's magic formula.*) Hence,

$$\langle L_X\alpha \,|\, Y\rangle = \langle \mathbf{d}\langle \alpha \,|\, X\rangle \,|\, Y\rangle + \mathbf{d}\alpha(X,Y), \tag{4.5}$$

$$\langle \alpha \,|\, [X,Y]\rangle = -\langle \mathbf{d}\langle \alpha \,|\, X\rangle \,|\, Y\rangle + \langle L_Y\alpha \,|\, X\rangle \tag{4.6}$$

for all vector fields X, Y and one-forms α on $\mathcal{X}$.

Now take arbitrary $(X_1,\alpha_1), (X_2,\alpha_2), (X_3,\alpha_3) \in \mathcal{D}$. Then

$$\begin{aligned}
&\langle \mathbf{i}_{X_1}\mathbf{d}\alpha_2 - \mathbf{i}_{X_2}\mathbf{d}\alpha_1 + \mathbf{d}\langle \alpha_2 \,|\, X_1\rangle \,|\, X_3\rangle + \langle \alpha_3 \,|\, [X_1,X_2]\rangle \\
&\quad = \langle \mathbf{i}_{X_1}\mathbf{d}\alpha_2 \,|\, X_3\rangle - \langle \mathbf{i}_{X_2}\mathbf{d}\alpha_1 \,|\, X_3\rangle + \langle \mathbf{d}\langle \alpha_2 \,|\, X_1\rangle \,|\, X_3\rangle + \langle \alpha_3 \,|\, [X_1,X_2]\rangle \\
&\quad = \mathbf{d}\alpha_2(X_1,X_3) + \langle \mathbf{d}\langle \alpha_2 \,|\, X_1\rangle \,|\, X_3\rangle - \mathbf{d}\alpha_1(X_2,X_3) + \langle \alpha_3 \,|\, [X_1,X_2]\rangle \\
&\quad = \langle L_{X_1}\alpha_2 \,|\, X_3\rangle + \mathbf{d}\alpha_1(X_3,X_2) + \langle L_{X_2}\alpha_3 \,|\, X_1\rangle - \langle \mathbf{d}\langle \alpha_3 \,|\, X_1\rangle \,|\, X_2\rangle \\
&\quad = \langle L_{X_1}\alpha_2 \,|\, X_3\rangle + \langle L_{X_2}\alpha_3 \,|\, X_1\rangle + \langle L_{X_3}\alpha_1 \,|\, X_2\rangle \qquad (4.7)
\end{aligned}$$

since $\mathbf{d}\alpha_1(X_2,X_3) = -\mathbf{d}\alpha_1(X_3,X_2)$, and $\mathbf{d}\langle \alpha_3 \,|\, X_1\rangle + \mathbf{d}\langle \alpha_1 \,|\, X_3\rangle = 0$ because (X_1,α_1), $(X_3,\alpha_3) \in \mathcal{D}$. Thus,

$$\begin{gathered}
\mathcal{D} \text{ is closed} \\
\Updownarrow \\
\langle L_{X_1}\alpha_2 \,|\, X_3\rangle + \langle L_{X_2}\alpha_3 \,|\, X_1\rangle + \langle L_{X_3}\alpha_1 \,|\, X_2\rangle = 0 \\
\text{for all } (X_1,\alpha_1), (X_2,\alpha_2), (X_3,\alpha_3) \in \mathcal{D} \\
\Updownarrow \\
\langle \mathbf{i}_{X_1}\mathbf{d}\alpha_2 - \mathbf{i}_{X_2}\mathbf{d}\alpha_1 + \mathbf{d}\langle \alpha_2 \,|\, X_1\rangle \,|\, X_3\rangle + \langle \alpha_3 \,|\, [X_1,X_2]\rangle = 0 \\
\text{for all } (X_1,\alpha_1), (X_2,\alpha_2), (X_3,\alpha_3) \in \mathcal{D} \\
\Updownarrow \\
([X_1,X_2], \mathbf{i}_{X_1}\mathbf{d}\alpha_2 - \mathbf{i}_{X_2}\mathbf{d}\alpha_1 + \mathbf{d}\langle \alpha_2 \,|\, X_1\rangle) \in \mathcal{D} \\
\text{for all } (X_1,\alpha_1), (X_2,\alpha_2) \in \mathcal{D},
\end{gathered}$$

where the last equivalence follows from the fact that $\mathcal{D} = \mathcal{D}^{\perp}$. □

Remark 4.1. *Courant* [C1] *uses property* (4.2) *as the definition of closedness* (*or integrability*) *of a generalized Dirac structure.*

Closedness needs only to be checked on a set of pairs (X_i,α_i) which span the generalized Dirac structure $\mathcal{D}$, as follows from the following lemma.

Lemma 4.1. *Consider a generalized Dirac structure $\mathcal{D}$ on a manifold $\mathcal{X}$. Let*

$$(X_1,\alpha_1), \dots, (X_n,\alpha_n) \in \mathcal{D}$$

and suppose that

$$\big([X_i,X_j], \mathbf{i}_{X_i}\mathbf{d}\alpha_j - \mathbf{i}_{X_j}\mathbf{d}\alpha_i + \mathbf{d}\langle \alpha_j \,|\, X_i\rangle\big) \in \mathcal{D}, \quad i,j = 1,\dots,n. \tag{4.8}$$

Then also $([X,Y], \mathbf{i}_X\mathbf{d}\beta - \mathbf{i}_Y\mathbf{d}\alpha + \mathbf{d}\langle \beta \,|\, X\rangle) \in \mathcal{D}$, *where*

$$(X,\alpha) = \sum_{i=1}^{n}\zeta_i(X_i,\alpha_i), \quad (Y,\beta) = \sum_{i=1}^{n}\eta_i(X_i,\alpha_i) \tag{4.9}$$

for arbitrary $\zeta_i, \eta_i \in C^\infty(\mathcal{X})$, $i = 1,\dots,n$.

Proof. Let $\gamma = \mathbf{i}_X\mathbf{d}\beta - \mathbf{i}_Y\mathbf{d}\alpha + \mathbf{d}\langle\beta \,|\, X\rangle$. A straightforward calculation then gives

$$[X, Y] = \sum_{i,j=1}^{n} [\zeta_i X_i(\eta_j)X_j + \zeta_i\eta_j[X_i, X_j] - \eta_j X_j(\zeta_i)X_i], \tag{4.10}$$

$$\gamma = \sum_{i,j=1}^{n} [\zeta_i X_i(\eta_j)\alpha_j + \zeta_i\eta_j(\mathbf{i}_{X_i}\mathbf{d}\alpha_j - \mathbf{i}_{X_j}\mathbf{d}\alpha_i + \mathbf{d}\langle\alpha_j \,|\, X_i\rangle) - \eta_j X_j(\zeta_i)\alpha_i]. \tag{4.11}$$

Thus, from (4.8) it follows that $([X, Y], \gamma) \in \mathcal{D}$. □

A smooth function $H \in C^\infty(\mathcal{X})$ is said to be *admissible* (see [C1]) if there exists a (smooth) vector field X such that $(X, \mathbf{d}H) \in \mathcal{D}$. From the definition of the codistribution P_1 in (3.2) we see that the space of all admissible functions is given by

$$\mathcal{A}_\mathcal{D} = \{H \in C^\infty(\mathcal{X}) \,|\, \mathbf{d}H \in P_1\}. \tag{4.12}$$

There is a well-defined generalized Poisson bracket on $\mathcal{A}_\mathcal{D}$ given by the formula

$$\{H_1, H_2\}_\mathcal{D} = \langle\mathbf{d}H_1 \,|\, X_2\rangle = -\langle\mathbf{d}H_2 \,|\, X_1\rangle, \tag{4.13}$$

where $(X_1, \mathbf{d}H_1), (X_2, \mathbf{d}H_2) \in \mathcal{D}$. To show that $\{\,,\,\}_\mathcal{D}$ as defined in (4.13) is a generalized Poisson bracket is straightforward. Bilinearity of $\{\,,\,\}_\mathcal{D}$ follows from bilinearity of $\langle\ |\ \rangle$. Skew-symmetry is a consequence of (4.13). Finally, take arbitrary $(X_1, \mathbf{d}H_1), (X_2, \mathbf{d}H_2), (X_3, \mathbf{d}H_3) \in \mathcal{D}$. Then

$$\begin{aligned}\{H_1, H_2H_3\}_D &= -\langle\mathbf{d}(H_2H_3) \,|\, X_1\rangle = -\langle H_3\mathbf{d}H_2 + H_2\mathbf{d}H_3 \,|\, X_1\rangle \\ &= H_3\{H_1, H_2\}_\mathcal{D} + H_2\{H_1, H_3\}_\mathcal{D}\end{aligned} \tag{4.14}$$

so $\{\,,\,\}_\mathcal{D}$ also satisfies the Leibniz identity. For a Dirac structure given by representation II (see Theorem 3.1), $\{\,,\,\}_\mathcal{D}$ is given as follows:

$$\{H_1, H_2\}_\mathcal{D}(x) = \left[\frac{\partial H_1}{\partial x}(x)\right]^T J(x)\, \frac{\partial H_2}{\partial x}(x), \quad H_1, H_2 \in \mathcal{A}_\mathcal{D}. \tag{4.15}$$

We will now characterize closedness of a generalized Dirac structure $\mathcal{D}$ in terms of the bracket $\{\,,\,\}_\mathcal{D}$ and the admissible functions $\mathcal{A}_\mathcal{D}$. The following necessary conditions for closedness follow from Theorem 4.1.

Corollary 4.1 (cf. [C1, D2]). *If $\mathcal{D}$ is closed, then*

1. *G_0 and G_1 are involutive distributions*;
2. $\{H_1, H_2\}_D \in \mathcal{A}_\mathcal{D}$;
3. $\{H_1, \{H_2, H_3\}_\mathcal{D}\}_D + \{H_2, \{H_3, H_1\}_\mathcal{D}\}_D + \{H_3, \{H_1, H_2\}_\mathcal{D}\}_D = 0$

for all $H_1, H_2, H_3 \in \mathcal{A}_\mathcal{D}$.

Proof.

1. Let $X_1, X_2 \in G_0$, i.e., $(X_1, 0), (X_2, 0) \in \mathcal{D}$. Then by Theorem 4.1 $([X_1, X_2], 0) \in \mathcal{D}$, which means that $[X_1, X_2] \in G_0$. Involutivity of G_1 also follows directly from Theorem 4.1.
2. Take $H_1, H_2 \in \mathcal{A}_\mathcal{D}$ so that $(X_1, \mathbf{d}H_1), (X_2, \mathbf{d}H_2) \in \mathcal{D}$. Then we have $([X_1, X_2], \mathbf{d}\langle\mathbf{d}H_2 \,|\, X_1\rangle) \in \mathcal{D}$, which means that

$$\mathbf{d}\langle\mathbf{d}H_2 \,|\, X_1\rangle = \mathbf{d}\{H_2, H_1\}_\mathcal{D} \in P_1 \;\; \Rightarrow \;\; \{H_1, H_2\}_\mathcal{D} \in \mathcal{A}_\mathcal{D}. \tag{4.16}$$

3. Take $H_1, H_2, H_3 \in \mathcal{A}_{\mathcal{D}}$ so that $(X_1, \mathbf{d}H_1), (X_2, \mathbf{d}H_2), (X_3, \mathbf{d}H_3) \in \mathcal{D}$. Then

$$\begin{aligned} 0 &= \langle L_{X_1}\mathbf{d}H_2 \,|\, X_3\rangle + \langle L_{X_2}\mathbf{d}H_3 \,|\, X_1\rangle + \langle L_{X_3}\mathbf{d}H_1 \,|\, X_2\rangle \\ &= \langle \mathbf{d}\langle \mathbf{d}H_2 \,|\, X_1\rangle \,|\, X_3\rangle + \langle \mathbf{d}\langle \mathbf{d}H_3 \,|\, X_2\rangle \,|\, X_1\rangle + \langle \mathbf{d}\langle \mathbf{d}H_1 \,|\, X_3\rangle \,|\, X_2\rangle \\ &= \langle \mathbf{d}\{H_2, H_1\}_{\mathcal{D}} \,|\, X_3\rangle + \langle \mathbf{d}\{H_3, H_2\}_{\mathcal{D}} \,|\, X_1\rangle + \langle \mathbf{d}\{H_1, H_3\}_{\mathcal{D}} \,|\, X_2\rangle \qquad (4.17) \\ &= \{\{H_2, H_1\}_{\mathcal{D}}, H_3\}_{\mathcal{D}} + \{\{H_3, H_2\}_{\mathcal{D}}, H_1\}_{\mathcal{D}} + \{\{H_1, H_3\}_{\mathcal{D}}, H_2\}_{\mathcal{D}} \\ &= \{H_1, \{H_2, H_3\}_{\mathcal{D}}\}_{\mathcal{D}} + \{H_2, \{H_3, H_1\}_{\mathcal{D}}\}_{\mathcal{D}} + \{H_3, \{H_1, H_2\}_{\mathcal{D}}\}_{\mathcal{D}}. \qquad \square \end{aligned}$$

If in addition the codistribution P_1 (see (3.2)) is constant-dimensional, the following theorem gives necessary *and* sufficient conditions for closedness in terms of $\{\,,\,\}_{\mathcal{D}}$ and $\mathcal{A}_{\mathcal{D}}$.

THEOREM 4.2. *Consider a generalized Dirac structure $\mathcal{D}$ on a manifold $\mathcal{X}$. Let P_1 denote the codistribution on $\mathcal{X}$ defined by* (3.2). *Assume that P_1 is constant-dimensional. Then $\mathcal{D}$ is closed if and only if the following three conditions are satisfied*:

1. $G_0 = \ker P_1$ *is involutive*;
2. $\{H_1, H_2\}_{\mathcal{D}} \in \mathcal{A}_{\mathcal{D}}$;
3. $\{H_1, \{H_2, H_3\}_{\mathcal{D}}\}_D + \{H_2, \{H_3, H_1\}_{\mathcal{D}}\}_D + \{H_3, \{H_1, H_2\}_{\mathcal{D}}\}_D = 0$

for all $H_1, H_2, H_3 \in \mathcal{A}_{\mathcal{D}}$.

Proof. The necessity of these three conditions follows from Corollary 4.1 so we have to show only the sufficiency part here. First note that by using Proposition 3.1 we have that $P_1 = \operatorname{ann} G_0$. Since $G_0 = \ker P_1$ is involutive and P_1 is constant-dimensional, by Frobenius's theorem in a neighborhood of any point $x_0 \in \mathcal{X}$ there exist local coordinates $x = (x_1, \dots, x_n)$ such that

$$P_1 = \operatorname{ann} G_0 = \operatorname{span}\{\mathbf{d}x_1, \dots, \mathbf{d}x_{n-m}\}, \qquad (4.18)$$

where $m = \dim \ker P_1$ $(= \dim G_0)$. In the following, every computation is done in such a neighborhood using local coordinates.

Take now arbitrary $(X_1, \alpha_1), (X_2, \alpha_2) \in \mathcal{D}$. Then, since $\alpha_1, \alpha_2 \in P_1$, we have that

$$\alpha_1 = \sum_{i=1}^{n-m} \zeta_i \mathbf{d}x_i, \qquad (4.19)$$

$$\alpha_2 = \sum_{i=1}^{n-m} \eta_i \mathbf{d}x_i, \qquad (4.20)$$

where ζ_i, η_i are smooth functions. Now, let the vector fields $Y_1, \dots, Y_{n-m}$ be such that

$$(Y_i, \mathbf{d}x_i) \in \mathcal{D}, \quad 1 \le i \le n-m. \qquad (4.21)$$

Since also $(X_1, \alpha_1) \in \mathcal{D}$ it follows that

$$\langle \mathbf{d}x_k \,|\, X_1\rangle + \left\langle \sum_{i=1}^{n-m} \zeta_i \mathbf{d}x_i \,\middle|\, Y_k \right\rangle = 0 \qquad (4.22)$$

so

$$\langle \mathbf{d}x_k \,|\, X_1\rangle = -\sum_{i=1}^{n-m} \zeta_i \{x_i, x_k\}_{\mathcal{D}} \qquad (4.23)$$

for $1 \leq k \leq n-m$. Define the vector fields Z_1, Z_2 as

$$Z_1 = X_1 - \sum_{i=1}^{n-m} \zeta_i Y_i, \tag{4.24}$$

$$Z_2 = X_2 - \sum_{i=1}^{n-m} \eta_i Y_i. \tag{4.25}$$

Then

$$\begin{aligned}
\langle \mathbf{d}x_k \,|\, Z_1 \rangle &= \left\langle \mathbf{d}x_k \middle| X_1 - \sum_{i=1}^{n-m} \zeta_i Y_i \right\rangle \\
&= -\sum_{i=1}^{n-m} \zeta_i \{x_i, x_k\}_{\mathcal{D}} - \sum_{i=1}^{n-m} \zeta_i \{x_k, x_i\}_{\mathcal{D}} \\
&= 0 \qquad (4.26)
\end{aligned}$$

for all $1 \leq k \leq n-m$ since $\{x_i, x_k\}_{\mathcal{D}} = -\{x_k, x_i\}_{\mathcal{D}}$. This means that $Z_1 \in \ker P_1 = G_0$ and

$$X_1 = \sum_{i=1}^{n-m} \zeta_i Y_i + Z_1, \tag{4.27}$$

$$X_2 = \sum_{i=1}^{n-m} \eta_i Y_i + Z_2, \tag{4.28}$$

where $Z_1, Z_2 \in G_0$.

Now we want to calculate the term

$$\alpha_{12} = \mathbf{i}_{X_1} \mathbf{d}\alpha_2 - \mathbf{i}_{X_2} \mathbf{d}\alpha_1 + \mathbf{d}\langle \alpha_2 \,|\, X_1 \rangle. \tag{4.29}$$

We have $\mathbf{d}\alpha_2 = \mathbf{d}(\sum_{i=1}^{n-m} \eta_i \mathbf{d}x_i) = \sum_{i=1}^{n-m} \mathbf{d}\eta_i \wedge \mathbf{d}x_i$, so

$$\begin{aligned}
\mathbf{i}_{X_1} \mathbf{d}\alpha_2 &= \mathbf{i}_{X_1} \left(\sum_{i=1}^{n-m} \mathbf{d}\eta_i \wedge \mathbf{d}x_i \right) \\
&= \sum_{i=1}^{n-m} \left[\mathbf{i}_{X_1} \mathbf{d}\eta_i \wedge \mathbf{d}x_i - \mathbf{d}\eta_i \wedge \mathbf{i}_{X_1} \mathbf{d}x_i \right] \\
&= \sum_{i=1}^{n-m} \left[\left\langle \mathbf{d}\eta_i \middle| \sum_{j=1}^{n-m} \zeta_j Y_j + Z_1 \right\rangle \mathbf{d}x_i - \left\langle \mathbf{d}x_i \middle| \sum_{j=1}^{n-m} \zeta_j Y_j + Z_1 \right\rangle \mathbf{d}\eta_i \right] \\
&= \sum_{i,j=1}^{n-m} \left[\zeta_j Y_j(\eta_i) \mathbf{d}x_i - \zeta_j \{x_i, x_j\}_D \mathbf{d}\eta_i \right] + \sum_{i=1}^{n-m} Z_1(\eta_i) \mathbf{d}x_i, \qquad (4.30)
\end{aligned}$$

where we used the fact that $\langle \mathbf{d}x_i \,|\, Z_1 \rangle = 0$ since $\mathbf{d}x_i \in \operatorname{ann} G_0$. Similarly we obtain

$$\mathbf{i}_{X_2} \mathbf{d}\alpha_1 = \sum_{i,j=1}^{n-m} [\eta_j Y_j(\zeta_i) \mathbf{d}x_i - \eta_j \{x_i, x_j\}_{\mathcal{D}} \mathbf{d}\zeta_i] + \sum_{i=1}^{n-m} Z_2(\zeta_i) \mathbf{d}x_i. \tag{4.31}$$

Moreover,

$$\begin{aligned}\mathbf{d}\langle \alpha_2 \,|\, X_1\rangle &= \mathbf{d}\left\langle \sum_{i=1}^{n-m} \eta_i \mathbf{d}x_i \,\middle|\, \sum_{j=1}^{n-m} \zeta_j Y_j + Z_1 \right\rangle \\ &= \mathbf{d}\left[\sum_{i,j=1}^{n-m} \eta_i \zeta_j \left\langle \mathbf{d}x_i \,|\, Y_j \right\rangle\right] \\ &= \mathbf{d}\left[\sum_{i,j=1}^{n-m} \eta_i \zeta_j \{x_i, x_j\}_{\mathcal{D}}\right] \\ &= \sum_{i,j=1}^{n-m} [\eta_i \zeta_j \mathbf{d}\{x_i, x_j\}_{\mathcal{D}} + \{x_i, x_j\}_{\mathcal{D}}(\zeta_j \mathbf{d}\eta_i + \eta_i \mathbf{d}\zeta_j)] \\ &= \sum_{i,j=1}^{n-m} [\eta_i \zeta_j \mathbf{d}\{x_i, x_j\}_{\mathcal{D}} + \{x_i, x_j\}_{\mathcal{D}}(\zeta_j \mathbf{d}\eta_i - \eta_j \mathbf{d}\zeta_i)], \qquad (4.32)\end{aligned}$$

where the last equation follows from skew-symmetry of $\{\,,\,\}_{\mathcal{D}}$. Inserting (4.30), (4.31), and (4.32) in (4.29) gives

$$\begin{aligned}\alpha_{12} &= \mathbf{i}_{X_1}\mathbf{d}\alpha_2 - \mathbf{i}_{X_2}\mathbf{d}\alpha_1 + \mathbf{d}\langle \alpha_2 \,|\, X_1\rangle \\ &= \sum_{i,j=1}^{n-m} [(\zeta_j Y_j(\eta_i) - \eta_j Y_j(\zeta_i))\mathbf{d}x_i + \eta_i \zeta_j \mathbf{d}\{x_i, x_j\}_{\mathcal{D}}] \\ &\quad + \sum_{i=1}^{n-m} (Z_1(\eta_i) - Z_2(\zeta_i))\mathbf{d}x_i. \qquad (4.33)\end{aligned}$$

From (4.33) we immediately see that $\alpha_{12} \in P_1$ since $\mathbf{d}\{x_i, x_j\}_{\mathcal{D}} \in P_1$ when $1 \le i, j \le n - m$.

Now we have to take a closer look at the term $[X_1, X_2]$. A direct calculation yields

$$\begin{aligned}[X_1, X_2] &= \left[\sum_{i=1}^{n-m} \zeta_i Y_i + Z_1, \sum_{j=1}^{n-m} \eta_j Y_j + Z_2\right] \qquad (4.34) \\ &= \sum_{i,j=1}^{n-m} \{(\zeta_j Y_j(\eta_i) - \eta_j Y_j(\zeta_i))Y_i + \eta_i \zeta_j [Y_j, Y_i]\} + \sum_{i=1}^{n-m} (Z_1(\eta_i) - Z_2(\zeta_i))Y_i + Z_{12},\end{aligned}$$

where the vector field Z_{12} is given by

$$Z_{12} = \sum_{i=1}^{n-m} (\zeta_i [Y_i, Z_2] - \eta_i [Y_i, Z_1]) + [Z_1, Z_2]. \qquad (4.35)$$

Now take arbitrary $Z \in G_0$ and consider $\mathbf{d}\{x_i, x_j\}_{\mathcal{D}} \in \operatorname{ann} G_0$ for $1 \le i, j \le n - m$.

Then

$$\begin{aligned} 0 &= \langle \mathbf{d}\{x_i, x_j\}_{\mathcal{D}} \,|\, Z\rangle \\ &= -Z(\{x_j, x_i\}_D) \\ &= -Z(Y_i(x_j)) \\ &= -Y_i(Z(x_j)) + [Y_i, Z](x_j) \\ &= \langle \mathbf{d}x_j \,|\, [Y_i, Z]\rangle \end{aligned} \tag{4.36}$$

for all $1 \le i, j \le n-m$, which means that $[Y_i, Z] \in \ker P_1 = G_0$ for all $1 \le i \le n-m$. Since G_0 is involutive we immediately see from (4.35) that also $Z_{12} \in G_0$.

Now we want to show that $([Y_j, Y_i], \mathbf{d}\{x_i, x_j\}_{\mathcal{D}}) \in \mathcal{D}$. We know that $\{x_i, x_j\}_{\mathcal{D}} \in \mathcal{A}_{\mathcal{D}}$, which means that there exist vector fields Y_{ij} such that $(Y_{ij}, \mathbf{d}\{x_i, x_j\}_{\mathcal{D}}) \in \mathcal{D}$, $i, j = 1, \ldots, n-m$. Then

$$\begin{aligned} \langle \mathbf{d}x_k \,|\, [Y_j, Y_i] - Y_{ij}\rangle &= [Y_j, Y_i](x_k) - Y_{ij}(x_k) \\ &= Y_j(\langle \mathbf{d}x_k \,|\, Y_i\rangle) - Y_i(\langle \mathbf{d}x_k \,|\, Y_j\rangle) - \langle \mathbf{d}x_k \,|\, Y_{ij}\rangle \\ &= \langle \mathbf{d}\{x_k, x_i\}_{\mathcal{D}} \,|\, Y_j\rangle - \langle \mathbf{d}\{x_k, x_j\}_{\mathcal{D}} \,|\, Y_i\rangle - \{x_k, \{x_i, x_j\}_{\mathcal{D}}\}_{\mathcal{D}} \\ &= \{\{x_k, x_i\}_{\mathcal{D}}, x_j\}_{\mathcal{D}} - \{\{x_k, x_j\}_{\mathcal{D}}, x_i\}_{\mathcal{D}} - \{x_k, \{x_i, x_j\}_{\mathcal{D}}\}_{\mathcal{D}} \\ &= \{x_j, \{x_i, x_k\}_{\mathcal{D}}\}_{\mathcal{D}} + \{x_i, \{x_k, x_j\}_{\mathcal{D}}\}_{\mathcal{D}} + \{x_k, \{x_j, x_i\}_{\mathcal{D}}\}_{\mathcal{D}} \\ &= 0 \end{aligned} \tag{4.37}$$

when $1 \le i, j, k \le n-m$, which means that $[Y_j, Y_i] - Y_{ij} \in \ker P_1 = G_0$. Thus,

$$([Y_j, Y_i], \mathbf{d}\{x_i, x_j\}) \in \mathcal{D}, \quad i, j = 1, \ldots, n-m, \tag{4.38}$$

and by inspection of (4.33) and (4.34) we see that $([X_1, X_2], \alpha_{12}) \in \mathcal{D}$, and closedness of $\mathcal{D}$ follows from Theorem 4.1. □

In the following we will explicitly characterize closedness in the three different *representations* of a Dirac structure.

THEOREM 4.3 (representation I). *Consider a generalized Dirac structure $\mathcal{D}$ on a manifold $\mathcal{X}$ given locally in representation* I (*see* (3.8), (3.9)). *Define $(X_i, \alpha_i) \in \mathcal{D}$ in local coordinates by*

$$X_i = E_i^T(x), \tag{4.39}$$

$$\alpha_i = -F_i^T(x), \tag{4.40}$$

*where $E_i^T(x)$ and $F_i^T(x)$ denote the i*th *column of the matrices $E^T(x)$ and $F^T(x)$, respectively. Then $\mathcal{D}$ is closed if and only if*

$$\big([X_i, X_j], \mathbf{i}_{X_i}\mathbf{d}\alpha_j - \mathbf{i}_{X_j}\mathbf{d}\alpha_i + \mathbf{d}\langle \alpha_j \,|\, X_i\rangle\big) \in \mathcal{D}(x) \tag{4.41}$$

for all $x \in \mathcal{X}$, $i, j = 1, \ldots, n$.

Proof. The proof follows from (3.27), Theorem 4.1, and Lemma 4.1. □

THEOREM 4.4 (representation II). *Let $\mathcal{X}$ be an n-dimensional manifold. Let G be a constant-dimensional distribution on $\mathcal{X}$, and $J(x) : T_x^*\mathcal{X} \to T_x\mathcal{X}$, $x \in \mathcal{X}$, be a skew-symmetric vector bundle map. Moreover, let $\{\,,\,\}$ denote the generalized Poisson bracket corresponding to J. Then the generalized Dirac structure given by* (*see Theorem* 3.1)

$$\mathcal{D} = \{(X, \alpha) \in T\mathcal{X} \oplus T^*\mathcal{X} \,|\, X(x) - J(x)\alpha(x) \in G(x), x \in \mathcal{X}, \alpha \in \text{ann}\, G\} \tag{4.42}$$

is closed if and only if

1. *G is involutive*;
2. $\{H_1, H_2\} \in \mathcal{A}_\mathcal{D}$;
3. $\{H_1, \{H_2, H_3\}\} + \{H_2, \{H_3, H_1\}\} + \{H_3, \{H_1, H_2\}\} = 0$

for all $H_1, H_2, H_3 \in \mathcal{A}_\mathcal{D} = \{H \in C^\infty(\mathcal{X}) \mid \mathbf{d}H \in \operatorname{ann} G\}$.

Proof. The result follows from Theorem 4.2 using the facts that $G_0 = G$ and that $\{H_1, H_2\}_\mathcal{D} = \{H_1, H_2\}$ for all $H_1, H_2 \in \mathcal{A}_\mathcal{D}$. □

THEOREM 4.5 (representation III). *Let $\mathcal{X}$ be an n-dimensional manifold. Let P be a constant-dimensional codistribution on $\mathcal{X}$, and $\omega(x) : T_x\mathcal{X} \to T_x^*\mathcal{X}$, $x \in \mathcal{X}$, be a skew-symmetric vector bundle map. Then the generalized Dirac structure given by* (*see Theorem* 3.2)

$$\mathcal{D} = \{(X, \alpha) \in T\mathcal{X} \oplus T^*\mathcal{X} \mid \alpha(x) - \omega(x)X(x) \in P(x), x \in \mathcal{X}, X \in \ker\, P\} \tag{4.43}$$

is closed if and only if

1. *$\mathit{ker} P$ is involutive*;
2. $\mathbf{d}\omega(X_1, X_2, X_3) = 0$ *for all* $X_1, X_2, X_3 \in \mathit{ker} P$.

Proof. Let $(X_1, \alpha_1), (X_2, \alpha_2) \in \mathcal{D}$, i.e.,

$$\alpha_i = \mathbf{i}_{X_i}\omega + p_i, \quad p_i \in P,\ X_i \in \ker P, \quad i = 1, 2. \tag{4.44}$$

Define as in (4.29) the one-form $\alpha_{12} = \mathbf{i}_{X_1}\mathbf{d}\alpha_2 - \mathbf{i}_{X_2}\mathbf{d}\alpha_1 + \mathbf{d}\langle\alpha_2 \mid X_1\rangle$. Now, using Cartan's magic formula, we get

$$\mathbf{d}\mathbf{i}_{X_1}\omega = L_{X_1}\omega - \mathbf{i}_{X_1}\mathbf{d}\omega, \tag{4.45}$$

$$\mathbf{d}\mathbf{i}_{X_1}\mathbf{i}_{X_2}\omega = L_{X_1}\mathbf{i}_{X_2}\omega - \mathbf{i}_{X_1}L_{X_2}\omega + \mathbf{i}_{X_1}\mathbf{i}_{X_2}\mathbf{d}\omega \tag{4.46}$$

for all vector fields X_1, X_2 on $\mathcal{X}$. Hence

$$\begin{aligned} \alpha_{12} &= \mathbf{i}_{X_1}\mathbf{d}\alpha_2 - \mathbf{i}_{X_2}\mathbf{d}\alpha_1 + \mathbf{d}\langle\alpha_2 \mid X_1\rangle \\ &= \mathbf{i}_{X_1}\mathbf{d}(\mathbf{i}_{X_2}\omega + p_2) - \mathbf{i}_{X_2}\mathbf{d}(\mathbf{i}_{X_1}\omega + p_1) + \mathbf{d}\mathbf{i}_{X_1}(\mathbf{i}_{X_2}\omega + p_2) \\ &= \mathbf{i}_{X_1}\mathbf{d}\mathbf{i}_{X_2}\omega + \mathbf{i}_{X_1}\mathbf{d}p_2 - \mathbf{i}_{X_2}\mathbf{d}\mathbf{i}_{X_1}\omega - \mathbf{i}_{X_2}\mathbf{d}p_1 + \mathbf{d}\mathbf{i}_{X_1}\mathbf{i}_{X_2}\omega + \mathbf{d}\mathbf{i}_{X_1}p_2 \\ &= -\mathbf{i}_{X_2}L_{X_1}\omega + L_{X_1}\mathbf{i}_{X_2}\omega + \mathbf{i}_{X_1}\mathbf{d}p_2 - \mathbf{i}_{X_2}\mathbf{d}p_1 + \mathbf{i}_{X_2}\mathbf{i}_{X_1}\mathbf{d}\omega \\ &= \mathbf{i}_{[X_1,X_2]}\omega + \mathbf{i}_{X_1}\mathbf{d}p_2 - \mathbf{i}_{X_2}\mathbf{d}p_1 + \mathbf{i}_{X_2}\mathbf{i}_{X_1}\mathbf{d}\omega \qquad (4.47) \end{aligned}$$

since $\mathbf{i}_{[X_1,X_2]}\omega = L_{X_1}\mathbf{i}_{X_2}\omega - \mathbf{i}_{X_2}L_{X_1}\omega$. Thus, using Theorem 4.1 and the definition of $\mathcal{D}$, we have that

$$\begin{gathered} \mathcal{D} \text{ is closed} \\ \Updownarrow \\ ([X_1, X_2], \mathbf{i}_{[X_1,X_2]}\omega + \mathbf{i}_{X_1}\mathbf{d}p_2 - \mathbf{i}_{X_2}\mathbf{d}p_1 + \mathbf{i}_{X_2}\mathbf{i}_{X_1}\mathbf{d}\omega) \in \mathcal{D} \\ \text{for all } p_1, p_2 \in P, \text{ for all } X_1, X_2 \in \ker P. \\ \Updownarrow \\ \left.\begin{array}{l} [X_1, X_2] \in \ker P \\ \mathbf{i}_{X_1}\mathbf{d}p_2 - \mathbf{i}_{X_2}\mathbf{d}p_1 + \mathbf{i}_{X_2}\mathbf{i}_{X_1}\mathbf{d}\omega \in P \end{array}\right\} \text{ for all } p_1, p_2 \in P, \text{ for all } X_1, X_2 \in \ker P. \end{gathered}$$

Now, if P is a constant-dimensional codistribution and $\ker P$ is involutive, it follows that for every $p \in P$ there exists $\bar{p} \in P$ and a one-form η such that $\mathbf{d}p = \eta \wedge \bar{p}$. Thus, $\mathbf{i}_X\mathbf{d}p = \eta(X)\bar{p} \in P$ for all $X \in \ker P$. Moreover, $\mathbf{i}_{X_2}\mathbf{i}_{X_1}\mathbf{d}\omega(X_3) = \mathbf{d}\omega(X_1, X_2, X_3)$ which means that $\mathbf{i}_{X_2}\mathbf{i}_{X_1}\mathbf{d}\omega \in P$ if and only if $\mathbf{d}\omega(X_1, X_2, X_3) = 0$ for all $X_3 \in \ker\, P$ since P is constant-dimensional. □

Remark 4.2. *In* [C1] *it is shown that closedness of D implies condition* 2 *in Theorem* 4.5.

We will now apply the above theory to mechanical systems with kinematic constraints (see Example 3.2).

Proposition 4.1. *Consider the mechanical system with kinematic constraints $A^T(q)\dot{q} = 0$ as given in Example* 3.2. *Let $\{\, , \,\}$ denote the Poisson bracket defined (locally) by the structure matrix J. Then the following statements are equivalent*:

1. *$\mathcal{D}_A$ is closed*;
2. *the constraints* (3.21) *are holonomic*;
3. *$\mathbf{d}\{H_1, H_2\} \in \operatorname{ann} G$ for all H_1, H_2 such that $\mathbf{d}H_1, \mathbf{d}H_2 \in \operatorname{ann} G$.*

Proof. $1 \Leftrightarrow 2$: From Theorem 4.5 it follows that $\mathcal{D}_A$ is closed if and only if $\ker P$ is involutive which is equivalent to the constraints (3.21) being holonomic. $1 \Leftrightarrow 3$: This follows from Theorem 4.4 since G is involutive and $\{\, , \,\}$ satisfies the Jacobi identity in this case. □

The next proposition gives an interesting interpretation of closedness of generalized Dirac structures that come up in connection with Lie-Poisson structures.

Proposition 4.2. *Let G be any n-dimensional Lie group* (*e.g.*, $SE(3)$), *with Lie algebra g, and the dual Lie algebra g^* with the Lie-Poisson bracket $\{\, , \,\}$. Consider a constant distribution on g^*, that is a linear subspace $\mathcal{V} \subset g^*$. Define the Dirac structure $\mathcal{D}$ on g^* as*

$$\mathcal{D} = \{(X, \alpha) \in Tg^* \oplus T^*g^* \mid X(x) - J(x)\alpha(x) \in \mathcal{V},\ \alpha(x) \in \mathcal{V}^\perp,\ x \in g^*\}, \tag{4.48}$$

where $J(x)$ is the structure matrix of the Lie-Poisson bracket $\{\, , \,\}$. Then $\mathcal{D}$ is closed if and only if $\mathcal{V}^\perp \subset g$ is a subalgebra.

Proof. The proof follows more or less directly from results obtained in [MR, p. 287]. □

Example 4.1 ($\mathcal{X} = se^*(3) \backsimeq \mathbb{R}^6$). *The motion of a rigid body with respect to a body-fixed rotation reference frame in the center of mass is given* (*in the absence of gravity*) *by*

$$M\dot{\omega} + \omega \times M\omega = \tau, \tag{4.49}$$

$$m\dot{v} + \omega \times mv = F, \tag{4.50}$$

where $v, \omega \in \mathbb{R}^3$ are, respectively, the linear and the angular velocities, M is the inertia tensor, and $\tau, F \in \mathbb{R}^3$ are, respectively, the torques and the forces. By defining $\Pi, p \in \mathbb{R}^3$ as

$$\Pi = M\omega,\ \ \Pi = [\Pi_x, \Pi_y, \Pi_z]^T, \tag{4.51}$$

$$p = mv,\ \ p = [p_x, p_y, p_z]^T \tag{4.52}$$

and the Hamiltonian $H(\Pi, p)$ as

$$H(\Pi, p) = \frac{1}{2}\Pi^T M^{-1}\Pi + \frac{1}{2m}p^T p, \tag{4.53}$$

it follows that (4.49) *and* (4.50) *can be written as*

$$\begin{bmatrix} \dot{\Pi} \\ \dot{p} \end{bmatrix} = \underbrace{\begin{bmatrix} S(\Pi) & S(p) \\ S(p) & 0 \end{bmatrix}}_{J(\Pi,p)} \begin{bmatrix} \frac{\partial H}{\partial \Pi} \\ \frac{\partial H}{\partial p} \end{bmatrix} + \begin{bmatrix} \tau \\ F \end{bmatrix}. \tag{4.54}$$

Here $\Pi = [\Pi_x, \Pi_y, \Pi_z]^T$ *and* $p = [p_x, p_y, p_z]^T$ *are the body angular and linear momentum, respectively.* $S(\,\cdot\,)$ *is defined by* $S(a)b = a \times b$ *for* $a, b \in \mathbb{R}^3$*.* $J(\Pi, p)$ *is the structure matrix of the Lie-Poisson bracket on* $\mathcal{X} = se^*(3) \simeq \mathbb{R}^6$*.*

Assume that the following constraints are imposed on the system:

$$p_y = p_z = 0. \tag{4.55}$$

Let $e_x = [1\;0\;0]^T$, $e_y = [0\;1\;0]^T$, $e_z = [0\;0\;1]^T$*. Then*

$$\mathcal{V}^\perp = \ker \begin{bmatrix} 0 & 0 \\ e_y & e_z \end{bmatrix}, \tag{4.56}$$

which is not a subalgebra of $se(3) \simeq \mathbb{R}^6$ (*see, e.g.,* [MR])*. Hence, the corresponding generalized Dirac structure is not closed in this case. However, if the additional constraint* $p_x = 0$ *is imposed on the system* (*fixed center of mass*)*, it is easy to see that closedness of the corresponding generalized Dirac structure follows.*

Similarly to the case when the Jacobi-identity is satisfied for a generalized Poisson structure, one can show that if the closedness condition (4.1) is satisfied for a generalized Dirac structure then there exist local canonical coordinates around any *regular* point in which the geometric picture simplifies considerably (see Proposition 4.1.2 in [C1]). In our context (i.e., for generalized Dirac structures arising from physical systems) constant-dimensionality of the codistribution P_1 is often a reasonable assumption. Thus, in the next proposition we will draw attention to the existence and construction of canonical coordinates for Dirac structures that may be given in representation II (cf. Theorem 3.1). In essence, the proof of this proposition comes down to using Frobenius's theorem and a generalized version of Darboux's theorem and proceeds along the same general line as the proof of Proposition 4.1.2 in [C1]. However, we show directly how local canonical coordinates may be found for a Dirac structure in representation II. In addition, we show more explicitly where the three necessary conditions in Corollary 4.1 come into play which is interesting in itself.

Proposition 4.3. *Let* $\mathcal{D}$ *be a generalized Dirac structure on an* n*-dimensional manifold* $\mathcal{X}$*. Assume that the codistribution* P_1 (*see* (3.2)) *is constant-dimensional so that* $\mathcal{D}$ *can always be given in representation* II *as follows*:

$$\mathcal{D} = \{(X, \alpha) \in T\mathcal{X} \oplus T^*\mathcal{X} \mid X(x) - J(x)\alpha(x) \in G_0(x), x \in \mathcal{X}, \alpha \in \operatorname{ann} G_0\}, \tag{4.57}$$

where $J(x) : T_x^*\mathcal{X} \to T_x\mathcal{X}$, $x \in \mathcal{X}$, *is a skew-symmetric vector bundle map. Then, if* $\mathcal{D}$ *is closed, there exist around every regular point* $x_0 \in \mathcal{X}$ *local canonical coordinates*

$$(q, p, r, s) = (q_1, \ldots, q_k, p_1, \ldots, p_k, r_1, \ldots, r_l, s_1, \ldots, s_m), \quad 2k + l + m = n$$

for $\mathcal{X}$ *in which* $J(x)$ *and* G_0 *take the simple form*

$$J(x) = \begin{bmatrix} 0 & I_k & 0 & * \\ -I_k & 0 & 0 & * \\ 0 & 0 & 0 & * \\ * & * & * & * \end{bmatrix}, \quad G_0 = span\left\{\frac{\partial}{\partial s_1}, \ldots, \frac{\partial}{\partial s_m}\right\}, \tag{4.58}$$

where $*$ *denotes unspecified elements,* $m = n - \dim P_1$ *and* $l = n - \dim G_1(x_0)$*.*

Conversely, if $\mathcal{D}$ *is given by* (4.57), (4.58) *in a neighborhood of* $x_0 \in \mathcal{X}$*, then* $\mathcal{D}$ *is closed in this neighborhood.*

Proof. If $\mathcal{D}$ is closed, it follows from condition 1 in Corollary 4.1 that G_0 is involutive. Since $P_1 = \operatorname{ann} G_0$ is constant-dimensional, also G_0 is constant-dimensional with dimension equal to m. Thus, by Frobenius' theorem in a neighborhood N_{x_0} of any point $x_0 \in \mathcal{X}$ there exist local coordinates $(y, s) = (y_1, \dots, y_{n-m}, s_1, \dots, s_m)$, such that

$$G_0 = \operatorname{span}\left\{\frac{\partial}{\partial s_1}, \dots, \frac{\partial}{\partial s_m}\right\} \tag{4.59}$$

and

$$P_1 = \operatorname{ann} G_0 = \operatorname{span}\left\{\mathbf{d}y_1, \dots, \mathbf{d}y_{n-m}\right\}. \tag{4.60}$$

Now, $\{\,,\,\}_{\mathcal{D}}$ is given in terms of $J(x)$ as follows:

$$\{F, G\}_{\mathcal{D}}(x) = \left[\frac{\partial F}{\partial x}(x)\right]^T J(x)\,\frac{\partial G}{\partial x}(x) \tag{4.61}$$

for all $F, G \in C^\infty(\mathcal{X})$ such that $\mathbf{d}F, \mathbf{d}G \in \operatorname{ann} G_0$. Moreover, since $\mathcal{D}$ is closed, it follows from condition 2 in Corollary 4.1 that

$$\mathbf{d}\{y_i, y_j\}_{\mathcal{D}} \in \operatorname{ann} G_0, \quad i, j = 1, \dots, n-m, \tag{4.62}$$

which means that

$$\frac{\partial\{y_i, y_j\}_D}{\partial s_k} = 0, \quad k = 1, \dots, m,\ i, j = 1, \dots, n-m. \tag{4.63}$$

Hence, $J(x)$ takes the following form in the local coordinates (y, s):

$$J(y, s) = \begin{bmatrix} \bar{J}(y) & * \\ * & * \end{bmatrix} \tag{4.64}$$

where $\bar{J}(y) = [\{y_i, y_j\}_{\mathcal{D}}]$ is the $(n-m) \times (n-m)$ upper-left submatrix of $J(y, s)$. In addition, the distribution G_1 is given locally in the coordinates (y, s) as

$$G_1(y, s) = \operatorname{Im}\begin{bmatrix} \bar{J}(y) & 0 \\ 0 & I_m \end{bmatrix}. \tag{4.65}$$

If $x_0 \in \mathcal{X}$ is a regular point, then G_1 is by definition constant-dimensional in a neighborhood of x_0 which implies that $\bar{J}(y)$ has constant rank $2k = n - (l + m)$ in a neighborhood $\hat{N}_{x_0} \subset N_{x_0}$ of x_0. Define (without loss of generality) the submanifold $\mathcal{Y} \subset \mathcal{X}$ as

$$\mathcal{Y} = \{(y, s) \in \hat{N}_{x_0} \mid s = s(x_0)\}. \tag{4.66}$$

$y = (y_1, \dots, y_{n-m})$ are local coordinates for $\mathcal{Y}$ around $y_0 = y(x_0)$. Since $\mathcal{D}$ is closed, it follows from condition 3 in Corollary 4.1 that $\{\,,\,\}_{\mathcal{D}}$ defines a Poisson structure on $\mathcal{Y}$ with structure matrix $\bar{J}(y)$. Now, using the fact that $\bar{J}(y)$ has constant rank $2k \leq n - m$ for all $y \in \mathcal{Y}$, it follows from Theorem 6.22 in [O] (called the generalized Darboux's theorem; see also [W]), that around $y_0 \in \mathcal{Y}$ there exist local coordinates $(q, p, r) = (q_1, \dots, q_k, p_1, \dots, p_k, r_1, \dots, r_l)$ in which $\bar{J}(y)$ takes the form

$$\bar{J}(p, q, r) = \begin{bmatrix} 0 & I_k & 0 \\ -I_k & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{4.67}$$

Now (q,p,r,s) are local coordinates for $\mathcal{X}$ around $x_0 \in \mathcal{X}$ in which $J(x)$ and G_0 take the simple form (4.58).

Conversely, it is easy to check that a generalized Dirac structure given by (4.57), (4.58) in a neighborhood of $x_0 \in \mathcal{X}$, satisfies the sufficient conditions for closedness as given in Theorem 4.2 in this neighborhood. □

The equations of an implicit generalized Hamiltonian system corresponding to the local representation (4.57), (4.58) and a Hamiltonian H take the form

$$\begin{array}{rcl} \dot{q} &=& \frac{\partial H}{\partial p}(q,p,r,s), \\ \dot{p} &=& -\frac{\partial H}{\partial q}(q,p,r,s), \\ \dot{r} &=& 0, \\ 0 &=& \frac{\partial H}{\partial s}(q,p,r,s). \end{array} \tag{4.68}$$

Comparing (4.68) with (2.7) we see that while (2.7) makes explicit the conserved quantities, (4.68) also makes explicit the algebraic *constraints*

$$\begin{array}{rcl} 0 &=& \frac{\partial H}{\partial s_1}(q,p,r,s), \\ & \vdots & \\ 0 &=& \frac{\partial H}{\partial s_m}(q,p,r,s). \end{array} \tag{4.69}$$

If H is nondegenerate in the energy-variables $s_1, \dots, s_m$, that is,

$$\operatorname{rank} \left[\frac{\partial^2 H}{\partial s_i \partial s_j} \right] = m, \tag{4.70}$$

then by the implicit function theorem one may locally express the variables $s_1, \dots, s_m$ as functions of $q,p,r,$ i.e., $s_i = s_i(q,p,r),\ i = 1, \dots, m.$ Defining the *constrained* Hamiltonian

$$H_c(q,p,r) := H(q,p,r,s(q,p,r)) \tag{4.71}$$

it follows that (4.68) reduces to the same format as (2.7):

$$\begin{array}{rcl} \dot{q} &=& \frac{\partial H_c}{\partial p}(q,p,r), \\ \dot{p} &=& -\frac{\partial H_c}{\partial q}(q,p,r), \\ \dot{r} &=& 0, \end{array} \tag{4.72}$$

which is an explicit Hamiltonian dynamics on the constrained state space $\mathcal{X}_c = \{(q,p,r,s) \mid \frac{\partial H}{\partial s_i}(q,p,r,s) = 0,\ i = 1, \cdots, m\}$. Also note that while under the assumption (4.70) the variables $s_1, \dots, s_m$ together with the Hamiltonian H define a (constraint) submanifold $\mathcal{X}_c$ of $\mathcal{X}$, dually the level sets of the variables $r_1, \dots, r_\ell$ define a *foliation* of $\mathcal{X}$. Both the constraint submanifold $\mathcal{X}_c$ and the foliation are *invariant* for the Hamiltonian dynamics. However, as shown in this section, there are cases of interest where the generalized Dirac structure does *not* satisfy the closedness condition (e.g., mechanical systems with *nonholonomic* constraints). Furthermore, also if the closedness condition is satisfied the actual *construction* of the canonical coordinates $q_i, p_i, r_i, s_i,$ may be very involved, and preferably should be avoided.

We remark that the representation (4.68) of an implicit Hamiltonian system with regard to a closed Dirac structure is quite amenable for *stability analysis*, at least when the nondegeneracy condition (4.70) is satisfied. Indeed, let (q_0, p_0, r_0, s_0) be an equilibrium of (4.68), that is,

$$\frac{\partial H}{\partial q}(q_0, p_0, r_0, s_0) = 0,\ \frac{\partial H}{\partial p}(q_0, p_0, r_0, s_0) = 0,\ \frac{\partial H}{\partial s}(q_0, p_0, r_0, s_0) = 0, \tag{4.73}$$

and let us also assume that $\frac{\partial H}{\partial r}(q_0, p_0, r_0, s_0) = 0$ (see later). Under the nondegeneracy condition (4.70) the implicit function theorem allows us to express the variables s locally around q_0, p_0, r_0, s_0 as functions of q, p, r leading as above to the explicit Hamiltonian dynamics (4.72). Note that in general the implicit function theorem only provides an *existence* result, and that finding the *actual* expression of s as function of q, p, r is in general not possible or preferably should be avoided.

Now, if the Hessian matrix of H_c at (q_0, p_0, r_0) is positive (or negative) definite it follows that (q_0, p_0, r_0) is a *stable* equilibrium of (4.72) (see, e.g., [MR]). On the other hand, this Hessian matrix of H_c can be easily expressed in the *original* Hamiltonian H as

$$\begin{bmatrix} \frac{\partial^2 H}{\partial q^2} & \frac{\partial^2 H}{\partial q \partial p} & \frac{\partial^2 H}{\partial q \partial r} \\ \frac{\partial^2 H}{\partial p \partial q} & \frac{\partial^2 H}{\partial p^2} & \frac{\partial^2 H}{\partial p \partial r} \\ \frac{\partial^2 H}{\partial r \partial q} & \frac{\partial^2 H}{\partial r \partial p} & \frac{\partial^2 H}{\partial r^2} \end{bmatrix} - \begin{bmatrix} \frac{\partial^2 H}{\partial q \partial s} \\ \frac{\partial^2 H}{\partial p \partial s} \\ \frac{\partial^2 H}{\partial r \partial s} \end{bmatrix} \left[\frac{\partial^2 H}{\partial s^2}\right]^{-1} \begin{bmatrix} \frac{\partial^2 H}{\partial s \partial q} & \frac{\partial^2 H}{\partial s \partial p} & \frac{\partial^2 H}{\partial s \partial r} \end{bmatrix} \tag{4.74}$$

evaluated at (q_0, p_0, r_0, s_0). Thus this way of checking stability can be performed *without* the actual computation of H_c. Furthermore, note that for checking definiteness of (4.74) only the variables s need to be explicitly computed; we may use other coordinates instead of q, p, r.

Since the variables $r_1, \ldots, r_l$ are invariants (or *Casimirs*) we may also replace in the stability analysis the constrained Hamiltonian H_c by $H_c(q, p, r) + \Phi(r)$, with Φ any function of $r = (r_1, \ldots, r_l)$. Hence we may also replace $H(q, p, r, s)$ with

$$\bar{H}_\Phi(q, p, r, s) := H(q, p, r, s) + \Phi(r) \tag{4.75}$$

and substitute $\bar{H}_\Phi$ into (4.74) in order to check definiteness. (The addition of a function $\Phi(r)$ to H_c when checking the definiteness of the Hessian is known as the energy-Casimir method; see, e.g., [MR].)

5. Implicit port-controlled generalized Hamiltonian systems.

As already alluded to in section 2, if we interconnect port-controlled Hamiltonian systems (2.1) in such a way that some of the external variables remain free port variables, then we will end up with an implicit generalized Hamiltonian system *with* external (or port) variables. In order to make this precise we give the following definition (see [SM2]).

DEFINITION 5.1. *Let $\mathcal{X}$ be an n-dimensional manifold of energy variables, and let $H : \mathcal{X} \to \mathbb{R}$ be a Hamiltonian. Furthermore, let $\mathcal{F}$ be the linear space $\mathbb{R}^m$ of external flows f, with dual the space $\mathcal{F}^*$ of external efforts e. Consider a Dirac structure on the product space $\mathcal{X} \times \mathcal{F}$, only depending on x. The implicit port-controlled generalized Hamiltonian system corresponding to $\mathcal{X}$, H, $\mathcal{D}$, and $\mathcal{F}$ is defined by the specification*

$$\left(\dot{x}, f, \frac{\partial H}{\partial x}(x), -e\right) \in \mathcal{D}(x). \tag{5.1}$$

REMARK 5.1. *The minus sign in front of the effort e comes from the natural identification $(\alpha, e) \in T^*\mathcal{X} \times \mathcal{F}^* \to (\alpha, -e) \in (T\mathcal{X} \times \mathcal{F})^*$. Physically this means that the ingoing power is counted positively.*

Since by definition of a Dirac structure (cf. (2.9)) $\langle \alpha \,|\, X\rangle - \langle e \,|\, f\rangle = 0$ for all $(X, f, \alpha, -e) \in \mathcal{D}$, it follows that an implicit port-controlled Hamiltonian system satisfies the energy balance

$$\frac{dH}{dt} = e^T f. \tag{5.2}$$

Definition 5.1 generalizes the notion of an (explicit) port-controlled generalized Hamiltonian system (2.1) by noting that in this case the Dirac structure $\mathcal{D}$ on $\mathcal{X} \times \mathcal{F}$ is given by the specification $(X, f, \alpha, -e) \in \mathcal{D}$ iff

$$\begin{aligned} X(x) &= J(x)\alpha(x) + g(x)f, \\ e &= g^T(x)\alpha(x), \quad x \in \mathcal{X}. \end{aligned} \tag{5.3}$$

Indeed, let $(X, f, \alpha, -e) \in \mathcal{D}^\perp$; that is,

$$\langle \hat{\alpha} \,|\, X\rangle + \langle \alpha \,|\, \hat{X}\rangle - \langle \hat{e} \,|\, f\rangle - \langle e \,|\, \hat{f}\rangle = 0 \tag{5.4}$$

for all $(\hat{X}, \hat{f}, \hat{\alpha}, -\hat{e})$ satisfying (5.3). By first taking $\hat{f} = 0$ we obtain

$$\hat{\alpha}^T(x)X(x) + \alpha^T(x)J(x)\hat{\alpha}(x) - \hat{\alpha}^T(x)g(x)f = 0 \tag{5.5}$$

for all $\hat{\alpha}$, and thus $X(x) = J(x)\alpha(x) + g(x)f$, and substitution in (5.4) yields

$$\hat{\alpha}^T(x)g(x)f + \alpha^T(x)g(x)\hat{f} - \hat{\alpha}^T(x)g(x)f - e^T\hat{f} = 0 \tag{5.6}$$

for all $\hat{f}$, implying that $e = g^T(x)\alpha(x)$, and thus that $(X, f, \alpha, -e) \in \mathcal{D}$.

Now let us consider, as in section 2, k port-controlled generalized Hamiltonian systems, see (2.15), with $\mathcal{E}_j = \mathcal{F}_j^*$, $j = 1, \dots, k$. A power-conserving *partial* interconnection is obtained by writing a direct sum decomposition

$$\mathcal{F}_1 \times \cdots \times \mathcal{F}_k = \mathcal{F}^i \oplus \mathcal{F}^p \tag{5.7}$$

with the subspace $\mathcal{F}^i$ denoting the flows to be interconnected, and $\mathcal{F}^p$ the remaining flows at the external ports of the partially interconnected system. By defining $\mathcal{E}^i := (\mathcal{F}^p)^\perp$ and $\mathcal{E}^p := (\mathcal{F}^i)^\perp$ we obtain the dual direct sum decomposition

$$\mathcal{E}_1 \times \cdots \times \mathcal{E}_k = \mathcal{E}^i \oplus \mathcal{E}^p. \tag{5.8}$$

Proposition 5.1. *Consider as in* (2.15) *k port-controlled generalized Hamiltonian systems, with direct sum decomposition* (5.7), (5.8). *Consider a power-conserving partial interconnection given by a subspace* (*possibly parametrized by $x_1, \dots, x_k$*)

$$I(x_1, \dots, x_k) \subset \mathcal{F}^i \times \mathcal{E}^i \tag{5.9}$$

with $\dim I(x_1, \dots, x_k) = \dim \mathcal{F}^{\mathrm{i}}$, *having the property*

$$(f^i, e^i) \in I(x_1, \dots, x_k) \;\Rightarrow\; \langle e^i \,|\, f^i\rangle = 0. \tag{5.10}$$

Then the resulting partially interconnected system is an implicit port-controlled generalized Hamiltonian system with state space $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$, Hamiltonian

$H(x_1,\dots,x_k) := H_1(x_1) + \cdots + H_k(x_k)$, and generalized Dirac structure on $\mathcal{X} \times \mathcal{F}^p$ given as

$$\begin{aligned}
&(X, f^p, \alpha, -e^p) = (X_1,\dots,X_k, f^p, \alpha_1,\dots,\alpha_k, -e^p) \in \mathcal{D} \iff \\
&X_j(x_j) = J_j(x_j)\alpha_j(x_j) + g_j(x_j) f_j, \\
&e_j = g_j^T(x_j)\alpha_j(x_j), \quad x_j \in \mathcal{X}_j,\ j = 1,\dots,k, \\
&(f_1,\dots,f_k, e_1,\dots,e_k) = (f^i, f^p, e^i, e^p) \text{ such that } (f^i, e^i) \in I(x_1,\dots,x_k).
\end{aligned} \tag{5.11}$$

Proof. The proof is very similar to the proof of Proposition 2.2. Let $(X, f^p, \alpha, -e^p)$ be in $\mathcal{D}^\perp$, that is,

$$\langle \hat{\alpha} \,|\, X \rangle + \langle \alpha \,|\, \hat{X} \rangle - \langle \hat{e}^p \,|\, f^p \rangle - \langle e^p \,|\, \hat{f}^p \rangle = 0 \tag{5.12}$$

for all $(\hat{X}, \hat{f}^p, \hat{\alpha}, -\hat{e}^p)$ satisfying (5.11). First, letting $\hat{f}_j = \hat{e}_j = 0$ for all $j = 1,\dots,k$ we obtain (2.21), and by substitution in (5.12) we obtain, similar to (2.22),

$$0 = \sum_{j=1}^{k} \left(\hat{e}_j^T f_j + e_j^T \hat{f}_j \right) - \langle \hat{e}^p \,|\, f^p \rangle - \langle e^p \,|\, \hat{f}^p \rangle = \langle \hat{e}^i \,|\, f^i \rangle + \langle e^i \,|\, \hat{f}^i \rangle \tag{5.13}$$

for all $\hat{f}^i, \hat{e}^i$. By definition of $I(x_1,\dots,x_k)$ in (5.10) this implies, as in Proposition 2.2 (adding if necessary flow vectors in the kernel of $g_j(x_i)$) that $(f^i, e^i) \in I(x_1,\dots,x_k)$, and thus $(X, f^p, \alpha, -e^p) \in \mathcal{D}$. Since it is readily seen that $\mathcal{D} \subset \mathcal{D}^\perp$ it follows that $\mathcal{D}$ defines a Dirac structure. □

REMARK 5.2. *An interesting open problem is the* variational *interpretation of Proposition* 5.1 (*and Proposition* 2.2). *Indeed, if all the Hamiltonian subsystems admit a variational characterization* (*as Euler–Lagrange equations*) *one could conjecture that also the* (*partially*) *interconnected Hamiltonian system admits "some kind of" variational characterization. It is to be expected, however, that the closedness conditions as treated in this and the previous section will play an important role in such a characterization, since already for classical mechanical systems with kinematic constraints it is known* (*see e.g.,* [AKN, BC]) *that they cannot be formulated as standard Euler–Lagrange equations in case the constraints are* nonholonomic. *Also, the formulation* (4.68) *of an implicit Hamiltonian system satisfying the closedness condition suggests a connection with variational principles via the first-order condition of Pontryagin's maximum principle. In the case of electrical circuits, where the interconnections are defined by Kirchhoff's laws and the closedness conditions are trivially satisfied* (*see Example* 3.1), *some important work concerning a variational formulation of Kirchhoff's laws and the resulting variational characterization of the overall circuit has been done* (*see, e.g.,* [JE, M1]), *and it seems of interest to extend these ideas to the general situation considered in Proposition* 5.1.

In the rest of this section we will not elaborate on general implicit port-controlled Hamiltonian systems and their different representations, but instead concentrate on a special subclass which arises naturally in the control of mechanical systems. Consider the following port-controlled generalized Hamiltonian system *with constraints* given by

$$\begin{aligned} \dot{x} &= J(x)\tfrac{\partial H}{\partial x}(x) + g(x)f + b(x)\lambda, \\ e &= g^T(x)\tfrac{\partial H}{\partial x}(x), \\ 0 &= b^T(x)\tfrac{\partial H}{\partial x}(x), \end{aligned} \tag{5.14}$$

where $x \in \mathcal{X}$, $f \in \mathcal{F} := \mathbb{R}^m$ and $g(x) = [g_1(x) \dots g_m(x)]$ is the $n \times m$ matrix of input vector fields g_j. $b(x) = [b_1(x) \dots b_k(x)]$ is the $n \times k$ matrix of constraint vector fields. Throughout this section we will assume that $b(x)$ has rank equal to k everywhere. It is easily seen that, e.g., an *actuated* mechanical system with kinematic constraints will fit into the description (5.14). By rewriting (5.14) as

$$\begin{aligned} \begin{bmatrix} \dot{x} \\ f \end{bmatrix} &= \underbrace{\begin{bmatrix} J(x) & 0 \\ 0 & 0 \end{bmatrix}}_{\tilde{J}(x)} \begin{bmatrix} \frac{\partial H}{\partial x}(x) \\ -e \end{bmatrix} + \begin{bmatrix} g(x) & b(x) \\ I_m & 0 \end{bmatrix} \tilde{\lambda}, \\ 0 &= \begin{bmatrix} g^T(x) & I_m \\ b^T(x) & 0 \end{bmatrix} \begin{bmatrix} \frac{\partial H}{\partial x}(x) \\ -e \end{bmatrix}, \end{aligned} \tag{5.15}$$

where $\tilde{\lambda} \in \mathbb{R}^{m+k}$, it follows from Theorem 3.1 that (5.15) defines representation II of a generalized Dirac structure $\mathcal{D}$ on $\mathcal{X} \times \mathcal{F}$. Thus (5.14) is an implicit port-controlled generalized Hamiltonian system.

We will now study $\mathcal{D}$ further as given in representation (5.15). In what follows we will use $\{\,,\,\}$ and $\{\,,\,\}_{\mathcal{X}\times\mathcal{F}}$ to denote the generalized Poisson brackets on $\mathcal{X}$ and $\mathcal{X} \times \mathcal{F}$, respectively, with structure matrices $J(x)$ and $\tilde{J}(x)$ (see (5.15)), respectively. In addition we will let B denote the constant-dimensional distribution on $\mathcal{X}$ given by

$$B(x) = \operatorname{Im} b(x), \quad x \in \mathcal{X}. \tag{5.16}$$

From (5.15) we immediately see that the distribution G_0 on $\mathcal{X} \times \mathcal{F}$ defined by $\mathcal{D}$ (see (3.1)) is given by

$$G_0(x,y) = \operatorname{Im} \begin{bmatrix} g(x) & b(x) \\ I_m & 0 \end{bmatrix}, \quad (x,y) \in \mathcal{X} \times \mathcal{F}. \tag{5.17}$$

Note that G_0 is constant-dimensional with dimension equal to $m+k$ since $\operatorname{rank} b(x) = k$ for all $x \in \mathcal{X}$. The following lemma, for which a proof is straightforward, gives necessary and sufficient conditions for G_0 being involutive.

LEMMA 5.1. *G_0 is involutive if and only if $[X,Y] \in B$ for all $X,Y \in \{g_1,\dots,g_m,b_1,\dots,b_k\}$.*

The next lemma gives three necessary conditions for the closedness of $\mathcal{D}$.

LEMMA 5.2. *If the generalized Dirac structure $\mathcal{D}$ on $\mathcal{X} \times \mathcal{F}$ is closed, then*

1. $\{H_1,\{H_2,H_3\}\} + \{H_2,\{H_3,H_1\}\} + \{H_3,\{H_1,H_2\}\} = 0$;
2. $L_{g_j}\{H_1,H_2\} = \{L_{g_j}H_1,H_2\} + \{H_1,L_{g_j}H_2\}$, $j = 1,\dots,m$;
3. $\mathbf{d}\{H_1,H_2\} \in \operatorname{ann} B$

for all $H_1,H_2,H_3 \in C^\infty(\mathcal{X})$ such that $\mathbf{d}H_1, \mathbf{d}H_2, \mathbf{d}H_3 \in \operatorname{ann} B$.

Proof. Assume that $\mathcal{D}$ is closed, i.e., satisfies (4.1). Using *Cartan's magic formula*, the closedness condition (4.1) can be written as

$$\begin{aligned} &\langle \mathbf{d}\langle \alpha_2 \,|\, X_1\rangle \,|\, X_3\rangle + \langle \mathbf{d}\langle \alpha_3 \,|\, X_2\rangle \,|\, X_1\rangle + \langle \mathbf{d}\langle \alpha_1 \,|\, X_3\rangle \,|\, X_2\rangle \\ &\qquad + \mathbf{d}\alpha_2(X_1,X_3) + \mathbf{d}\alpha_3(X_2,X_1) + \mathbf{d}\alpha_1(X_3,X_2) = 0. \end{aligned} \tag{5.18}$$

Consider H_1, H_2, and $H_3 \in C^\infty(\mathcal{X})$, where $\mathbf{d}H_i \in \operatorname{ann} B$, $i = 1,2,3$. Let $j \in \{1,\dots,m\}$ and define

$$X_1(x,y) = \begin{bmatrix} X_{H_1}(x) \\ 0 \end{bmatrix}, \qquad \alpha_1(x,y) = \begin{bmatrix} \frac{\partial H_1}{\partial x}(x) \\ -L_g^T H_1(x) \end{bmatrix}, \tag{5.19}$$

$$X_2(x,y) = \begin{bmatrix} X_{H_2}(x) \\ 0 \end{bmatrix}, \qquad \alpha_2(x,y) = \begin{bmatrix} \frac{\partial H_2}{\partial x}(x) \\ -L_g^T H_2(x) \end{bmatrix}, \tag{5.20}$$

$$X_3(x,y) = \begin{bmatrix} X_{H_3}(x) + \rho g_j(x) \\ \rho Y_j \end{bmatrix}, \quad \alpha_3(x,y) = \begin{bmatrix} \frac{\partial H_3}{\partial x}(x) \\ -L_g^T H_3(x) \end{bmatrix} \tag{5.21}$$

for $(x,y) \in \mathcal{X} \times \mathcal{F}$, where $Y_j = \frac{\partial}{\partial y_j}$, $\rho \in \mathbb{R}$ and

$$X_{H_i}(x) = J(x)\frac{\partial H_i}{\partial x}(x), \; L_g^T H_i(x) = g^T(x)\frac{\partial H_i}{\partial x}(x). \tag{5.22}$$

Thus, $(X_i, \alpha_i) \in \mathcal{D}$, $i = 1,2,3$. Now, it is easy to see that $\langle \alpha_i \,|\, X_j \rangle = \{H_i, H_j\}$, $i,j = 1,2,3$, which implies that

$$\langle \mathbf{d}\langle \alpha_2 \,|\, X_1 \rangle \,|\, X_3 \rangle = \{\{H_2, H_1\}, H_3\} + \rho L_{g_j}\{H_2, H_1\}, \tag{5.23}$$

$$\langle \mathbf{d}\langle \alpha_3 \,|\, X_2 \rangle \,|\, X_1 \rangle = \{\{H_3, H_2\}, H_1\}, \tag{5.24}$$

$$\langle \mathbf{d}\langle \alpha_1 \,|\, X_3 \rangle \,|\, X_2 \rangle = \{\{H_1, H_3\}, H_2\}. \tag{5.25}$$

Moreover, we have that

$$\alpha_i = \frac{\partial H_i}{\partial x_1}\mathbf{d}x_1 + \cdots + \frac{\partial H_i}{\partial x_n}\mathbf{d}x_n - L_{g_1}H_i\mathbf{d}y_1 - \cdots - L_{g_m}H_i\mathbf{d}y_m, \quad i = 1,\dots,3, \tag{5.26}$$

which means that

$$\mathbf{d}\alpha_i = -\sum_{l=1}^{m}\sum_{k=1}^{n} \frac{\partial L_{g_l}H_i}{\partial x_k}\mathbf{d}x_k \wedge \mathbf{d}y_l, \quad i = 1,\dots,3. \tag{5.27}$$

Hence,

$$\mathbf{d}\alpha_2(X_1, X_3) \quad = \quad -\rho\left[\frac{\partial L_{g_j}H_2}{\partial x}\right]^T X_{H_1} \quad = \quad -\rho\{L_{g_j}H_2, H_1\} \tag{5.28}$$

and similarly

$$\mathbf{d}\alpha_1(X_3, X_2) \quad = \quad -\mathbf{d}\alpha_1(X_2, X_3) \quad = \quad \rho\{L_{g_j}H_1, H_2\} \quad = \quad -\rho\{H_2, L_{g_j}H_1\}. \tag{5.29}$$

In addition, it follows that $\mathbf{d}\alpha_3(X_2, X_1) = 0$. Therefore, from the integrability condition (5.18) we have that

$$\begin{aligned} &\{\{H_2, H_1\}, H_3\} + \{\{H_3, H_2\}, H_1\} + \{\{H_1, H_3\}, H_2\} \\ &\qquad + \rho\left(L_{g_j}\{H_2, H_1\} - \{L_{g_j}H_2, H_1\} - \{H_2, L_{g_j}H_1\}\right) = 0 \end{aligned} \tag{5.30}$$

for all $\rho \in \mathbb{R}$, implying condition 1 for $\rho = 0$ and condition 2 for $\rho = 1$.

Now, a direct calculation yields

$$\begin{aligned}\alpha_{12} &= \mathbf{i}_{X_2}\mathbf{d}\alpha_1 - \mathbf{i}_{X_1}\mathbf{d}\alpha_2 + \mathbf{d}\langle \alpha_1 \mid X_2\rangle \\ &= -\sum_{l=1}^{m}(\{L_{g_l}H_1, H_2\} + \{H_1, L_{g_l}H_2\})\,\mathbf{d}y_l + \sum_{k=1}^{n}\frac{\partial\{H_1,H_2\}}{\partial x_k}\mathbf{d}x_k,\end{aligned} \tag{5.31}$$

from which condition 3 (and condition 2) follows directly since $\alpha_{12} \in \operatorname{ann} G_0$ (see Theorem 4.1). □

Before being able to give the sufficient and necessary conditions for $\mathcal{D}$ being closed, we also need the following result.

LEMMA 5.3. *If for arbitrary $H_1, H_2 \in C^\infty(\mathcal{X})$ such that $\mathbf{d}H_1, \mathbf{d}H_2 \in \operatorname{ann} B$ there holds*

1. $L_{g_j}\{H_1, H_2\} = \{L_{g_j}H_1, H_2\} + \{H_1, L_{g_j}H_2\}$, $j = 1, \ldots, m$,
2. $\mathbf{d}\{H_1, H_2\} \in \operatorname{ann} B$,

then $\{\tilde{H}_1, \tilde{H}_2\}_{\mathcal{X}\times\mathcal{F}} \in \mathcal{A}_{\mathcal{D}}$ for all $\tilde{H}_1, \tilde{H}_2 \in \mathcal{A}_{\mathcal{D}}$.

Proof. Take arbitrary $\tilde{H}_2, \tilde{H}_2 \in \mathcal{A}_{\mathcal{D}}$. From (5.17) we see that this is equivalent to

$$0 = b^T(x)\frac{\partial \tilde{H}_i}{\partial x}(x,y), \tag{5.32}$$

$$\frac{\partial \tilde{H}_i}{\partial y}(x,y) = -g^T(x)\frac{\partial \tilde{H}_i}{\partial x}(x,y) \tag{5.33}$$

for $i = 1, 2$. Let $j \in \{1, \ldots, m\}$ and define

$$\hat{g}_j(x,y) = \begin{bmatrix} g_j(x) \\ 0 \end{bmatrix}. \tag{5.34}$$

Then (5.33) can be written as

$$\frac{\partial \tilde{H}_i}{\partial y_k} = -L_{\hat{g}_k}\tilde{H}_i, \quad k = 1, \ldots, m, \tag{5.35}$$

for $i = 1, 2$. Now,

$$\{\tilde{H}_1, \tilde{H}_2\}_{\mathcal{X}\times\mathcal{F}} = \sum_{k,l=1}^{n} J_{kl}(x)\frac{\partial \tilde{H}_1}{\partial x_k}\frac{\partial \tilde{H}_2}{\partial x_l}, \tag{5.36}$$

so

$$\frac{\partial\{\tilde{H}_1, \tilde{H}_2\}_{\mathcal{X}\times\mathcal{F}}}{\partial y_j} = -\,\{L_{\hat{g}_j}\tilde{H}_1, \tilde{H}_2\}_{\mathcal{X}\times\mathcal{F}} - \{\tilde{H}_1, L_{\hat{g}_j}\tilde{H}_2\}_{\mathcal{X}\times\mathcal{F}}. \tag{5.37}$$

Since

$$\{H_1, H_2\} = \sum_{k,l=1}^{n} J_{kl}(x)\frac{\partial H_1}{\partial x_k}\frac{\partial H_2}{\partial x_l}, \quad H_1, H_2 \in C^\infty(\mathcal{X}), \tag{5.38}$$

it follows from condition 1 that

$$L_{\hat{g}_j}\{\tilde{H}_1, \tilde{H}_2\}_{\mathcal{X}\times\mathcal{F}} = \{L_{\hat{g}_j}\tilde{H}_1, \tilde{H}_2\}_{\mathcal{X}\times\mathcal{F}} + \{\tilde{H}_1, L_{\hat{g}_j}\tilde{H}_2\}_{\mathcal{X}\times\mathcal{F}}, \tag{5.39}$$

which inserted in (5.37) yields

$$\frac{\partial\{\tilde{H}_1,\tilde{H}_2\}_{\mathcal{X}\times\mathcal{F}}}{\partial y_j} = -L_{\hat{g}_j}\{\tilde{H}_1,\tilde{H}_2\}_{\mathcal{X}\times\mathcal{F}}. \tag{5.40}$$

Moreover, from condition 2 it follows that

$$b^T(x)\frac{\partial\{\tilde{H}_1,\tilde{H}_2\}_{\mathcal{X}\times\mathcal{F}}}{\partial x} = 0. \tag{5.41}$$

Thus, from (5.40) and (5.41) we see that $\{\tilde{H}_1,\tilde{H}_2\}_{\mathcal{X}\times\mathcal{F}} \in \mathcal{A}_{\mathcal{D}}$. □

We are now ready to present the necessary and sufficient conditions for (5.15) defining a Dirac structure on $\mathcal{X}\times\mathcal{F}$.

THEOREM 5.1. *The generalized Dirac structure $\mathcal{D}$ on $\mathcal{X}\times\mathcal{F}$ as defined by* (5.15) *is closed if and only if*

1. $[X,Y]\in B$ *for all vector fields* $X,Y\in\{g_1,\dots,g_m,b_1,\dots,b_k\}$;
2. $L_{g_j}\{H_1,H_2\} = \{L_{g_j}H_1,H_2\}+\{H_1,L_{g_j}H_2\}$, $j=1,\dots,m$;
3. $\mathbf{d}\{H_1,H_2\}\in \operatorname{ann} B$;
4. $\{H_1,\{H_2,H_3\}\}+\{H_2,\{H_3,H_1\}\}+\{H_3,\{H_1,H_2\}\}=0$

for all $H_1,H_2,H_3\in C^\infty(\mathcal{X})$ *such that* $\mathbf{d}H_1,\mathbf{d}H_2,\mathbf{d}H_3\in\operatorname{ann} B$.

Proof. The necessary and sufficient conditions for closedness of $\mathcal{D}$ follows immediately by combining the results in Lemma 5.1, Lemma 5.2, Lemma 5.3 and then using Theorem 4.4. □

COROLLARY 5.1 ($B=0$). *Let* $b(x)=0$ (*no constraints*) *in* (5.14). *Then the generalized Dirac structure $\mathcal{D}$ on $\mathcal{X}\times\mathcal{F}$ as defined by* (5.15) (*with* $b(x)=0$) *is closed if and only if*

1. $[g_i,g_j]=0$, $i,j=1,\dots,m$;
2. $L_{g_j}\{H_1,H_2\} = \{L_{g_j}H_1,H_2\}+\{H_1,L_{g_j}H_2\}$ *for all* $H_1,H_2\in C^\infty(\mathcal{X})$, $j=1,\dots,m$;
3. $\{\,,\,\}$ *satisfies the Jacobi identity.*

Hence, the closedness condition (4.1) for the generalized Dirac structure on $\mathcal{X}\times\mathcal{F}$ arising from the constrained port-controlled Hamiltonian system (5.14) translates (among other things) into strong conditions on the input vector fields g_j.

Conditions 2–4 in Theorem 5.1 may be succinctly expressed by requiring that the generalized Poisson bracket $\{\,,\,\}$ of $F,G\in C^\infty(\mathcal{X})$ where $\mathbf{d}F$, $\mathbf{d}G\in\operatorname{ann}B$ is preserved by the dynamics of (5.14) for *every* choice of internal energy H such that $\mathbf{d}H\in\operatorname{ann}B$ and for every $f\in\mathcal{F}$. Indeed, requiring that

$$\frac{d}{dt}\{F,G\} = \left\{\frac{d}{dt}F,G\right\}+\left\{F,\frac{d}{dt}G\right\} \tag{5.42}$$

for all $F,G\in C^\infty(\mathcal{X})$ such that $\mathbf{d}F,\mathbf{d}G\in\operatorname{ann}B$, where $\frac{d}{dt}$ denotes the time-derivative along (5.14), is equivalent to

$$\begin{aligned}\{\{F,G\},H\}&+L_g\{F,G\}f+L_b\{F,G\}\lambda\\ &=\{\{F,H\},G\}+\{(L_gF)f,G\}+\{F,\{G,H\}\}+\{F,(L_gG)f\}\end{aligned} \tag{5.43}$$

for all $H\in C^\infty(\mathcal{X})$ such that $\mathbf{d}H\in\operatorname{ann}B$ and $f\in\mathcal{F}$. Letting $H=0$ and $f=0$, we obtain

$$L_b\{F,G\}\lambda = 0 \quad \text{for all } \lambda\in\mathbb{R}^k, \tag{5.44}$$

which means that $\mathbf{d}\{F, G\} \in \operatorname{ann} B$. Moreover, letting $f = 0$ leads to

$$\{\{F, G\}, H\} = \{\{F, H\}, G\} + \{F, \{G, H\}\}, \tag{5.45}$$

which is none other than the Jacobi-identity. Thus, (5.43) amounts to

$$L_g\{F, G\}f = \{(L_g F)f, G\} + \{F, (L_g G)f\} \quad \text{for all } f \in \mathcal{F}, \tag{5.46}$$

which is equivalent to

$$L_{g_j}\{F, G\} = \{L_{g_j} F, G\} + \{F, L_{g_j} G\}, \quad j = 1, \dots, m. \tag{5.47}$$

The next example should give an idea of what the conditions in Theorem 5.1 imply for the (local) mathematical structure of system (5.14).

Example 5.1. *Consider the port-controlled generalized Hamiltonian system with constraints given in* (5.14). *Assume that conditions* 1–4 *in Theorem* 5.1 *are all satisfied. By condition* 1 *it follows that the constant-dimensional distribution B is involutive. Hence, by Frobenius' theorem in a neighborhood of any point $x_0 \in \mathcal{X}$ there exist local coordinates $(y, s) = (y_1, \dots, y_{n-k}, s_1, \dots, s_k)$, such that*

$$\operatorname{ann} B = \operatorname{span}\{\mathbf{d}y_1, \dots, \mathbf{d}y_{n-k}\} \tag{5.48}$$

and

$$B = \operatorname{span}\left\{\frac{\partial}{\partial s_1}, \dots, \frac{\partial}{\partial s_k}\right\}. \tag{5.49}$$

Condition 3 *implies that*

$$\frac{\partial\{y_i, y_j\}}{\partial s_l} = 0, \quad l = 1, \dots, k, \ i, j = 1, \dots, n - k. \tag{5.50}$$

Hence, $J(x)$ takes the following form in the local coordinates (y, s):

$$J(y, s) = \begin{bmatrix} J_{yy}(y) & * \\ * & * \end{bmatrix}, \tag{5.51}$$

where $J_{yy}(y) = [\{y_i, y_j\}]$ is the $(n - k) \times (n - k)$ upper-left submatrix of $J(y, s)$. From condition 1 *it also follows that $[b_i, g_j] \in B$ which implies that in the coordinates (y, s) the matrix of input vector fields takes the form*

$$g(x, y) = \begin{bmatrix} g_y(y) \\ g_s(y, s) \end{bmatrix}. \tag{5.52}$$

Furthermore, since

$$[g_i, g_j](y, s) = \begin{bmatrix} [g_{y_i}, g_{y_j}](y) \\ * \end{bmatrix} \tag{5.53}$$

while $[g_i, g_j] \in B$, it follows that $[g_{y_i}, g_{y_j}] = 0$, $i, j = 1, \dots, m$. Assume additionally that the distribution $B + G$ is constant-dimensional with dimension equal to $m + k$. Then the submatrix $g_y(y)$ of $g(x, y)$ has constant rank equal to $m \leq n - k$. Thus *(see e.g., Theorem* 2.36 *in* [NS]), *there exists a local transformation* $(y_1, \dots, y_{n-k}) \to (\tilde{y}_1, \dots, \tilde{y}_{n-k})$ *such that*

$$g_{y_j} = \frac{\partial}{\partial \tilde{y}_j}, \quad j = 1, \dots m. \tag{5.54}$$

In these coordinates condition 2 *amounts to*

$$\frac{\partial \{\tilde{y}_i, \tilde{y}_j\}}{\partial \tilde{y}_l} = \left\{\frac{\partial \tilde{y}_i}{\partial \tilde{y}_l}, \tilde{y}_j\right\} + \left\{\tilde{y}_i, \frac{\partial \tilde{y}_j}{\partial \tilde{y}_l}\right\} = 0, \quad l = 1, \dots, m, \ i, j = 1, \dots n-k, \tag{5.55}$$

which means that $\{\tilde{y}_i, \tilde{y}_j\}$ *is independent of the first* m *local coordinates* $\tilde{y}_1, \dots, \tilde{y}_m$*. Let now* $z = (\tilde{y}_{m+1}, \dots, \tilde{y}_{n-k})$ *and* $w = (\tilde{y}_1, \dots, \tilde{y}_m)$*. Then from the discussion above we can conclude that* (z, w, s) *are local coordinates for* $\mathcal{X}$ *around* x_0 *in which* (5.14) *takes the form*

$$\begin{aligned} \begin{bmatrix} \dot{z} \\ \dot{w} \end{bmatrix} &= \begin{bmatrix} J_{zz}(z) & J_{zw}(z) \\ -J_{zw}^T(z) & J_{ww}(z) \end{bmatrix} \begin{bmatrix} \frac{\partial H}{\partial z}(z, w, s) \\ \frac{\partial H}{\partial w}(z, w, s) \end{bmatrix} + \begin{bmatrix} 0 \\ I_m \end{bmatrix} f, \\ \dot{s} &= J_{sz}(z, w, s)\tfrac{\partial H}{\partial z}(z, w, s) + J_{sw}(z, w, s)\tfrac{\partial H}{\partial w}(z, w, s) \\ &\quad + g_s(z, w, s)f + b_s(z, w, s)\lambda, \\ e &= \tfrac{\partial H}{\partial w}(z, w, s), \\ 0 &= \tfrac{\partial H}{\partial s}(z, w, s), \end{aligned} \tag{5.56}$$

where the last equation follows from the fact that the $k \times k$ *matrix* $b_s(z, w, s)$ *has full rank. Note that the equation for* $\dot{s}$ *can be* left out *from* (5.56) *because it is needed only to determine the Lagrange multipliers* $\lambda \in \mathbb{R}^k$*. Finally from condition* 4 *it follows that the matrix*

$$\begin{bmatrix} J_{zz}(z) & J_{zw}(z) \\ -J_{zw}^T(z) & J_{ww}(z) \end{bmatrix} \tag{5.57}$$

satisfies the Jacobi-identity (*in the* (*z,w*)*-coordinates*).

Finally, in the next example we will relate the results in this paper (in particular this section) to "passivity-based control" of actuated mechanical systems with kinematic constraints.

Example 5.2. *Consider a mechanical system with kinematic constraints* $A^T(q)\dot{q} = 0$ *as in Example* 3.2. *Additionally, let the system be actuated by generalized external forces* $u = (u_1, \dots, u_m)$ *corresponding to generalized configuration coordinates* $C_1(q), \dots, C_m(q)$*. The dynamical equations of motion are given as*

(5.58)

$$\begin{aligned} \begin{bmatrix} \dot{q} \\ \dot{p} \end{bmatrix} &= \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix} \begin{bmatrix} \frac{\partial H}{\partial q}(q,p) \\ \frac{\partial H}{\partial p}(q,p) \end{bmatrix} + \textstyle\sum_{i=1}^m \begin{bmatrix} 0 \\ \frac{\partial C_i}{\partial q}(q) \end{bmatrix} u_i + \begin{bmatrix} 0 \\ A(q) \end{bmatrix} \lambda, \\ y_i &= \left[\tfrac{\partial C_i}{\partial q}(q)\right]^T \tfrac{\partial H}{\partial p}(q,p) \ \left(= \tfrac{dC_i}{dt}(q)\right), \quad i = 1, \dots, m, \\ 0 &= A^T(q)\tfrac{\partial H}{\partial p}(q,p) \ (= A^T(q)\dot{q}). \end{aligned}$$

This is a port-controlled generalized Hamiltonian system with constraints as in (5.14) *with external flows the vector* $(u_1, \ldots, u_m)$ *of external forces, and external efforts the vector* $(y_1, \ldots, y_m)$ *of corresponding generalized velocities. It can be verified, as in Proposition* 4.1, *that the underlying generalized Dirac structure satisfies the conditions of Theorem* 5.1 (*that is, is closed*) *if and only if the kinematic constraints* $A^T(q)\dot{q} = 0$ *are* holonomic.

Now, consider an additional port-controlled Hamiltonian system (*the* "*controller*")

$$\begin{aligned} \dot{\xi} &= u_c, \\ y_c &= \tfrac{\partial P}{\partial \xi}(\xi), \quad \xi, u_c, y_c \in \mathbb{R}^m, \end{aligned} \tag{5.59}$$

with Hamiltonian P. (*Note that this is of the type* (2.1) *with* $J = 0$, $g =$ *identity matrix*, $x = \xi$, $f = u_c$, *and* $e = y_c$.) *Feedback interconnection as in Example* 2.1 *leads to the implicit generalized Hamiltonian system*

$$\begin{aligned} \begin{bmatrix} \dot{q} \\ \dot{p} \\ \dot{\xi} \end{bmatrix} &= \begin{bmatrix} 0 & I_n & 0 \\ -I_n & 0 & -\frac{\partial C}{\partial q}(q) \\ 0 & \frac{\partial^T C}{\partial q}(q) & 0 \end{bmatrix} \begin{bmatrix} \frac{\partial H}{\partial q}(q,p) \\ \frac{\partial H}{\partial p}(q,p) \\ \frac{\partial P}{\partial \xi}(\xi) \end{bmatrix} + \begin{bmatrix} 0 \\ A(q) \\ 0 \end{bmatrix} \lambda, \\ 0 &= A^T(q)\tfrac{\partial H}{\partial p}(q,p), \end{aligned} \tag{5.60}$$

with $\frac{\partial C}{\partial q}(q)$ *denoting the matrix with ith column* $\frac{\partial C_i}{\partial q}(q)$. *The codistribution* P_0 *of the underlying generalized Dirac structure can be readily seen to be given as*

$$P_0 = \operatorname{span}\{\mathbf{d}C_i - \mathbf{d}\xi_i \mid i = 1, \ldots, m\} \tag{5.61}$$

expressing the fact (*see also Remark* 3.1) *that the functions*

$$C_i(q) - \xi_i, \quad i = 1, \ldots, m, \tag{5.62}$$

are independent conserved quantities for the closed-loop dynamics (5.60). *It follows that along* (5.60)

$$\xi_i(t) = C_i(q(t)) + c_i, \quad \textit{for all } t, \quad i = 1, \ldots, m, \tag{5.63}$$

with the constants c_i *solely depending on the initial conditions of the* "*controller*" (5.59).

Substituting (5.63) *into* (5.60) *and noting that*

$$\frac{\partial C}{\partial q}(q)\frac{\partial P}{\partial \xi}(C_1(q) + c_1, \ldots, C_m(q) + c_m) = \frac{\partial P}{\partial q}(C_1(q) + c_1, \ldots, C_m(q) + c_m), \tag{5.64}$$

it follows that the dynamics of the (q,p)*-part of* (5.60) (*the original mechanical system*) *are given as*

$$\begin{aligned} \begin{bmatrix} \dot{q} \\ \dot{p} \end{bmatrix} &= \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix} \begin{bmatrix} \frac{\partial H_{new}}{\partial q}(q,p) \\ \frac{\partial H_{new}}{\partial p}(q,p) \end{bmatrix} + \begin{bmatrix} 0 \\ A(q) \end{bmatrix} \lambda, \\ 0 &= A^T(q)\tfrac{\partial H_{new}}{\partial p}(q,p), \end{aligned} \tag{5.65}$$

where H_{new} is the "new" Hamiltonian defined by

$$H_{new}(q,p) = H(q,p) + P(C_1(q) + c_1, \dots, C_m(q) + c_m). \tag{5.66}$$

Thus by appropriately choosing the Hamiltonian $P(\xi)$ of the "controller sub-system" (5.59), *we may* shape *the Hamiltonian $H(q,p)$ of the constrained mechanical system* (5.58) *by addition of the* potential energy $P(C_1(q) + c_1, \dots, C_m(q) + c_m)$, *with $c_1, \dots, c_m$ only depending on the initial condition of* (5.59) (*that is*, *with properly initialization we may set $c_1 = \dots = c_m = 0$*). *This idea of shaping the internal energy is one of the main ideas of "passivity-based control." We have thus demonstrated that this can be accomplished by power-conserving* (*in fact*, *feedback*) *interconnection of* (5.58) *with a controller sub-system* (5.59).

In particular, *if H and $C_1, \dots, C_m$ are such that P can be chosen in such a manner that H_{new} as defined by* (5.66) *has a strict minimum at some desired equilibrium point (q_0, p_0)*, *then (q_0, p_0) will be a* (*Lyapunov*) *stable equilibrium of* (5.65) (*and*, *because of* (5.63), *also the ξ-dynamics will be stable*). *To be more precise we only need the function H_{new}* restricted *to the constraint manifold $\{(q,p) \mid A^T(q)\frac{\partial H_{new}}{\partial p}(q,p) = 0\}$ to have a strict minimum at (q_0, p_0).*

It can be verified that the underlying generalized Dirac structure of (5.60) *is closed if and only if the kinematic constraints $A^T(q)\dot{q} = 0$ are* holonomic. *If this happens to be the case then checking that H_{new} restricted to the constraint manifold has a strict minimum may be performed as indicated at the end of section* 4.

Within the same philosophy one may pursue asymptotic *stability by adding*, *apart from the energy-shaping Hamiltonian controller* (2.24), *energy-dissipating elements to the system. In particular*, *one may replace the feedback interconnection $u_c = y$, $u = -y_c$ as above by the power-conserving* partial *interconnection* (*with free external flow v and external effort y*)

$$\begin{array}{rcl} u_c & = & y, \\ u & = & -y_c + v, \end{array} \tag{5.67}$$

and then terminate *this port by an energy-dissipating element*

$$v = -\frac{\partial R}{\partial y}(y) \tag{5.68}$$

for some (*Rayleigh*) *dissipation function R. For the asymptotic stability analysis of the resulting closed-loop system one again must distinguish between holonomic and nonholonomic kinematic constraints $A^T(q)\dot{q} = 0$.* (*In fact*, *in the nonholonomic case there is a fundamental obstruction to asymptotic stabilization*, *since Brockett's necessary conditions are not satisfied*; *see*, *e.g.*, [MS3] *for the references.*)

6. Conclusions. It has been shown that a power-conserving interconnection of port-controlled generalized Hamiltonian systems leads to an implicit generalized Hamiltonian system, and a power-conserving partial interconnection to an implicit port-controlled Hamiltonian system. The crucial concept is the notion of a (generalized) Dirac structure, defined on the space of energy-variables or on the product of the space of energy-variables and the space of flow-variables in the port-controlled case. Three natural representations of generalized Dirac structures have been treated. Necessary and sufficient conditions for closedness of a Dirac structure in all three representations have been obtained. This has been illustrated on mechanical systems with kinematic constraints and constrained systems on dual Lie algebras. Canonical coordinates for (closed) Dirac structures have been discussed, as well as their use for stability analysis of implicit Hamiltonian systems. Finally the theory has been applied to implicit port-controlled generalized Hamiltonian systems, such as actuated mechanical systems with kinematic constraints, and it has been shown in particular that the closedness condition for the Dirac structure leads to strong conditions on the input vector fields.

Acknowledgments. We would like to thank Bernhard Maschke (Conservatoire National des Arts et Métiers, Paris) for stimulating discussions. We also thank Peter Crouch (Arizona State University) and Tony Bloch (University of Michigan) for pointing out to the second author the relevance of Dirac structures.

REFERENCES

[AKN] V.I. Arnold, V.V. Kozlov, and A.I. Neishtadt, *Mathematical Aspects of Classical and Celestial Mechanics*, Springer-Verlag, Berlin, New York, 1997.

[AMR] R. Abraham, J.E. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications*, Springer-Verlag, Berlin, New York, 1988.

[BC] A.M. Bloch and P.E. Crouch, *Nonholonomic control systems on Riemannian manifolds*, SIAM J. Control Optim., 33 (1995), pp. 126–148.

[BKMM] A.M. Bloch, P.S. Krishnaprasad, J.E. Marsden, and R.M. Murray, *Nonholonomic mechanical systems with symmetry*, Arch. Rational Mech. Anal., 136 (1996), pp. 21–99.

[C1] T.J. Courant, *Dirac manifolds*, Trans. Amer. Math. Soc., 319 (1990), pp. 631–661.

[C2] P.E. Crouch, Handwritten notes, 1994.

[CW] T.J. Courant and A. Weinstein, *Beyond Poisson structures*, in Seminaire sud-rhodanien de geometrie 8, Travaux en Cours 27, Hermann, Paris, 1988.

[D1] I. Dorfman, *Dirac structures of integrable evolution equations*, Phys. Lett. A, 125 (1987), pp. 240–246.

[D2] I. Dorfman, *Dirac Structures and Integrability of Nonlinear Evolution Equations*, John Wiley, Chichester, 1993.

[D3] P. Dirac, *Generalized Hamiltonian dynamics*, Canad. J. Math., 2 (1950), pp. 129–148.

[JE] D.L. Jones and F.J. Evans, *Variational analysis of electrical networks*, J. Franklin Inst., 295 (1973), pp. 9–23.

[M1] A.G.J. MacFarlane, *An integral formulation of a canonical equation set for non-linear electrical networks*, Internat. J. Control, 11 (1970), pp. 449–470.

[M2] B.M. Maschke, *Elements on the modelling of multibody systems*, in Modelling and Control of Mechanisms and Robot Systems, C. Melchiorri and A. Tornambe, eds., World Scientific Publishing Ltd., River Edge, NJ, 1996, pp. 1–38.

[MBS] B.M. Maschke, C. Bidard, and A.J. van der Schaft, *Screw-vector bond graphs for the kinestatic and dynamic modeling of multibody systems*, Dynamic Systems and Control, 55(1994), pp. 637–644.

[MR] J.E. Marsden and T. Ratiu, *Introduction to Mechanics and Symmetry*, Springer-Verlag, Berlin, New York, 1994.

[MS1] B.M. Maschke and A.J. van der Schaft, *Port-controlled Hamiltonian systems: Modelling origins and system theoretic properties*, in Proc. IFAC Symp. NOLCOS, Bordeaux, France, 1992, M. Fliess, ed., International Federation of Automatic Control, pp. 282–288.

[MS2] B.M. Maschke and A.J. van der Schaft, *System-theoretic properties of port-controlled Hamiltonian systems*, in Systems and Networks: Mathematical Theory and Applications, vol. II, Akademie Verlag, Berlin, 1994, pp. 349–352.

[MS3] B.M. Maschke and A.J. van der Schaft, *A Hamiltonian approach to stabilization of nonholonomic mechanical systems*, in Proc. 33rd IEEE CDC, Orlando, FL, 1994, IEEE Control Systems Society, pp. 2950–2954.

[MSB1] B.M. Maschke, A.J. van der Schaft, and P.C. Breedveld, *An intrinsic Hamiltonian formulation of network dynamics: Non-standard Poisson structures and gyrators*, J. Franklin Inst., 329 (1992), pp. 923–966.

[MSB2] B.M. Maschke, A.J. van der Schaft, and P.C. Breedveld, *An intrinsic Hamiltonian formulation of the dynamics of LC-circuits*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 42 (1995), pp. 73–82.

[NS] H. Nijmeijer and A.J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, Berlin, New York, 1990.

[O] P.J. Olver, *Applications of Lie Groups to Differential Equations*, 2nd ed., Springer-Verlag, Berlin, New York, 1993.

[OS] R. Ortega and M.W. Spong, *Adaptive motion control of rigid robots: A tutorial*, Automatica J. IFAC, 25 (1989), pp. 877–888.

[S] J.J. Slotine, *Putting Physics in Control: The Example of Robotics*, IEEE Control Systems Mag., 8 (1988), pp. 12–18.

[SM1] A.J. van der Schaft and B.M. Maschke, *On the Hamiltonian formulation of nonholonomic mechanical systems*, Rep. Math. Phys., 34 (1994), pp. 225–233.

[SM2] A.J. van der Schaft and B.M. Maschke, *The Hamiltonian formulation of energy conserving physical systems with external ports*, Arch. für Elektronik und Übertragungstechnik, 49 (1995), pp. 362–371.

[SM3] A.J. van der Schaft and B.M. Maschke, *Mathematical modelling of constrained Hamiltonian systems*, in Proc. IFAC Symp. NOLCOS, Tahoe City, CA, 1995, International Federation of Automatic Control, pp. 678–683.

[TA] M. Takegaki and S. Arimoto, *A new feedback method for dynamic control of manipulators*, Trans. ASME J. Dynam. System Meas. Control, 103 (1981), pp. 119–125.

[W] A. Weinstein, *The local structure of Poisson manifolds*, J. Differential Geom., 18 (1983), pp. 523–557.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 1, pp. 92–117

BOUNDED POWER SIGNAL SPACES FOR ROBUST CONTROL AND MODELING[*]

P. M. MÄKILÄ[†], J. R. PARTINGTON[‡], AND T. NORLANDER[§]

Abstract. The nonlinear space of signals allowing Wiener's generalized harmonic analysis (GHA), the linear bounded power signal spaces of Beurling, Marcinkiewicz, and Wiener, and a new linear bounded power space are studied from a control and systems theory perspective. Specifically, it is shown that the system power gain is given by the H_∞ norm of the system transfer function in each of these spaces for a large class of (power) stable finite and infinite dimensional systems. The GHA setup is shown to possess several limitations for the purpose of robustness analysis which motivates the use of the other more general (nonstationary) signal spaces. The natural double-sided time axis versions of bounded power signal spaces are shown to break the symmetry between Hardy space H_∞ methods and bounded power operators; e.g., the system transfer function being in H_∞ does not imply that a causal-linear time-invariant (LTI) system is bounded as an operator on any of the double-sided versions of the studied bounded power signal spaces.

Key words. bounded power signals, generalized harmonic analysis, robust control, infinite dimensional systems, H_∞

AMS subject classifications. 93A30, 93B36, 93D09, 43A15, 47B35

PII. S0363012997316664

1. Introduction.

It is popular to use bounded power signals (Boyd and Barratt (1991); Doyle, Francis, and Tannenbaum (1992); Zhou, Doyle, and Glover (1996); Gardner (1988)) as a physically well-motivated persistent signal setup in control and system identification studies. These provide also an alternative to finite energy L_2 signal spaces for introducing robust H_∞ control (Zames (1981)). Unfortunately, the most common definitions of the set of signals with bounded power do not result in a linear vector space setup as pointed out by Mari (1996). It is possible to eliminate this problem by considering some appropriate subset of the set of all bounded power signals, such as periodic or almost periodic signals; (see, e.g., Mäkilä (1990), Partington and Mäkilä (1996)). In this way, however, some generality is lost. An attractive alternative is to generalize the usual definition of the power size measure of a signal. It turns out that several interesting generalizations are possible and have in fact been treated in the functional analytic literature by famous mathematicians, such as Marcinkiewicz and Beurling (see Chen and Lau (1989) for some historical background). It is the purpose of this paper to treat both the standard setup and some of the generalizations of bounded power signals from a control and systems perspective.

A milestone in the development of rich nonprobabilistic signal models is the work of Wiener on GHA summarized in his book (Wiener (1933)); see also Wiener (1979). In his work the now standard definition of bounded power signals for the continuous-time case is given using the notion of autocorrelation function. This work generalizes essentially the Fourier integral to a large class of persistent signals in a rigorous way

[*] Received by the editors February 13, 1997; accepted for publication (in revised form) November 12, 1997; published electronically September 25, 1998.
http://www.siam.org/journals/sicon/37-1/31666.html

[†] Automation and Control Department, Tampere University of Technology, P.O. Box 692, FIN-33101 Tampere, Finland (pmakila@ae.tut.fi).

[‡] School of Mathematics, University of Leeds, Leeds LS2 9JT, UK (j.r.partington@leeds.ac.uk).

[§] Luleå University of Technology, Department of Computer Science and Electrical Engineering, SE-971 87 Luleå, Sweden (torben@sm.luth.se). The research of this author was supported by the Swedish Research Council for Engineering Sciences.

and gives a functional analytic basis for spectral analysis. In Wiener (1927) GHA is considered for sequences (i.e., the discrete-time case is covered). It was well known to Wiener that the set of bounded power signals as introduced by him (the set S in Wiener (1933)) is not a linear space; see, e.g., his treatment of so-called coherency issues (Wiener, (1930), (1949)). Wiener's work on GHA has inspired Gardner (1988) to develop a complete nonprobabilistic paradigm for spectral analysis.

Robust H_∞ control is fundamental to modern feedback control theory and control engineering. Due to mathematical difficulties with standard bounded power signal setups, robust H_∞ control and estimation theory have been developed mostly based on a finite energy L_2 (ℓ_2) signal space setting, although much of the engineering motivation for robust disturbance rejection and estimation comes from persistent bounded power–like (stationary and nonstationary) signal settings; see, e.g., the seminal paper by Zames (1981). Several recent books and papers (Boyd and Barratt (1991); Doyle, Francis, and Tannenbaum (1992); Zhou et al. (1994); Zhou, Doyle, and Glover (1996)) on robust control emphasize the importance of persistent bounded power signal setups for robust control but give only results for finite dimensional systems using some subset of the set of signals allowing GHA. In the present paper we generalize, among other things, system gain results to full GHA setups and beyond to Marcinkiewicz and related spaces for large classes of stable finite dimensional and infinite dimensional systems. Furthermore, we introduce a new bounded power–like signal space. The equivalency between the Hardy space H_∞ of bounded analytic transfer functions and spaces of power stable causal LTI operators is lost in the double-sided time axis case.

The paper is organized as follows. In section 2 GHA methods are studied. Several results are proved here on discrete- and continuous-time GHA analysis, including Theorems 2.1–2.4. Technical difficulties restricting the generality of GHA methods in, e.g., robustness analysis are demonstrated. More general bounded power signal spaces are studied in section 3. Specifically, bounded power signal spaces that can be traced back to Marcinkiewicz, Wiener, and Beurling are studied here. It is shown that for a large class of stable, causal, LTI finite and infinite dimensional systems, the system power gain is given by the H_∞ norm of the system transfer function in each of the considered generalized bounded power signal spaces. Conclusions are drawn in section 4.

2. The average power setup—autocorrelation function.

2.1. Discrete-time case. Consider now the standard form of the bounded power signal setup, which is based on an average power measure. A (real or complex) sequence $x = \{x_k\}_{k=-\infty}^{\infty}$ is said to have average (symmetric) bounded power if

$$\|x\|_A = \left(\lim_{n\to\infty} \frac{1}{2n+1} \sum_{k=-n}^{n} |x_k|^2 \right)^{1/2} < \infty. \tag{2.1}$$

This is the discrete analogue of the continuous-time definition in Doyle, Francis, and Tannenbaum (1992); Zhou, Doyle, and Glover (1996). It has been recently pointed out in a control context in Mari (1996) that the set of all finite power signals is not a linear vector space in the continuous-time case. The same is true in the discrete case (Mäkilä and Partington (1996)), even if it is required that the allowable signals must, in addition, have bounded amplitude (i.e., must belong to ℓ_∞) and be such that they possess a well-defined autocorrelation function.

A well-motivated starting point for a reexamination of such issues is to study Wiener's GHA (Wiener (1927), (1930), (1933)). It is appropriate to start by stating some results related to GHA but which actually apply under more general assumptions than those used in GHA. To do this we introduce a seminorm which coincides with the seminorm $\|u\|_A$ given in (2.1), whenever that exists. Furthermore, we introduce a norm which will turn out to be most useful later. Let $\{u_k\}$ denote a real or complex sequence. Introduce the seminorm

$$\|u\|_B = \left(\limsup_{n\to\infty} \frac{1}{2n+1} \sum_{k=-n}^{n} |u_k|^2 \right)^{1/2} \tag{2.2}$$

and the norm

$$\|u\|_{BP} = \left(\sup_{n} \frac{1}{2n+1} \sum_{k=-n}^{n} |u_k|^2 \right)^{1/2} . \tag{2.3}$$

We use the notations u_k and $u(k)$ to denote the general element of the sequence $u = \{u_k\} = \{u(k)\}$.

Note that the seminorm $\|u\|_B$ is finite if and only if the norm $\|u\|_{BP}$ is finite. Usually, in the single-sided time axis case, it will be more natural to replace $2n+1$ with $n+1$ in the above definitions, since we take summation from 0 to n (see Mäkilä and Partington (1996), where this was done for *BP*). The new seminorms are uniformly equivalent to the corresponding old ones.

In a certain sense, a more appropriate seminorm to take on the space of bounded power signals would be

$$\|u\|_{BL} = \left(\mathrm{Blim}_{n\to\infty} \frac{1}{2n+1} \sum_{k=-n}^{n} |u_k|^2 \right)^{1/2} , \tag{2.4}$$

where Blim denotes any Banach or generalized limit (see Rudin (1973)). However, the existence of such limits is not constructive, and this definition would be useless for most practical purposes, so we reject it.

THEOREM 2.1. *Let $u = \{u_k\}$ be a (double-sided) sequence of complex (or real) numbers.*

(a) *If $\|u\|_{BP} < \infty$, then*

$$|u(0)|^2 + \sum_{k\neq 0} \frac{|u(k)|^2}{k^2} \leq 6\|u\|_{BP}^2 . \tag{2.5}$$

(b) *If $\|u\|_B < \infty$, then*

$$|u(0)|^2 + \sum_{k\neq 0} \frac{|u(k)|^2}{k^2} < \infty . \tag{2.6}$$

Furthermore, there does not exist any bound of the form

$$|u(0)|^2 + \sum_{k\neq 0} \frac{|u(k)|^2}{k^2} \leq C\|u\|_B^2 \tag{2.7}$$

for some constant $C > 0$.

Proof. Consider part (a). Write

$$|u(0)|^2 + \sum_{k=-n,k\neq 0}^{n} \frac{|u(k)|^2}{k^2} = \sum_{k=1}^{n} \frac{|u(k)|^2 + |u(-k)|^2 + \delta(k-1)|u(0)|^2}{k^2}, \tag{2.8}$$

where $\delta(\cdot)$ is the Kronecker function, i.e., $\delta(i) = 1$ if $i = 0$, and $\delta(i) = 0$ otherwise. Now apply Abel's partial summation formula (Apostol (1967)). This gives

$$\begin{aligned} &|u(0)|^2 + \sum_{k=-n,k\neq 0}^{n} \frac{|u(k)|^2}{k^2} \\ &\quad = \frac{1}{(n+1)^2} \sum_{k=-n}^{n} |u(k)|^2 + \sum_{k=1}^{n} \sum_{l=-k}^{k} |u(l)|^2 \left(\frac{1}{k^2} - \frac{1}{(k+1)^2} \right) \\ &\quad \leq \frac{2n+1}{(n+1)^2} \frac{1}{2n+1} \sum_{k=-n}^{n} |u(k)|^2 + \sum_{k=1}^{n} \left(\frac{1}{2k+1} \sum_{l=-k}^{k} |u(l)|^2 \right) \frac{2(2k+1)}{k(k+1)^2} \\ &\quad \leq \frac{2n+1}{(n+1)^2} \|u\|_B^2 + 2 \left(\sum_{k=1}^{n} \frac{2k+1}{k(k+1)^2} \right) \|u\|_B^2 \leq 6\|u\|_B^2 \end{aligned} \tag{2.9}$$

for sufficiently large n. This proves part (a). We could prove part (b) by first noting that $\|u\|_B < \infty$ implies that $\|u\|_{BP} < \infty$ and then rely on part (a). Or, independently of this, we could split the sum

$$|u(0)|^2 + \sum_{k=-n,k\neq 0}^{n} \frac{|u(k)|^2}{k^2} \tag{2.10}$$

into two parts, the latter part being

$$\sum_{m<|k|\leq n} \frac{|u(k)|^2}{k^2}, \tag{2.11}$$

choosing n and m sufficiently large, applying Abel's partial summation formula to the latter part, so that one does not need a uniform bound on

$$\frac{1}{2n+1} \sum_{k=-n}^{n} |u_k|^2, \tag{2.12}$$

and estimate terms otherwise in an analogous manner to part (a). This idea of proof makes it plausible that no bound analogous to the bound in part (a) can be found. In fact, as we can find u with $u(0) \neq 0$ but $\|u\|_B = 0$, it follows that the latter claim of part (b) is true. This completes the proof of the theorem.

Wiener (1927) gives a weaker form of part (a) of the above result. Using this result it follows that if $\{u_k\}$ satisfies the condition $\|u\|_B < \infty$ (or $\|u\|_{BP} < \infty$), then

$$\{u_0, u(\pm 1)/(\pm 1), u(\pm 2)/(\pm 2), \ldots, u(\pm k)/(\pm k), \ldots\} \tag{2.13}$$

is a square summable sequence, and hence the function $s(\omega)$ defined by

$$s(\omega) = \frac{1}{2\pi} u(0)\omega + \frac{1}{2\pi} \sum_{k\neq 0} \frac{u(k)}{-ik} e^{-ik\omega}, \qquad \omega \in [-\pi, \pi] \tag{2.14}$$

is square integrable on $[-\pi,\pi]$, i.e., $s \in L_2(-\pi,\pi)$. Note that (2.14) also defines s outside the interval $[-\pi,\pi]$.

We have the following result. (Wiener (1949) mentions this result in a more general form without proof. We add here a proof for completeness.)

THEOREM 2.2. *Let $u = \{u_k\}$ be such that $\|u\|_A < \infty$. Then*

$$\|u\|_A^2 = \lim_{\epsilon \to 0} \frac{1}{2\epsilon} \int_{-\pi}^{\pi} |s(\omega+\epsilon) - s(\omega-\epsilon)|^2 \, d\omega. \tag{2.15}$$

Proof. We have that

$$s(\omega+\epsilon) - s(\omega-\epsilon) = \frac{1}{\pi} u(0)\epsilon - \frac{i}{\pi} \sum_{k \neq 0} u(k) e^{-ik\omega} \frac{\sin k\epsilon}{k}. \tag{2.16}$$

By Parseval's theorem

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |s(\omega+\epsilon) - s(\omega-\epsilon)|^2 d\omega = \left(\frac{1}{\pi} u(0)\epsilon\right)^2 + \left(\frac{1}{\pi}\right)^2 \sum_{k \neq 0} |u(k)|^2 \left(\frac{\sin k\epsilon}{k}\right)^2. \tag{2.17}$$

Define the continuous-time signal $u(t)$ as $u(t) = u([t])$, where $[t]$ denotes the largest integer satisfying $[t] \leq t$. We now proceed to show that

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \sum_{k \neq 0} |u(k)|^2 \left(\frac{\sin k\epsilon}{k}\right)^2 = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_{-\infty}^{\infty} |u(t)|^2 \left(\frac{\sin \epsilon t}{t}\right)^2 dt. \tag{2.18}$$

Note that we can neglect here in the left-hand side (LHS) sum the terms corresponding to $k = -1, 0$, and in the right-hand side (RHS) integral the part $\int_{-1}^{1}$, as these tend to zero when $\epsilon \to 0$. We thus need to estimate (taking $\epsilon > 0$ to simplify the notation)

$$\begin{aligned}
&\left| \frac{1}{\epsilon} \sum_{k \neq -1,0} |u(k)|^2 \left(\frac{\sin k\epsilon}{k}\right)^2 - \frac{1}{\epsilon} \int_{|t|>1} |u(t)|^2 \left(\frac{\sin \epsilon t}{t}\right)^2 dt \right| \\
&\leq \frac{1}{\epsilon} \sum_{k \neq -1,0} |u(k)|^2 \int_k^{k+1} \left| \left(\frac{\sin \epsilon k}{k}\right)^2 - \left(\frac{\sin \epsilon t}{t}\right)^2 \right| dt \\
&\leq \frac{1}{\epsilon} \sum_{k \neq -1,0}^{-n \leq k \leq n} |u(k)|^2 \int_k^{k+1} + \frac{1}{\epsilon} \sum_{|k|>n} |u(k)|^2 \int_k^{k+1},
\end{aligned} \tag{2.19}$$

where $n > 1$ and the RHS expression is divided into two separate parts to facilitate the remaining size estimation. Substitute in the integrals above $t = k + x$, $0 \leq x \leq 1$. It is convenient to estimate (for $k \neq -1, 0$)

$$\begin{aligned}
&\int_k^{k+1} \left| \left(\frac{\sin \epsilon k}{k}\right)^2 - \left(\frac{\sin \epsilon t}{t}\right)^2 \right| dt \\
&= \int_0^1 \frac{|(k+x)^2 \sin^2 k\epsilon - k^2 \sin(k+x)\epsilon|}{k^2(k+x)^2} dx \\
&\leq \int_0^1 \frac{(x^2 + 2kx) \sin^2 k\epsilon}{k^2(k+x)^2} dx + \int_0^1 \frac{\sin^2 x\epsilon \sin^2 k\epsilon}{(k+x)^2} dx \\
&\quad + \frac{1}{2} \int_0^1 \frac{|\sin 2k\epsilon \sin 2x\epsilon|}{(k+x)^2} dx + \int_0^1 \frac{\cos^2 k\epsilon \sin^2 x\epsilon}{(k+x)^2} dx.
\end{aligned} \tag{2.20}$$

Now we continue estimating terms in (2.19). Consider first the term with $|k| \leq n$. Using (2.20) and the relationships $|\sin y| \leq |y|$ (for any real y) and $|k+x| \geq 1$, we then get

$$\frac{1}{\epsilon} \sum_{k \neq -1,0}^{-n \leq k \leq n} |u(k)|^2 \int_k^{k+1} \left| \left(\frac{\sin \epsilon k}{k} \right)^2 - \left(\frac{\sin \epsilon t}{t} \right)^2 \right| dt$$

$$\leq \epsilon \left[\frac{1}{3} + n + n^2 + \frac{1}{2} n + \frac{1}{3} \right] \sum_{k=-n}^{n} |u(k)|^2. \tag{2.21}$$

Similarly, we get for the term with $|k| > n$ in (2.19), taking $n > 10$

$$\frac{1}{\epsilon} \sum_{|k|>n} |u(k)|^2 \int_k^{k+1} \left| \left(\frac{\sin \epsilon k}{k} \right)^2 - \left(\frac{\sin \epsilon t}{t} \right)^2 \right| dt \leq (2+\epsilon) \sum_{|k|>n} \frac{|u(k)|^2}{k^2} \to 0 \tag{2.22}$$

when $n \to \infty$ by Theorem 2.1. Hence this, (2.21), and (2.19) imply the relationship (2.18). Observe that

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} |u(t)|^2 \, dt = \lim_{n \to \infty} \frac{1}{2n+1} \sum_{k=-n}^{n} |u(k)|^2 \tag{2.23}$$

by construction. Hence the RHS integral in (2.18) is equal to $\pi/2$ times $\|u\|_A^2$ by an application of Theorem 21 in Wiener (1933). The result then follows by (2.18).

Let us assume for the moment that the sequence $u = \{u_k\}$ is such that the autocorrelation function (or covariance sequence) defined by

$$R_{uu}(\nu) = \lim_{n \to \infty} \frac{1}{2n+1} \sum_{k=-n}^{n} u(k+\nu)\overline{u(k)} \tag{2.24}$$

exists for any $\nu = 0, \pm 1, \pm 2, \ldots$. (This is not always the case, as Example 2.2 below shows.) Note that this is essentially the same definition as used in Wiener (1927).

We shall state here for later reference a necessary condition for the existence of the limit defining $R_{uu}(0)$.

Proposition 2.1. *A necessary condition for $R_{uu}(0)$ to exist is that*

$$\lim_{n \to \pm\infty} \frac{|u(n)|^2}{n} = 0. \tag{2.25}$$

Proof. This is easy, as a necessary condition for the existence of the required limit is given by

$$\lim_{n \to \infty} \left(\frac{1}{2n+3} \sum_{k=-n-1}^{n+1} |u(k)|^2 - \frac{1}{2n+1} \sum_{k=-n}^{n} |u(k)|^2 \right) = 0. \tag{2.26}$$

The result then follows readily.

Let S_A denote the set of all sequences having an autocorrelation function. Note that $R_{uu}(0) = \|u\|_A^2$. Furthermore, let the notation P_X mean the set defined by

$$P_X = \{u \,|\, \|u\|_X < \infty\}, \tag{2.27}$$

where X denotes a signal size symbol, e.g., $X = B$ in the case of the seminorm (2.2). Therefore we have the set inclusions $S_A \subset P_A \subset P_B = P_{BP}$. Note that the space P_B is the discrete analogue of the Marcinkiewicz space (Chen and Lau (1989)); see also section 2.2. Furthermore, P_{BP} is the discrete analogue of an important Banach space usually credited (Chen and Lau (1989)) to Beurling (1964); see also section 2.2.

Introduce for $u \in S_A$ the function

$$F_u(\omega) = \frac{1}{2\pi} R_{uu}(0)\omega + \frac{1}{2\pi} \sum_{k \neq 0} \frac{R_{uu}(k)}{-ik} e^{-ik\omega}, \qquad \omega \in [-\pi, \pi]. \tag{2.28}$$

(Again this is essentially the same definition as used in Wiener (1927).) It is easily seen that $|R_{uu}(k)| \leq R_{uu}(0)$ (using Schwarz's inequality). Hence $\{R_{uu}(k)\}$ is a bounded sequence. It therefore follows that $F_u(\omega)$ is a square integrable function, i.e., $F_u(\omega) \in L_2(-\pi, \pi)$. Following Wiener's terminology (Wiener (1927), (1930), (1933)), F_u will be called the spectrum of u. (The term *spectral distribution function* is often used instead, especially in the stochastic literature.) We just state some fairly standard properties of R_{uu} and F_u.

FACT 2.1. *Let* $u \in S_A$. *Then*

(a) $R_{uu}(-\nu) = \overline{R_{uu}(\nu)}$.

(b) R_{uu} *is positive semidefinite*; *i.e., for any complex sequence* $\{x_i\}$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} x_i \bar{x}_j R_{uu}(i-j) \geq 0, \qquad n = 1, 2, \dots . \tag{2.29}$$

(c) F_u *is real-valued.*

It can be shown (Wiener (1927)) that F_u can be defined to be a nondecreasing function (note that our choice of interval of definition for F_u from $-\pi$ to π means that F_u can obtain both negative and positive values) so that

$$R_{uu}(\nu) = \int_{-\pi}^{\pi} e^{i\nu\omega} dF_u(\omega) \tag{2.30}$$

holds. Then furthermore,

$$R_{uu}(0) = F_u(\pi) - F_u(-\pi). \tag{2.31}$$

Note that the essential property for the existence of the spectrum, i.e., of a nondecreasing function satisfying (2.30), is the positive semidefiniteness of the covariance sequence. This is the same property that shows up in spectral analysis of stationary stochastic processes (Caines (1988)). Hence Herglotz's theorem (Caines (1988)) in the theory of positive semidefinite sequences proves the existence of the spectrum (spectral distribution).

Now (2.31) shows that it is possible to recover the signal size $\|u\|_A$ from the spectrum in the discrete case (the continuous case is more difficult). However, it is common to use the so-called spectral density function $f_u(\omega)$ introduced as

$$f_u(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} R_{uu}(k) e^{-ik\omega}, \qquad -\pi \leq \omega \leq \pi \tag{2.32}$$

in spectral analysis. The recovery of the signal size $\|u\|_A$ from the spectral density function f_u is, however, a rather technical issue. The fact below illustrates that smoothness of the spectral density function guarantees recovery of the signal size $\|u\|_A$ from f_u.

FACT 2.2. *Let $u \in S_A$. If $\sum_{k=-\infty}^{\infty} |R_{uu}(k)| < \infty$, then the spectral density $f_u(\omega)$ is a continuous function and*

$$\int_{-\pi}^{\pi} f_u(\omega) d\omega = R_{uu}(0). \tag{2.33}$$

Unfortunately, in general

$$R_{uu}(0) = F_u(\pi) - F_u(-\pi) \geq \int_{-\pi}^{\pi} f_u(\omega) d\omega, \tag{2.34}$$

where strict inequality applies if F_u contains a singular part (i.e., a part that is continuous but has vanishing derivative almost everywhere) (Kolmogorov and Fomin (1970), Wiener (1927)), even if jumps in F_u are accounted for in f_u by allowing f_u to include impulses (Dirac delta functions).

We thus see that it need not be possible to recover the signal size correctly ($R_{uu}(0)$ is equal to $\|u\|_A^2$) from the spectral density f_u. This provides yet another motivation for studying alternative formulations of bounded power–like signals.

Input-output analysis for LTI (causal) systems is the topic of the next result with S_A as the signal set.

THEOREM 2.3. *Let G be a causal LTI system with (unit) impulse response $\{g(k)\}_{k \geq 0}$. Let $y = Gu$ denote the output of G to the input u defined by $y(t) = \sum_{k \geq 0} g(k) u(t-k)$. Let $u \in S_A$ and suppose that*

$$\sum_{k \geq 0} k^{1/2} |g(k)| < \infty. \tag{2.35}$$

Then $y \in S_A$, and the covariance sequence of the output y is given by

$$R_{yy}(\nu) = \sum_{k \geq 0, l \geq 0} g(k) \overline{g(l)} R_{uu}(\nu - k + l), \qquad \nu = 0, \pm 1, \pm 2, \dots . \tag{2.36}$$

Furthermore, $R_{yy}(\nu)$ can be expressed in terms of the spectrum F_u of u as

$$R_{yy}(\nu) = \int_{-\pi}^{\pi} e^{i\nu\omega} |G(e^{-i\omega})|^2 \, dF_u(\omega), \tag{2.37}$$

where $G(z) = \sum_{k \geq 0} g(k) z^k$.

Remark 2.1. Note that it is easy to see that condition (2.35) cannot be relaxed here to $\sum_{k \geq 0} |g(k)| < \infty$; cf. Mäkilä and Partington (1996). The condition of *strict stability*, i.e.,

$$\sum_{k \geq 0} k |g(k)| < \infty \tag{2.38}$$

(see Ljung (1987) for the terminology) would at first inspection appear more natural than condition (2.35) here because we get for an arbitrary $u \in S_A$, by an application of Schwarz's inequality,

$$\begin{aligned} |y(t)| &\leq \sum_{k \geq 0} |g(k)| |u(t-k)| \\ &\leq |g(0)||u(t)| + \left(\sum_{k \geq 1} k^2 |g(k)|^2 \right)^{1/2} \left(\sum_{k \geq 1} \frac{|u(t-k)|^2}{k^2} \right)^{1/2}, \end{aligned} \tag{2.39}$$

where the RHS is finite by Theorem 2.1 and by (2.38) (this is easy to verify). Hence strict stability guarantees that the output y is well defined. Obviously strict stability implies (2.35). Note that strict stability implies that the frequency response $G(e^{i\theta})$ of the system is continuously differentiable, while (2.35) does not even imply differentiability of $G(e^{i\theta})$.

Proof of Theorem 2.3. Introduce the quantity

$$r_{yy}(\nu) = \sum_{k\geq 0, l\geq 0} g(k)\overline{g(l)} R_{uu}(\nu - k + l), \qquad \nu = 0, \pm 1, \pm 2, \dots . \tag{2.40}$$

Now $r_{yy}(\nu)$ clearly exists for any ν as $\{g(k)\}$ is absolutely summable and $\{R_{uu}(\tau)\}$ is a bounded sequence. We now show that the output covariance sequence $\{R_{yy}(\nu)\}$ is given by $\{r_{yy}(\nu)\}$ when (2.35) holds. We can take

$$\|G\|_1 \equiv \sum_{k\geq 0} |g(k)| > 0, \tag{2.41}$$

as otherwise there is nothing to prove. We estimate

$$\begin{aligned}
&\left| \frac{1}{2n+1} \sum_{t=-n}^{n} y(t+\nu)\overline{y(t)} - r_{yy}(\nu) \right| \\
&\quad = \left| \sum_{k\geq 0, l\geq 0} g(k)\overline{g(l)} \left[\frac{1}{2n+1} \sum_{t=-n}^{n} u(t+\nu-k)\overline{u(t-l)} - R_{uu}(\nu-k+l) \right] \right| \\
&\quad \leq \left| \sum_{k=0,l=0}^{k=N,l=N} g(k)\overline{g(l)} \left[\frac{1}{2n+1} \sum_{t=-n}^{n} u(t+\nu-k)\overline{u(t-l)} - R_{uu}(\nu-k+l) \right] \right| \\
&\quad + \left| \sum_{\max(k,l)>N} g(k)\overline{g(l)} \left[\frac{1}{2n+1} \sum_{t=-n}^{n} u(t+\nu-k)\overline{u(t-l)} - R_{uu}(\nu-k+l) \right] \right| .
\end{aligned} \tag{2.42}$$

Consider the last term in (2.42). Clearly

$$\sum_{\max(k,l)>N} g(k)\overline{g(l)} R_{uu}(\nu-k+l) \to 0 \tag{2.43}$$

when $N \to \infty$ (recall that $|R_{uu}(\nu-k+l)| \leq R_{uu}(0)$). Furthermore, by Schwarz's inequality

$$\begin{aligned}
&\left| \sum_{\max(k,l)>N} g(k)\overline{g(l)} \frac{1}{2n+1} \sum_{t=-n}^{n} u(t+\nu-k)\overline{u(t-l)} \right| \\
&\quad \leq \sum_{\max(k,l)>N} |g(k)||g(l)| \left(\frac{1}{2n+1} \sum_{t=-n}^{n} |u(t+\nu-k)|^2 \right)^{1/2} \\
&\quad \times \left(\frac{1}{2n+1} \sum_{t=-n}^{n} |u(t-l)|^2 \right)^{1/2}
\end{aligned}$$

$$\leq \sum_{\max(k,l)>N} |g(k)||g(l)| \left(\frac{2(n+|\nu|+k)+1}{2n+1}\right)^{1/2} \left(\frac{2(n+l)+1}{2n+1}\right)^{1/2}$$

$$\times \left(\frac{1}{2(n+|\nu|+k)+1} \sum_{t=-n-|\nu|-k}^{n+|\nu|+k} |u(t)|^2\right)^{1/2}$$

$$\times \left(\frac{1}{2(n+l)+1} \sum_{t=-n-l}^{n+l} |u(t)|^2\right)^{1/2}. \tag{2.44}$$

Now the RHS of the last inequality above tends to zero when $N \to \infty$ by (2.35) and as

$$\frac{1}{2n+1} \sum_{t=-n}^{n} |u(t)|^2 \tag{2.45}$$

tends to $R_{uu}(0)$ (i.e., to a finite limit) when $n \to \infty$.

It therefore remains only to show that the first term on the RHS of the inequality (2.42) tends to zero when n tends to infinity for any finite N, for the first part of the theorem to follow. Hence we need to estimate

$$\left| \sum_{k=0,l=0}^{k=N,l=N} g(k)\overline{g(l)} \left[\frac{1}{2n+1} \sum_{t=-n}^{n} u(t+\nu-k)\overline{u(t-l)} - R_{uu}(\nu-k+l) \right] \right|$$

$$\leq \left| \sum_{k,l=0}^{k,l=N} g(k)\overline{g(l)} \left[\frac{1}{2n+1} \left(\sum_{t=-n}^{n} u(t+\nu-k)\overline{u(t-l)} \right.\right.\right.$$

$$\left.\left.\left. - \sum_{\tau=-n}^{n} u(\tau+\nu-k+l)\overline{u(\tau)} \right) \right] \right|$$

$$+ \left| \sum_{k,l=0}^{k,l=N} g(k)\overline{g(l)} \left[\frac{1}{2n+1} \sum_{t=-n}^{n} u(t+\nu-k+l)\overline{u(t)} - R_{uu}(\nu-k+l) \right] \right|. \tag{2.46}$$

Consider the last term above. Let $\epsilon > 0$ be given. As $u \in S_A$ it follows that there exists a positive integer $n(\epsilon, N, \nu)$ such that

$$\left| \frac{1}{2n+1} \sum_{t=-n}^{n} u(t+\nu-k+l)\overline{u(t)} - R_{uu}(\nu-k+l) \right|$$

$$\leq \epsilon/(2\|G\|_1^2), \quad 0 \leq k, l \leq N, \tag{2.47}$$

for any $n \geq n(\epsilon, N, \nu)$. Letting $n \geq n(\epsilon, N, \nu)$ we thus get

$$\left| \sum_{k,l=0}^{k,l=N} g(k)\overline{g(l)} \left[\frac{1}{2n+1} \sum_{t=-n}^{n} u(t+\nu-k+l)\overline{u(t)} - R_{uu}(\nu-k+l) \right] \right|$$

$$\leq \sum_{k,l=0}^{k,l=N} |g(k)||g(l)|\epsilon/(2\|G\|_1^2) \leq \epsilon/2. \tag{2.48}$$

Consider now the first term on the RHS of the inequality (2.46). Take $n_1 > \min[N + |\nu|, n(\epsilon, N, \nu)]$ such that

$$|u(n)|^2 \leq \frac{\epsilon\, n}{2\|G\|_1^2 N(N+|\nu|)} \tag{2.49}$$

for any $n \geq n_1$. This is possible by Proposition 2.1. Take $n \geq n_1$. Then

$$\begin{aligned}&\left|\sum_{t=-n}^{n} u(t+\nu-k)\overline{u(t-l)} - \sum_{\tau=-n}^{n} u(\tau+\nu-k+l)\overline{u(\tau)}\right| \\ &\qquad \leq l\,\{\epsilon[n+|\nu|+\max(k,l)]\}/[2\|G\|_1^2 N(N+|\nu|)].\end{aligned} \tag{2.50}$$

Therefore, finally,

$$\begin{aligned}&\left|\sum_{k,l=0}^{k,l=N} g(k)\overline{g(l)}\left[\frac{1}{2n+1}\left(\sum_{t=-n}^{n} u(t+\nu-k)\overline{u(t-l)} - \sum_{\tau=-n}^{n} u(\tau+\nu-k+l)\overline{u(\tau)}\right)\right]\right| \\ &\qquad \leq \sum_{k,l=0}^{k,l=N} |g(k)||g(l)|[\epsilon\, N(n+|\nu|+N)]/[2\|G\|_1^2 N(N+|\nu|)(2n+1)] \\ &\qquad \leq \|G\|_1^2[\epsilon\, N(|\nu|+N)]/[2\|G\|_1^2 N(N+|\nu|)] \leq \epsilon/2.\end{aligned} \tag{2.51}$$

As $\epsilon > 0$ is arbitrary, it follows that it is possible to make the RHS of the inequality in (2.46) as small as one likes by choosing n large enough. As this conclusion holds for any N, this concludes the proof of the first part of the theorem; i.e.,

$$R_{yy}(\nu) = \lim_{n\to\infty} \frac{1}{2n+1}\sum_{t=-n}^{n} y(t+\nu)\overline{y(t)} \tag{2.52}$$

exists and is given by $r_{yy}(\nu)$. The latter part of the theorem, i.e., (2.37), follows readily by properties of Lebesgue–Stieltjes integrals by directly computing

$$\begin{aligned}\int_{-\pi}^{\pi} e^{i\nu\omega}|G(e^{-i\omega})|^2 dF_u &= \int_{-\pi}^{\pi} e^{i\nu\omega}\sum_{k,l\geq 0} g(k)\overline{g(l)}e^{-ik\omega}e^{il\omega}dF_u \\ &= \sum_{k,l\geq 0} g(k)\overline{g(l)}\int_{-\pi}^{\pi} e^{i(\nu-k+l)\omega}\,dF_u \\ &= \sum_{k,l\geq 0} g(k)\overline{g(l)}R_{uu}(\nu-k+l) = R_{yy}(\nu),\end{aligned} \tag{2.53}$$

where we have used absolute summability of $\{g(k)\}$, which implies continuity of the transfer function $G(e^{-i\omega})$, (2.30), and finally (2.36).

Theorem 2.3 shows that it is the set S_A, not the larger set P_A, that is natural in input-output analysis when the limit operation is used in the signal set definition. Note that if we restrict the input to the intersection $S_A \cap \ell_\infty(\mathbb{Z})$, the condition of strict stability can be relaxed to the condition $\sum_{k\geq 0}|g(k)| < \infty$. This we see easily following the steps of the proof above; cf. also Theorem 2.2 in Ljung (1987). Ljung (1987) defines a setup of quasi-stationary signals which reduces to the single-sided time axis variant of $S_A \cap \ell_\infty(\mathbb{Z})$ (Ljung (1987) considers the single-sided time axis case) when deterministic signal modeling is used only.

Theorem 2.3 has the following important corollary.

COROLLARY 2.1. *Let the conditions of Theorem* 2.3 *hold. Then*

$$\|G\|_{S_A} \equiv \sup_{u \in S_A, \|u\|_A \neq 0} \frac{\|Gu\|_A}{\|u\|_A} = \|G\|_\infty, \tag{2.54}$$

where $\|G\|_\infty = \sup_\omega |G(e^{-i\omega})|$.

This result is obvious from Theorem 2.3. It guarantees that the H_∞ norm of the system transfer function gives the induced system gain in S_A under mild assumptions, e.g., for strictly stable systems and hence for exponentially stable systems specifically. Note that this result is obtained with the help of the machinery of Lebesgue–Stieltjes integrals. In the literature (Boyd and Barratt (1991); Doyle, Francis, and Tannenbaum (1992); Zhou et al. (1994)) similar results have been derived for stable finite dimensional systems in some subset of S_A only, e.g., for signals such that $R_{uu}(0) = \int_{-\pi}^{\pi} f_u(\omega)\, d\omega$.

Remark 2.2. An equivalent problem to that of determining the $\|\cdot\|_A$ gain is the following. Let $v_m = [g(0), \ldots, g(m)]^T$ and consider the optimization problem

$$\sup_T \frac{v_m^H T v_m}{T_{11}}, \tag{2.55}$$

where the symbol H denotes complex conjugate transpose and where the supremum is taken over all positive semidefinite Hermitian Toeplitz matrices with $T_{11} > 0$. This supremum is given by the square of the H_∞ norm of $G_m(z) = \sum_{k=0}^{m} g(k) z^k$, since the supremization problem is equivalent to the problem of determining $\|G_m\|_A^2$. (It suffices to observe that a finite segment of the covariance sequence of $u \in S_A$ generates a positive semidefinite Hermitian Toeplitz matrix and to each such positive semidefinite Hermitian Toeplitz matrix there corresponds a $\tilde{u} \in S_A$, e.g., by standard results from the theory of stationary stochastic processes (Karlin and Taylor (1975)).)

Let us now study whether we could get linear vector space setups by changing from a quadratic norm to some other norm (as we observed that in the quadratic case we needed cross-correlations (i.e., generalizations, R_{uv}, of the autocorrelation function R_{uu} to signal pairs u and v in a natural way) to exist to get a subset of S_A, which is a linear vector space).

We get the following result.

PROPOSITION 2.2. *Consider signals* $u = \{u_k\}$ *satisfying*

$$\lim_{n\to\infty} \frac{1}{2n+1} \sum_{k=-n}^{n} |u(k)| < \infty. \tag{2.56}$$

There exist signals u *and* v *satisfying* (2.56) *whose sum* $u + v$ *does not have a limit of the form in* (2.56). *Hence, the set defined by condition* (2.56) *is not a linear vector space.*

Proof. Define the sequence $u \in S_A$ as $u(k) = +1$ for all k, and the sequence $v \in S_A$ as $v(0) = -1$, $v(k) = +1$ for $|k| = 2^1 - 1$ to $2(2^1 - 1)$, $v(k) = -1$ for $|k| = 2^2 - 1$ to $2(2^2 - 1)$, $v(k) = +1$ for $|k| = 2^3 - 1$ to $2(2^3 - 1)$, etc. Evaluating

$$s_n(u,v) = \frac{1}{2n+1} \sum_{k=-n}^{n} |u(k) + v(k)| \tag{2.57}$$

for different n, we see that the generated s_n has two points of accumulation (4/3 and 2/3), and so s_n does not tend to any limit when $n \to \infty$. The result follows.

It is easy to generalize this result to other ℓ_p norms. Our conclusion is that the problem is in the use of the limit operation in the definition of the signal set, so the natural remedy is to study more general operations on signals.

The lack of linear vector space structure is a disturbing property of formulations of bounded power signals using the limit operation, and there are some other difficulties as well.

Let $y = Gu$ denote the output of the system G to the input $u = [u_1, \ldots, u_m]^T$ (each u_i is a complex (or real) sequence). Furthermore, let the size of u be measured by some function $\|u\|$ of $\|u_i\|_A$, $i = 1, \ldots, m$, with the property that in the single-variable case this function reduces to the $\| \cdot \|_A$ function. Let the size of y be measured with the same function as u.

FACT 2.3. *There is a causal, stable, finite dimensional, LTI multivariable system G for which $\|u\| < \infty$ does not imply that $\|y\|$ exists, where $y = Gu$.*

To get an example of such a multivariable system it suffices to consider the two-input and one-output system $G = [1, 1]$. The output $y = Gu$ is then given by $y = u_1 + u_2$. Hence the claim of the fact follows as S_A is not a linear vector space.

Note that this result is true for any function $\| \cdot \|$ as defined above. The only way to get around this property would be to require that in order that u be an admissible input signal, the cross-covariances of the elements of u must exist. However, there does not seem to be any good physical reason to make such a restriction. Anyway, in the standard GHA formulation for the multivariable case (Wiener (1949)) the existence of the covariance matrix sequence

$$R_{uu}(\nu) = \lim_{n \to \infty} \frac{1}{2n+1} \sum_{k=-n}^{n} u(k+\nu)\overline{u(k)}^T, \qquad \nu = 0, \pm 1, \pm 2, \ldots \tag{2.58}$$

is required. This is, however, somewhat counterintuitive as a basis for a signal set for robustness analysis, e.g., as the existence of cross-covariances between the components of an input vector means that different components must know what the other components are for them to be admissible (which is not a very natural requirement in many applications, although technically not very restrictive). For example, if $m = 2$ (i.e., the input is two-dimensional) and we take $u_1 = 0$ and $u_2 = v$, where v is defined in the proof of Proposition 2.2, then $u = [u_1, u_2]$ has a covariance matrix sequence, but if $u_1 = 1$ and $u_2 = v$, then u has not, and so this input is not admissible from the GHA point of view.

We shall now consider an example demonstrating that a more serious shortcoming of the standard GHA signal setup for robustness analysis is that, in general, it is not appropriate for non-LTI systems or when non-LTI system perturbations are present.

Example 2.1. Consider the causal linear time-varying system

$$y(t) = \sum_{k \geq 0} g(t,k) u(t-k). \tag{2.59}$$

Let $u \in S_A$ have the constant value 1. Define $g(0,0) = 1$, $g(t,0) = 0$ for $|t| = 2^1 - 1$ to $2(2^1 - 1)$, $g(t,0) = 1$ for $|t| = 2^2 - 1$ to $2(2^2 - 1)$, $g(t,0) = 0$ for $|t| = 2^3 - 1$ to $2(2^3 - 1)$, etc. Let $g(t,k) = 0$ for any $k > 0$ and any t. (Obviously, $g(t,k) = 0$ for

$k < 0$ by causality.) Then the sequence

$$s_n(y) = \frac{1}{2n+1} \sum_{t=-n}^{n} y(t)\overline{y(t)} \tag{2.60}$$

has two points of accumulation (1/3 and 2/3), and so y does not belong to S_A.

2.2. Continuous-time case. Let u denote a real- or complex-valued signal which is locally square integrable on the real axis. Let us specify two signal norms as follows:

$$\|u\|_S = \left(\sup_{T>0} \frac{1}{2T} \int_{-T}^{T} |u(t)|^2 \, dt \right)^{1/2}, \tag{2.61}$$

$$\|u\|_{BP} = \left(\sup_{T\geq 1} \frac{1}{2T} \int_{-T}^{T} |u(t)|^2 \, dt \right)^{1/2}, \tag{2.62}$$

and the seminorm

$$\|u\|_B = \left(\limsup_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} |u(t)|^2 \, dt \right)^{1/2}. \tag{2.63}$$

As in the discrete case, a further seminorm could be defined using Banach limits, but it would not be of practical use.

Wiener (1933) gave several results related to GHA in terms of a boundedness condition on $\| \cdot \|_S$. Using an analogous notation as in the discrete-time case, we let P_X mean the set defined by

$$P_X = \{u \in L^2_{loc}(R) \mid \|u\|_X < \infty\}, \tag{2.64}$$

where X denotes a signal size symbol and $L^2_{loc}(R)$ the set of all locally square integrable functions on the real axis. P_S and P_{BP} are Banach spaces. Many functional analytic results related to these spaces are given in Beurling (1964) and Chen and Lau (1989).

The continuous-time analogue of the B-seminorm, i.e., $\|u\|_B$ as defined above, is discussed in Wiener (1979, p. 334), where Masani refers to the vector space that it defines as the *Marcinkiewicz vector space.* It is worth noting that taking the completion of the space of almost periodic functions under $\| \cdot \|_B$, which is a norm on it, produces some objects which are not functions in the usual sense (cf. Jacob, Larsen, and Zwart (1996); Larsen (1996)). See also Chen and Lau (1989) for further historical remarks and additional references.

The following set inclusions are clearly valid: $P_S \subset P_{BP} = P_B$. We get the following result.

THEOREM 2.4. *Let u be a locally square integrable function on the real axis.*

(a) *If $\|u\|_S < \infty$, then*

$$\int_{-\infty}^{\infty} \frac{|u(t)|^2}{1+t^2} \, dt \leq \pi \|u\|_S^2. \tag{2.65}$$

(b) *If $\|u\|_{BP} < \infty$, then*

$$\int_{-\infty}^{\infty} \frac{|u(t)|^2}{1+t^2} \, dt \leq \left(2 + \frac{\pi}{2}\right) \|u\|_{BP}^2. \tag{2.66}$$

(c) *If* $\|u\|_B < \infty$, *then*

$$\int_{-\infty}^{\infty} \frac{|u(t)|^2}{1+t^2}\, dt < \infty. \tag{2.67}$$

Furthermore, there does not exist any bound of the form

$$\int_{-\infty}^{\infty} \frac{|u(t)|^2}{1+t^2}\, dt \leq C\|u\|_B^2 \tag{2.68}$$

for some constant $C > 0$.

Proof. Note that part (a) above is Theorem 20 in Wiener (1933) with a somewhat tighter bound. Integration by parts gives

$$\int_{-A}^{A} \frac{|u(t)|^2}{1+t^2}\, dt = \frac{1}{1+A^2}\int_{-A}^{A} |u(t)|^2\, dt + \int_0^A \frac{2t}{(1+t^2)^2}\, dt \int_{-t}^{t} |u(\tau)|^2\, d\tau. \tag{2.69}$$

Hence

$$\int_{-A}^{A} \frac{|u(t)|^2}{1+t^2}\, dt \leq \left(\frac{2A}{1+A^2} + \int_0^A \left(\frac{2t}{1+t^2}\right)^2 dt\right)\|u\|_S^2 = 2(\arctan A)\|u\|_S^2 \leq \pi\|u\|_S^2. \tag{2.70}$$

To prove part (b) we partition the interval of integration in the last term in (2.69) into two intervals as follows:

$$\int_0^A \frac{2t}{(1+t^2)^2}\, dt \int_{-t}^{t} |u(\tau)|^2\, d\tau = \int_0^1 + \int_1^A \tag{2.71}$$

for $A > 1$. Then we notice that

$$\int_{-t}^{t} |u(\tau)|^2\, d\tau \leq 2\|u\|_{BP}^2 \tag{2.72}$$

for $0 \leq t \leq 1$. Inserting these relationships in (2.69) and proceeding otherwise as in part (a) gives

$$\int_{-A}^{A} \frac{|u(t)|^2}{1+t^2}\, dt \leq \left(\frac{2A}{1+A^2} + 2 - \frac{\pi}{2} + 2\arctan A - \frac{2A}{1+A^2}\right)\|u\|_{BP}^2 \tag{2.73}$$

from which part (b) follows. Part (c) follows in a rather analogous manner, so we omit the proof here. The nonexistence of a bound of the form indicated follows as there are nontrivial signals with $\|u\|_B = 0$, while the integral on the LHS of the bound is greater than zero.

Introduce the autocovariance function (Wiener (1933)) of u by

$$R_{uu}(\tau) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} u(t+\tau)\overline{u(t)}dt. \tag{2.74}$$

when the indicated limit exists. Let S_A denote the set of all signals u for which $R_{uu}(\tau)$ exists for all real τ. Wiener (1933) has shown that $|R_{uu}(\tau)| \leq R_{uu}(0)$ for all τ for any $u \in S_A$. Let us define the average power signal measure in analogy with the discrete-time case by $\|u\|_A = \sqrt{R_{uu}(0)}$. The following example shows that there exist signals u which have a bounded average power but for which the autocovariance function is not always well defined.

Example 2.2. Let $u(t)$ be constructed to be real and to take the constant value $\epsilon_k = \pm 1$ on each interval of the form $(k, k+1)$, $k \in \mathbb{Z}$, where $\epsilon_{-1} = \epsilon_0 = \epsilon_1 = 1$ and ϵ_k is defined by induction on $|k|$ to satisfy

$$\epsilon_k \epsilon_{k+1} = \begin{cases} +1 & \text{if } 3^{2n} \leq |k| < 3^{2n+1}, \\ -1 & \text{if } 3^{2n+1} \leq |k| < 3^{2n+2} \end{cases} \tag{2.75}$$

for each $n \geq 0$. It is easily checked that u has unit average power but that if $R_{uu}(1)$ existed it would be given by

$$R_{uu}(1) = \lim_{N \to \infty} \frac{1}{2N} \sum_{k=-N}^{N-1} \epsilon_k \epsilon_{k+1}. \tag{2.76}$$

However, this limit does not exist.

This example can clearly be adapted to hold in the discrete-time case as well.

Wiener (1933) has shown that there exists for any $u \in S_A$ a nondecreasing function $\sigma_u(\omega)$ of bounded variation such that almost everywhere

$$R_{uu}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega\tau} d\sigma_u(\omega), \tag{2.77}$$

where the integral should be interpreted as a Lebesgue–Stieltjes integral. (It is worth noting here that the validity of this result does not require that $\|u\|_S < \infty$; it is enough to assume that $\|u\|_B < \infty$. This is important as S_A is not a subset of P_S.) This result follows also by Bochner's theory for positive semidefinite functions (Bochner (1959)). The function σ_u is called the spectrum of u.

It is also possible to introduce the spectral density f_u of $u \in S_A$ by

$$f_u(\omega) = \int_{-\infty}^{\infty} R_{uu}(\tau) e^{-i\omega\tau} d\tau, \qquad \omega \in R. \tag{2.78}$$

The correct interpretation of f_u is a bit technical in the general case but can be given in terms of generalized functions or distributions. Under certain conditions it can simply be interpreted as the derivative of the spectrum $\sigma_u(\omega)$.

The problem of recovery of the average power of a signal in S_A from its spectrum or from its spectral density is highly nontrivial. However, a fairly complete answer has been given by Wiener (1933). We state Wiener's results on this below. Let T_A denote the subset of S_A consisting of those signals having a continuous autocorrelation function. (Wiener (1933) shows that the autocorrelation function is continuous for all τ if it is continuous for $\tau = 0$, a remarkable result!)

FACT 2.4 (Wiener (1933)).

(a) *If $u \in S_A$, then*

$$\sigma_u(\infty) - \sigma_u(-\infty) \leq R_{uu}(0). \tag{2.79}$$

(b) *Let $u \in S_A$. Then*

$$\sigma_u(\infty) - \sigma_u(-\infty) = R_{uu}(0) \tag{2.80}$$

if and only if $u \in T_A$.

Conditions for recovery of the average power of a signal from the power density function are even trickier than those given above. Note that, e.g., in Zhou et al. (1994), it is suggested that one could use the formula

$$R_{uu}(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f_u(\omega)d\omega, \tag{2.81}$$

but unfortunately Fact 2.4 shows that this formula does not apply when the autocorrelation function of R_{uu} lacks continuity.

The problem in general is that the spectrum and the spectral density have smoothed out certain information about the signal u. Let us provide an example of this phenomenon.

Example 2.3. Consider the static nonlinearity $y = G(u) = |u|$. Introduce the system gain by

$$\|G\| = \sup_{u\in S_A, \|u\|_A \neq 0} \frac{\|G(u)\|_A}{\|u\|_A}. \tag{2.82}$$

Clearly $\|G\| = 1$. Now consider what one could conclude by applying formula (2.81). Let us choose the signal $u(t) = e^{it^2} \in S_A$. The autocorrelation function of u is given by (Wiener (1933)) $R_{uu}(0) = 1$, $R_{uu}(\tau) = 0$, elsewhere. But formula (2.81) gives erroneously that $R_{uu}(0) = 0$ as the spectral density function is identical to zero (the same result is obtained by evaluating the total variation of the spectrum). However, $y = G(u)$ is now the constant signal $y(t) = 1$, so its spectrum has a jump of 1 at $\omega = 0$ and its spectral density is the Dirac delta function (i.e., a generalized function). So now both the total variation of the spectrum and formula (2.81) give the correct value 1 for the average power of the output. Hence one could erroneously conclude that the system gain for the static nonlinearity must be unbounded!

We would now like to determine the induced system gain in a GHA (S_A) setup for some fairly large class of continuous-time LTI systems; cf. Corollary 2.1 for the discrete-time case. One such result is as follows. Let G map S_A into S_A. Define the induced system gain

$$\|G\|_{S_A} = \sup_{u\in S_A, \|u\|_A \neq 0} \frac{\|Gu\|_A}{\|u\|_A}. \tag{2.83}$$

Note that we use here (for simplicity) the same notation as in the discrete-time case.

THEOREM 2.5 (Mäkilä (1990)). *Let G be a causal LTI system given by convolution with a* (*unit*) *impulse response function $g(t)$, such that $tg(t)$ is absolutely integrable* (*on the nonnegative reals*) *and such that $(1+t)g(t)$ is square integrable* (*on the nonnegative reals*). *Then*

$$\|G\|_{S_A} = \|G\|_\infty, \tag{2.84}$$

where $\|G\|_\infty = \sup_{Re\, s\geq 0} |G(s)|$ denotes the H_∞ norm of the transfer function $G(s)$ of the system, the latter being defined by

$$G(s) = \int_0^\infty e^{-st} g(t)\, dt. \tag{2.85}$$

Proof. The proof of this result was sketched in Mäkilä (1990). We shall give a slightly more detailed proof here. This result follows from some important results in Wiener (1933). The output y of the system G to the input $u \in S_A$ is defined by

$$y(t) = \int_0^\infty g(\tau)u(t-\tau)\,d\tau. \tag{2.86}$$

By Lemma 29_6 in Wiener (1933), it follows by the assumptions on G that the output y has a continuous autocorrelation function; i.e., it belongs to T_A. This means that the autocorrelation function of y can be exactly recovered from the spectrum σ_y of y, and in particular that $\sigma_y(\infty) - \sigma_y(-\infty) = \|y\|_A^2$. Thus by Fact 2.4 and Theorems 30 and 31 in Wiener (1933)

$$\|G\|_{S_A}^2 \le \sup_{\sigma_u(\infty)-\sigma_u(-\infty)>0} \frac{\int_{-\infty}^{\infty} |G(j\omega)|^2\,d\sigma_u}{\int_{-\infty}^{\infty} d\sigma_u}, \tag{2.87}$$

where $\sigma_u(\omega)$ denotes the spectrum of u. (Note that we can take the supremum here over signals u with spectra having nonzero total variation; as for signals with $\|u\|_A \neq 0$ but having zero spectrum, the signal amplification is zero, i.e., then $\|y\|_A = 0$ under the assumptions of the theorem.) Here the integrals are Lebesgue–Stieltjes integrals. To see how (2.87) can be utilized, observe that, by the Schwarz inequality,

$$\int_0^\infty |g(t)|\,dt \le \left(\int_0^\infty |g(t)|^2(1+t)^2\,dt\right)^{1/2} \left(\int_0^\infty \frac{1}{(1+t)^2}\,dt\right)^{1/2} < \infty. \tag{2.88}$$

Hence g is absolutely integrable on the nonnegative reals, i.e., belongs to $L_1(R_+)$. It follows that the frequency response $G(j\omega)$ is a continuous function tending to zero when $|\omega| \to \infty$ (by the Riemann–Lebesgue lemma). We see then readily that the RHS of (2.87) is bounded from above by the square of the H_∞ norm of the transfer function of G. As there exists a periodic signal u in S_A, with frequency equal to a finite frequency at which $|G(j\omega)|$ attains its maximum, having the signal gain amplification ratio $\|Gu\|_A/\|u\|_A = \|G\|_\infty$, the proof of the theorem is completed.

Remark 2.3. Note that it is not possible to relax the conditions on G to absolute integrability of g on the nonnegative reals. This is seen easily by a counterexample analogous to the discrete-time example given in Mäkilä and Partington (1996). The above result means that the induced system gain in an S_A setup is given by the H_∞ norm of the system transfer function for a large class of systems, including, e.g., LTI systems having an exponentially decaying impulse response function.

3. Generalized bounded power signal setups. In this section we shall consider several generalized bounded power signal setups. There are several subtle issues when analyzing bounded power signals and corresponding induced input-output norms for LTI causal systems. Some of these subtle issues are related to differences between the two-sided (double-sided) time axis case and the single-sided time axis case.

3.1. Discrete-time case. Let F be a causal LTI operator mapping sequences $\{u(k)\}_{k\ge0}$ to $\{y(k)\}_{k\ge0}$.

THEOREM 3.1. *An operator F as above has finite gain with respect to the seminorm $\|\cdot\|_B$ if and only if it is bounded with respect to the BP norm; this happens if and only if F is ℓ_2 bounded and thus corresponds to a transfer function in H_∞.*

Proof. If F is ℓ_2 bounded, then it has finite $\|\cdot\|_B$ gain, since, as in Mäkilä and Partington (1996),

$$\frac{1}{n+1}\sum_{k=0}^{n}|y(k)|^2 \le \|F\|_2^2\frac{1}{n+1}\sum_{k=0}^{n}|u(k)|^2, \tag{3.1}$$

and the result follows on taking the limit superior of each side as $n \to \infty$.

For the converse we show the following "automatic continuity" result. *Any causal LTI operator that maps each one-sided sequence in BP into a sequence in BP is bounded as an operator on the one-sided BP space.* This will give the result, since, if F is $\|\cdot\|_B$ bounded, then F maps BP into BP; then, by Mäkilä and Partington (1996), F is ℓ_2 bounded. The method we use is a standard one and similar to the proof of Theorem 1 of Loy (1974).

Suppose to the contrary that there exist $u^{(n)}$ with $\|u^{(n)}\|_{BP}=1$ but $\|y^{(n)}\|_{BP}\ge n$. We shall establish a contradiction by constructing an input u with bounded power mapping to an output y with unbounded power, so we can certainly assume that each $y^{(n)}$ has bounded power.

We write S for the right shift so that $S(a_0,a_1,a_2,\dots)=(0,a_1,a_2,\dots)$, and consider $u^{(2)}$. There will be a constant K_1 such that the corresponding output y satisfies

$$\left(\frac{1}{K_1}\sum_{k=0}^{K_1-1}|y_k|^2\right)^{1/2} > 1. \tag{3.2}$$

We express this by saying that the power at time K_1-1 is greater than 1.

Now let $N_1>2$, consider $u^{(2)}+S^{K_1}u^{(N_1)}/2$, and estimate the power of the corresponding output y. It is the same as before at time K_1-1, and at time K_1+n it is

$$\left(\frac{1}{K_1+n+1}\sum_{k=0}^{K_1+n}|y_k|^2\right)^{1/2} \ge -p_1+q, \tag{3.3}$$

where p_1 is the output power due to the first input and q is that due to the second one. We assumed that p_1 was finite and we can make $q-p_1>2$ by choosing N_1 large since it requires

$$\left(\frac{n+1}{K_1+n+1}\right)^{1/2}\frac{N_1}{2} > p_1+2, \tag{3.4}$$

which holds, whatever n is, once $(1/(K_1+1))^{1/2}N_1/2 > p_1+2$.

So let K_2 be such that, with input $u^{(2)}+S^{K_1}u^{(N_1)}/2$ the corresponding output has power at least 2 by time K_2-1. Repeat, and now consider $u^{(2)}+S^{K_1}u^{(N_1)}/2+S^{K_2}u^{(N_2)}/4$, where N_2 is large enough that the inequality $(1/(K_2+1))^{1/2}N_2/4 > p_2+3$, where p_2 is the maximum output power due to the first two inputs.

Repeating this way we obtain an input $u^{(2)}+\sum_{j=1}^{\infty}S^{K_j}u^{(N_j)}/2^j$, whose power is bounded but which gives an output with unbounded power. This completes the proof.

Remark 3.1. A natural question here is whether for some $C>0$ there holds an inequality of the form $\|F\|_{BP\to BP}\le C\|F\|_{B\to B}$. For systems with ℓ_1 impulse responses, this is easily seen to be the case, since they map periodic sequences of the form $(u, u, u, u, \ldots)$, where u is a finite-length vector, into sequences of the form $(y_1, y_1 + y_2, y_1 + y_2 + y_3, \ldots)$, where each y_k is a vector of the same length as u and where $\sum \|y_k\|_2 < \infty$. The gain of F with respect to $\| \cdot \|_B$ is now at least

$$\frac{\|y_1\|_2 - \sum_{k=2}^{\infty} \|y_k\|_2}{\|u\|_2}, \tag{3.5}$$

which can be made arbitrarily close to the H_∞ norm of the transfer function corresponding to F, by a suitable choice of u. In general, however, such an inequality is not known to hold, and it is not possible to use the closed graph theorem (cf. Kolmogorov and Fomin (1970)) to deduce the existence of a constant, because the norm $\|F\|_{B \to B}$ is not known to be complete.

We now consider the two-sided case, where our inputs and outputs are defined on $\mathbb{Z}$. Let us define a signal size measure as follows. Let $u = \{u(k)\}$ denote an arbitrary (real or complex) sequence. Consider the set of all signals u defined by the condition

$$\|u\|_C^2 = \limsup_{n \to \infty} \left[\sup_{t_0} \frac{1}{2n+1} \sum_{k=-n+t_0}^{n+t_0} |u(k)|^2 \right] < \infty. \tag{3.6}$$

Introduce the notation $P_C = \{u \mid \|u\|_C < \infty\}$. Our first observation is that the definition in (3.6) can be simplified.

Proposition 3.1. *Whenever the norm $\|u\|_C$ is finite, then the limit superior in the defining equation* (3.6) *can be replaced by an infimum over $n \geq 0$ or a limit as $n \to \infty$.*

Proof. Suppose that $n \geq 0$ and let

$$A = \sup_{t_0} \frac{1}{2n+1} \sum_{k=-n+t_0}^{n+t_0} |u(k)|^2. \tag{3.7}$$

Then, for $m > n$ we can write $2m + 1 = q(2n + 1) + r$, where q and r are integers with $q \geq 1$ and $0 \leq r < 2n + 1$. Now for any $t \in \mathbb{Z}$ we have

$$\frac{1}{2m+1} \sum_{k=-m+t}^{m+t} |u(k)|^2 \leq (q+1) \frac{2n+1}{2m+1} \sup_{t_0} \frac{1}{2n+1} \sum_{k=-n+t_0}^{n+t_0} |u(k)|^2, \tag{3.8}$$

as we see by splitting the block of $2m + 1$ terms into at most $q + 1$ blocks of length $2n + 1$. Thus

$$\begin{aligned} \limsup_{m \to \infty} \left[\sup_{t} \frac{1}{2m+1} \sum_{k=-m+t}^{m+t} |u(k)|^2 \right] &\leq \limsup_{m \to \infty} (q+1) A \frac{2n+1}{2m+1} \\ &\leq \limsup_{q \to \infty} (q+1) A / q = A. \end{aligned}$$

Hence we have "lim sup ≤ inf," and since we always have "inf ≤ lim inf ≤ lim sup," the limit exists and equals the infimum, as required.

We are now ready for the following result.

Theorem 3.2. *The set P_C is a linear vector space, on which $\| \cdot \|_C$ is a seminorm. Furthermore, $u \in P_C$ if and only if $u \in \ell_\infty(\mathbb{Z})$. One has $\|u\|_C = \|u\|_A$ whenever u is periodic.*

Proof. We shall show that $\|\cdot\|_C$ satisfies the triangle inequality in P_C. This is the only nontrivial part to establish that P_C is a linear vector space and that $\|\cdot\|_C$ is a seminorm. Thus we estimate by Schwarz's inequality

(3.9)

$$\begin{aligned}
\|u+v\|_C^2 &= \limsup_{n\to\infty} \sup_{t_0} \frac{1}{2n+1} \sum_{k=-n+t_0}^{n+t_0} |u(k)+v(k)|^2 \\
&\le \limsup_{n\to\infty} \sup_{t_0} \frac{1}{2n+1} \left[\sum_{k=-n+t_0}^{n+t_0} |u(k)|^2 + \sum_{k=-n+t_0}^{n+t_0} |v(k)|^2 + 2 \sum_{k=-n+t_0}^{n+t_0} |u(k)||v(k)| \right] \\
&\le \limsup_{n\to\infty} \left[\sup_{t_0} \frac{1}{2n+1} \sum_{k=-n+t_0}^{n+t_0} |u(k)|^2 + \sup_{t_0} \frac{1}{2n+1} \sum_{k=-n+t_0}^{n+t_0} |v(k)|^2 \right. \\
&\qquad \left. + 2 \sup_{t_0} \frac{1}{2n+1} \left(\sum_{k=-n+t_0}^{n+t_0} |u(k)|^2 \right)^{1/2} \left(\sum_{k=-n+t_0}^{n+t_0} |v(k)|^2 \right)^{1/2} \right] \\
&\le \|u\|_C^2 + \|v\|_C^2 + 2 \limsup_{n\to\infty} \left[\sup_{t_0} \left(\frac{1}{2n+1} \sum_{k=-n+t_0}^{n+t_0} |u(k)|^2 \right)^{1/2} \right. \\
&\qquad \left. \times \sup_{t_0} \left(\frac{1}{2n+1} \sum_{k=-n+t_0}^{n+t_0} |v(k)|^2 \right)^{1/2} \right] \\
&\le \|u\|_C^2 + \|v\|_C^2 + 2\|u\|_C \|v\|_C = \left(\|u\|_C + \|v\|_C \right)^2 .
\end{aligned}$$

Hence P_C is a linear vector space on which $\|\cdot\|_C$ is a seminorm. The validity of the latter part of the theorem is seen as follows. Clearly if $u \in \ell_\infty(\mathbb{Z})$, then $u \in P_C$ as $\|u\|_C \le \|u\|_\infty$ $(= \sup_t |u(t)|)$ for any $u \in \ell_\infty(\mathbb{Z})$. Suppose now that $u \in P_C$ but that $u \notin \ell_\infty(\mathbb{Z})$. Clearly if $u \notin \ell_\infty(\mathbb{Z})$, then

$$\sup_{t_0} \frac{1}{2n+1} \sum_{k=-n+t_0}^{n+t_0} |u(k)|^2 = \infty \tag{3.10}$$

so that $u \notin P_C$, which is a contradiction. Hence $u \in P_C$ must imply that $u \in \ell_\infty(\mathbb{Z})$.

Finally, let u be a periodic signal, with period p, say. Then it is easily seen that

$$\|u\|_A^2 = \frac{1}{p} \sum_{j=1}^{p} |u(j)|^2. \tag{3.11}$$

Suppose now that $2n+1 = pq + r$, where q and r are nonnegative integers with $0 \le r < p$. Then summing any $2n+1$ consecutive values $|u(k)|^2$ gives a total of at least $q \sum_{j=1}^{p} |u(j)|^2$ and at most $(q+1) \sum_{k=1}^{p} |u(k)|^2$. Thus

$$\frac{q}{2n+1} \sum_{j=1}^{p} |u(j)|^2 \le \sup_{t_0} \frac{1}{2n+1} \sum_{k=-n+t_0}^{n+t_0} |u(k)|^2 \le \frac{q+1}{2n+1} \sum_{j=1}^{p} |u(j)|^2. \tag{3.12}$$

The result now follows, since $q/(2n+1)$ and $(q+1)/(2n+1)$ both tend to $1/p$ as $n \to \infty$.

If G takes P_C into P_C we may introduce the system gain, induced by $\|\cdot\|_C$, as follows:

$$\|G\|_C = \sup_{u \in P_C, \|u\|_C \neq 0} \frac{\|Gu\|_C}{\|u\|_C} \tag{3.13}$$

when it exists.

It is known (see Loy (1974), Maté (1989)) that any shift-invariant linear system on $\mathbb{Z}_+$ that takes bounded inputs to bounded outputs is automatically continuous as an operator on ℓ_∞, and hence it has an impulse response in ℓ_1. For systems on $\mathbb{Z}$, such an automatic continuity result does not hold, even for causal operators. We have not found this stated explicitly anywhere in the literature, but it follows from the observation that the closure of the set of all vectors of the form $u - Su$ in $\ell_\infty(\mathbb{Z}_+)$ is a subspace Y of infinite codimension which contains $c_0(\mathbb{Z}_+)$. The quotient space $X = \ell_\infty(\mathbb{Z}_+)/Y$ therefore supports a discontinuous linear functional ϕ, and then defining $(Fu)(k) = \phi(P_X Ru)$, where

$$(Ru)(k) = \begin{cases} u(-k) & \text{if } k \geq 0, \\ 0 & \text{if } k < 0 \end{cases} \tag{3.14}$$

and $P_X : \ell_\infty(\mathbb{Z}_+) \to X$ is the projection, produces a discontinuous operator which is easily seen to be causal since it is zero on all sequences u such that $u(k) \to 0$ as $k \to -\infty$. Similar ideas can be found in the book of Sinclair (1976).

Returning to P_C we consider the case with signals defined on $\mathbb{Z}$. Let G be a bounded input bounded output stable, causal, LTI system. To avoid pathological situations such as those given above we shall assume now that G is defined by a convolution kernel such that $\|G\|_1 \equiv \sum_{k\geq 0} |g(k)| < \infty$. Consider G as an operator from P_C into P_C. (Here $g = \{g(k)\}$ denotes the unit impulse response of G.) Then

$$\|G\|_\infty \leq \|G\|_C \leq \|G\|_1, \tag{3.15}$$

where $\|G\|_\infty$ denotes the H_∞ norm of the transfer function $G(z) = \sum_{k\geq 0} g(k) z^k$ of G, i.e., $\|G\|_\infty = \sup_{|z|<1} |G(z)|$.

This is seen as follows. Let $u \in P_C$. Let $y = Gu$. We estimate, using Schwarz's inequality,

(3.16)

$$\begin{aligned} \frac{1}{2n+1} \sum_{t=-n+t_0}^{n+t_0} |y(t)|^2 &= \frac{1}{2n+1} \sum_{t=-n+t_0}^{n+t_0} \left| \sum_{k\geq 0} g(k) u(t-k) \right|^2 \\ &\leq \frac{1}{2n+1} \sum_{k,l\geq 0} |g(k)\overline{g(l)}| \sum_{t=-n+t_0}^{n+t_0} |u(t-k)\overline{u(t-l)}| \\ &\leq \sum_{k,l\geq 0} |g(k)||g(l)| \left(\frac{1}{2n{+}1} \sum_{t=-n+t_0}^{n+t_0} |u(t{-}k)|^2 \right)^{1/2} \left(\frac{1}{2n{+}1} \sum_{t=-n+t_0}^{n+t_0} |u(t{-}l)|^2 \right)^{1/2}. \end{aligned}$$

This gives

$$\begin{aligned} \|y\|_C^2 &\leq \limsup_{n\to\infty} \sum_{k,l\geq 0} |g(k)||g(l)| \\ &\qquad \times \left[\sup_{t_0} \left(\frac{1}{2n+1} \sum_{t=-n+t_0}^{n+t_0} |u(t)|^2 \right)^{1/2} \sup_{t_0} \left(\frac{1}{2n+1} \sum_{t=-n+t_0}^{n+t_0} |u(t)|^2 \right)^{1/2} \right] \\ &= \|G\|_1^2 \|u\|_C^2. \end{aligned} \tag{3.17}$$

This in turn gives the RHS inequality in (3.15). The LHS inequality in (3.15) is obtained by observing that $y(t) = G(e^{-i\omega})u(t)$ for $u(t) = e^{i\omega t}$ so that $\|G\|_C \geq \|G\|_\infty$.

A further argument is needed to establish the following important improvement of (3.15).

THEOREM 3.3. *Let G be a causal LTI system defined by a convolution kernel, satisfying $\|G\|_1 < \infty$. Consider G as an operator from P_C into P_C. Then*

$$\|G\|_C = \|G\|_\infty. \tag{3.18}$$

Proof. Suppose that G has a finite impulse response $(g(0), \ldots, g(m), 0, 0, \ldots)$, say, so that

$$y(t) = \sum_{k=0}^{m} g(k)u(t-k) \tag{3.19}$$

for $t \in \mathbb{Z}$. It will be notationally convenient to suppose that m is even—say, $m = 2p$. For each $t_0 \in \mathbb{Z}$ we have

$$\sum_{k=t_0-n}^{t_0+n} |y(t)|^2 \leq \|G\|_\infty^2 \sum_{k=t_0-n-m}^{t_0+n} |u(k)|^2, \tag{3.20}$$

since the output that would result from the input $u(t_0-n-m), \ldots, u(t_0+n)$ includes the terms $y(t_0-n), \ldots, y(t_0+n)$. Thus

(3.21)

$$\sup_{t_0} \frac{1}{2n+1} \sum_{k=-n+t_0}^{n+t_0} |y(k)|^2 \leq \frac{2n+1+2p}{2n+1} \|G\|_\infty^2 \sup_{t_1} \frac{1}{2n+1+2p} \sum_{k=t_1-(n+p)}^{t_1+n+p} |u(k)|^2,$$

where $t_1 = t_0 - p$. Now, taking limit superiors of both sides of (3.22), as n and $n+p$ go to infinity, we obtain $\|y\|_C^2 \leq \|G\|_\infty^2 \|u\|_C^2$. The general case now follows, since we can approximate the general convolution kernel G by finite impulse response kernels $G^{(m)}$ and then use (3.15) to show that for each input u the corresponding sequence $y^{(m)}$ of outputs converges in the seminorm $\|\cdot\|_C$.

An alternative proof of the above result can also be given, based on Remark 2.2.

Note that in the double-sided axis case the assumption $\|G\|_1 < \infty$ is necessary as otherwise the output $y(t)$ may become infinite even for finite t.

3.2. Continuous-time case. Let F be a causal time-invariant operator mapping functions on $(0, \infty)$ to functions on $(0, \infty)$. The following result is the analogue of the discrete-time theorem proven in Mäkilä and Partington (1996), and the proof is the same.

THEOREM 3.4. *For an operator F as above there is a constant $K > 0$ such that $\|Fu\|_{BP} \leq K\|u\|_{BP}$ for all u such that $\|u\|_{BP} < \infty$, if and only if F is a bounded operator on $L_2(0, \infty)$ and thus corresponds to a transfer function in $H_\infty(\mathbb{C}_+)$. Moreover $\|F\|$ is the same whether we consider F as acting on the space BP or $L_2(0, \infty)$.*

It will also be noted that the proof goes through to yield the similar result for $\|\cdot\|_S$.

The continuous-time version of Theorem 3.1 is also valid, and we state it here for completeness.

THEOREM 3.5. *An operator F as above has finite gain with respect to the seminorm $\| \cdot \|_B$ if and only if it is bounded with respect to the BP norm; this happens if and only if F is L_2 bounded and thus corresponds to a transfer function in $H_\infty(\mathbb{C}_+)$.*

Proof. The proof is the same as that of Theorem 3.1, replacing sums by integrals. For example, if F is L_2 bounded, then

$$\frac{1}{T}\int_0^T |y(t)|^2\,dt \le \|F\|_2^2 \frac{1}{T}\int_0^T |u(t)|^2\,dt, \tag{3.22}$$

and thus F has finite $\| \cdot \|_B$ gain, as we see on taking limit superiors as $T \to \infty$.

Conversely, the automatic continuity result stated in the proof of Theorem 3.1 (*any causal LTI operator that maps each one-sided sequence in BP into a sequence in BP is bounded as an operator on the one-sided BP space*) still holds in continuous time. The same proof holds, with the obvious replacement of sums by integrals. Theorem 3.4 then completes the proof of the equivalence of finite B-gain and finite L_2-gain.

Some recent work on the representation of shift-invariant operators on function spaces on $(0, \infty)$ by transfer functions can be found in Partington and Ünalmış (1997).

Remark 3.2. For systems with L_1 impulse responses we again have the identity

$$\|F\|_{BP\to BP} = \|F\|_{B\to B}, \tag{3.23}$$

as in Remark 3.1. This is because periodic functions of the form $u + S_T u + S_T^2 u + \cdots$, where $u \in L_2(0,T)$ and S_T denotes the right shift by an amount T, are mapped to functions of the form $y_1 + S_T(y_1+y_2) + S_T^2(y_1+y_2+y_3) + \cdots$, where each $y_k \in L_2(0,T)$ and $\sum_{k=1}^\infty \|y_k\|_2 < \infty$. The gain of F with respect to $\| \cdot \|_B$ is now at least

$$\frac{\|y_1\|_2 - \sum_{k=2}^\infty \|y_k\|_2}{\|u\|_2}, \tag{3.24}$$

which can be made arbitrarily close to the H_∞ norm of the transfer function corresponding to T, by a suitable choice of u. Alternatively one could base a proof on the lemma of Amerio and Prouse (1971, p. 72), which asserts that an L_1 convolution kernel maps almost periodic functions to almost periodic functions, and gives a representation of the operator in terms of the corresponding transfer function.

Similarly for the case of signals defined on the whole of $\mathbb{R}$, we can define a seminorm $\|u\|_C$ in continuous time, analogously to equation (3.6). Namely, for locally L_2 functions for which the quantity below is finite, we define

$$\|u\|_C^2 = \limsup_{T\to\infty}\left[\sup_{t_0} \frac{1}{2T}\int_{k=-T+t_0}^{T+t_0} |u(t)|^2\,dt\right]. \tag{3.25}$$

Again the limit superior is actually a limit and equals the infimum over $T > 0$. There is one difference here from the discrete-time case, in that for any function which is in $L_2(\mathbb{R})$, including ones not in $L_\infty(\mathbb{R})$, the quantity given in (3.25) will still be finite.

The proof of (3.15) can be extended in the obvious way to show that in this case operators defined by convolution with L_1 impulse responses give systems with finite gain. It can also be shown, by methods analogous to those used in the proof of Theorem 3.3, that in this case the $\| \cdot \|_C$ gain is equal to the H_∞ norm of the Laplace transform of the impulse response.

4. Conclusions. Bounded power signal spaces due to Wiener, Marcinkiewicz, and Beurling have been studied from a control and systems perspective. Furthermore, a new bounded power signal space has been introduced. Specifically, it has been shown that for a large class of stable finite and infinite dimensional systems, the system power gain is given by the H_∞ norm of the system transfer function. This means that robust H_∞ control can be rigorously motivated based on very general, physically attractive, persistent signal spaces, which allow both stationary and nonstationary signals (nonstationarity meaning here signals with time-dependent spectral properties).

In the double-sided time axis case some care must be taken when using H_∞ methods to analyze robust control and estimation for bounded power signals. It follows from our results that, e.g., in the double-sided time axis case, the causal LTI system G with the H_∞ transfer function $G(z) = z\exp[-(1+z)/(1-z)]$ is not power stable (actually not even stabilizable by any controller with transfer function of the form $C(z) = X(z)/Y(z)$, where $Y \neq 0$ and X and Y are transfer functions of power stable causal LTI systems, by an application of a result in Partington and Mäkilä (1994). It is believed that the bounded power signal spaces studied here will turn out to be important in the active research area of system identification for robust control (see, e.g., Partington (1997)), as they allow for very rich and realistic nonprobabilistic signal models.

Acknowledgments. The authors would like to thank H. G. Dales for helpful discussions on automatic continuity theory.

REFERENCES

L. Amerio and G. Prouse (1971), *Almost-Periodic Functions and Functional Equations*, Van Nostrand, New York.

T. M. Apostol (1967), *Calculus*, Vol I, 2nd ed., Ginn & Blaisdell, Waltham, MA.

A. Beurling (1964), *Construction and analysis of some convolution algebras*, Ann. Inst. Fourier (Grenoble), 14, pp. 229–260.

S. Bochner (1959), *Lectures on Fourier Integrals*, Princeton University Press, Princeton, NJ.

S. P. Boyd and C. H. Barratt (1991), *Linear Controller Design. Limits of Performance*, Prentice-Hall, Englewood Cliffs, NJ.

P. E. Caines (1988), *Linear Stochastic Systems*, John Wiley, New York.

Y.-Z. Chen and K.-S. Lau (1989), *Some new classes of Hardy spaces*, J. Funct. Anal., 84, pp. 255–278.

J. C. Doyle, B. A. Francis, and A. R. Tannenbaum (1992), *Feedback Control Theory*, Macmillan, New York.

W. A. Gardner (1988), *Statistical Spectral Analysis. A Nonprobabilistic Theory*, Prentice-Hall, Englewood Cliffs, NJ.

B. Jacob, M. Larsen, and H. Zwart (1996), *Corrections and extensions of "Optimal control of linear systems with almost periodic inputs" by G. Da Prato and A. Ichikawa*, SIAM J. Control Optim., 36 (1998), pp. 1473–1480.

S. Karlin and H. M. Taylor (1975), *A First Course in Stochastic Processes*, 2nd ed., University of California Press, Berkeley.

A. N. Kolmogorov and S. V. Fomin (1970), *Introductory Real Analysis*, Prentice-Hall, Englewood Cliffs, NJ.

M. Larsen (1996), *$\mathcal{H}_\infty$ Control of Linear Systems with Almost Periodic Inputs*, Technical University of Denmark, Lyngby, Denmark, preprint.

L. Ljung (1987), *System Identification*, Prentice-Hall, Englewood Cliffs, NJ.

R. J. Loy (1974), *Continuity of linear operators commuting with shifts*, J. Funct. Anal., 16, pp. 48–60.

P. M. Mäkilä (1990), *H_∞-optimization and optimal rejection of persistent disturbances*, Automatica J. IFAC, 26, pp. 617–618.

P. M. Mäkilä and J. R. Partington (1996), *Lethargy results in LTI system modelling*, Automatica J. IFAC, to appear.

J. Mari (1996), *A counterexample in power signals space*, IEEE Trans. Automat. Control, AC-41, pp. 115–116.

L. Maté (1989), *On the continuity of causal operators*, Publicationes Math., 36, pp. 191–198.

J. R. Partington (1997), *Interpolation, Identification and Sampling*, Oxford University Press, New York.

J. R. Partington and P. M. Mäkilä (1994), *Worst-case analysis of identification–BIBO robustness for closed-loop data*, IEEE Trans Automat. Control, AC-39, pp. 2171–2176.

J. R. Partington and P. M. Mäkilä (1996), *Modeling of fading memory systems*, IEEE Trans. Automat. Control, AC-41, pp. 899–903.

J. R. Partington and B. Ünalmiş (1997), *On the representation of shift-invariant operators by transfer functions*, Systems Control Lett., 33(1998), pp. 25–30.

W. Rudin (1973), *Functional Analysis*, Tata McGraw-Hill, New Delhi.

A. M. Sinclair (1976), *Automatic Continuity of Linear Operators*, Cambridge University Press, London.

N. Wiener (1927), *The spectrum of an array and its application to the study of the translation properties of a simple class of arithmetical functions, Part* I, J. Math. Phys., 6, pp. 145–157.

N. Wiener (1930), *Generalized harmonic analysis*, Acta Math., 55, pp. 117–258.

N. Wiener (1933), *The Fourier Integral*, Cambridge University Press, London.

N. Wiener (1949), *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, MIT Press, Cambridge, MA.

N. Wiener (1979), *Collected Works with Commentaries*, Vol. II, P. Masani, ed., MIT Press, Cambridge, MA.

G. Zames (1981), *Feedback and optimal sensitivity*: *Model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Automat. Control, AC-26, pp. 301–320.

K. Zhou, J. C. Doyle, and K. Glover (1996), *Robust and Optimal Control*, Prentice-Hall, Upper Saddle River, NJ.

K. Zhou, K. Glover, B. Bodenheimer, and J. Doyle (1994), *Mixed H_2 and H_∞ performance objectives*: *Robust performance analysis*, IEEE Trans. Automat. Control, AC-39, pp. 1564–1587.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 1, pp. 118–130

EQUIVALENCE OF NONLINEAR SYSTEMS TO PRIME SYSTEMS UNDER GENERALIZED OUTPUT TRANSFORMATIONS*

E. ARANDA-BRICAIRE† AND R. M. HIRSCHORN‡

Abstract. Within a linear algebraic framework, we present a new characterization of the class of nonlinear systems which are equivalent to a prime system. We then introduce a class of generalized output transformations that can be thought of as a generalization to the nonlinear setting of a unimodular transformation in the output space. Our main result gives necessary and sufficient conditions for equivalence to a prime system under a certain group of transformations that includes generalized output transformations.

Key words. nonlinear systems, output transformation, prime systems, differential forms

AMS subject classifications. 93C10, 93B17, 58A10

PII. S0363012996307060

1. Introduction.

The problem of characterizing the class of linear systems that are equivalent to prime systems was first posed and solved by Morse [12]. The group of transformations considered in [12] included, besides state space change of coordinates and linear state feedback, output space change of coordinates. Marino, Respondek, and van der Schaft [10] generalized this result to the nonlinear case. They showed that the class of smooth affine nonlinear systems that are locally equivalent to prime systems can be characterized by the properties of two families of involutive distributions defined on the state manifold.

In this paper we consider the problem of equivalence to a prime system under a group of transformations that consist of state space diffeomorphism, regular static state feedback, and generalized output transformation (GOT). In the case where we restrict ourselves to the output space change of coordinates used by Marino, Respondek, and van der Schaft [10], we obtain a new and simpler characterization for the class of nonlinear systems which are (locally) equivalent to prime systems. We then introduce the notion of GOT to identify a larger class of systems equivalent to prime systems. We proceed in two steps. First, we enlarge the output space by considering a finite number of the time derivatives of the output as coordinates of new output space. Then we define a new set of outputs on this enlarged space. This transformation is invertible in the sense that the new outputs can be expressed as functions of a finite number of the time derivatives of the original output, and vice versa. That is, we can recover the original output without any integration. By analogy with the linear case, these transformations could be called unimodular, in the sense that their inverse belongs to the same class of transformations. This approach finds a natural application in control problems such as output tracking and output regulation.

The goal of this paper is to develop a framework and sound theory to study this

*Received by the editors July 22, 1996; accepted for publication (in revised form) August 5, 1997; published electronically September 25, 1998.
http://www.siam.org/journals/sicon/37-1/30706.html

†Sección de Control Automático, Departamento de Ingeniería Eléctrica, CINVESTAV–IPN, Apartado Postal 14-740, 07000 México, D.F., México (aranda@ctrl.cinvestav.mx). This work was done while this author was with the Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada.

‡Department of Mathematics and Statistics, Queen's University, Kingston, ON K7L 3N6, Canada (ron@rmh.mast.queensu.ca).

new group of transformations and then identify those systems which are equivalent to a prime system under this group of transformations.

Of course, a necessary condition for this equivalence is that the system be linearizable by static state feedback. It will be shown that the crucial step is the requirement that certain linear forms on the extended output space can be constructed such that their *pull-back* under the output map coincide with some suitable forms on the state space. In that respect, it is worth mentioning that the pull-back of a form is *always* a well-defined object, as opposed to the *push-forward* of a distribution, which may fail to be well defined. Therefore, one completely avoids the projectability-type conditions as stated in [10].

We present our results within the linear algebraic framework introduced by Di Benedetto, Grizzle, and Moog [6]. However, it will be shown that our results can be given a meaningful geometric interpretation in terms of jet bundles [15]. Finally, let us mention that this work was partially motivated by some results previously obtained for discrete-time systems [1]. One advantage of our formalism is that it allows a completely parallel treatment of both the continuous- and the discrete-time cases.

The paper is organized as follows. In section 2 we recall some basic definitions from the so-called linear algebraic approach [2, 6]. Our main results are contained in section 3. In subsection 3.1 we obtain new necessary and sufficient conditions for equivalence to prime system under regular static state feedback, state space diffeomorphism, and output space diffeomorphism. In subsection 3.2 we introduce the notion of GOT and study some of its properties. In subsection 3.3 we derive necessary and sufficient conditions for equivalence to prime system under regular static state feedback, state space diffeomorphism, and GOT. Finally, some conclusions and final remarks are offered in section 4.

2. Linear algebraic framework. To begin with, we recall some basic definitions from [2, 6]. Consider a nonlinear system Σ, described by equations of the form

$$\Sigma : \begin{cases} \dot{x} &= f(x) + \sum_{i=1}^{m} u_i g_i(x) = f(x) + g(x)u, \\ y &= h(x), \end{cases} \tag{1}$$

where the state $x \in M$, an open and connected subset of $\mathbb{R}^n$, the control $u \in \mathbb{R}^m$, and the output $y \in Y$, an open and connected subset of $\mathbb{R}^m$. Throughout the paper the following standing assumptions are made:

A1. The vector fields $f(x)$ and $g_i(x)$ and the mapping $h(x)$ are real analytic.

A2. For almost all $x \in M$, $\operatorname{rank} g(x) = \operatorname{rank} \mathrm{d}h(x) = m$.

Let $\mathcal{K}$ denote the field of meromorphic functions of a finite number of the variables $\{x, u^{(j)},\ j \geq 0\}$. The time derivative of a function $\varphi \in \mathcal{K}$ is defined by

$$\dot{\varphi} = \frac{d}{dt}\varphi = \frac{\partial \varphi}{\partial x}[f(x) + g(x)u] + \sum_{j \geq 0} \frac{\partial \varphi}{\partial u^{(j)}} u^{(j+1)}. \tag{2}$$

Notice that the sum in (2) involves only finitely many terms. Let $\mathcal{E}$ denote the $\mathcal{K}$-vector space spanned by $\{\mathrm{d}x, \mathrm{d}u^{(j)},\ j \geq 0\}$, where $\mathrm{d}x$ and $\mathrm{d}u^{(j)}$ stand, respectively, for $\{\mathrm{d}x_1, \ldots, \mathrm{d}x_n\}$ and $\{\mathrm{d}u_1^{(j)}, \ldots, \mathrm{d}u_m^{(j)}\}$. The elements of $\mathcal{E}$ are *differential forms* of degree one, or simply *one-forms*. The operator $\frac{d}{dt} : \mathcal{K} \to \mathcal{K}$ induces a derivation in $\mathcal{E}$ by

$$\omega = \sum_j a_j \mathrm{d}v_j \mapsto \dot{\omega} = \sum_j (\dot{a}_j \mathrm{d}v_j + a_j \mathrm{d}\dot{v}_j).$$

The *relative degree* r of a one-form $\omega \in \mathcal{E}$ is defined to be the least integer such that $\omega^{(r)} \not\in \operatorname{span}_{\mathcal{K}}\{\mathrm{d}x\}$. If such an integer does not exist, set $r = \infty$.

Introduce a sequence of subspaces $\{\mathcal{H}_k\}$ of $\mathcal{E}$ by

$$\begin{array}{rcl} \mathcal{H}_1 & = & \operatorname{span}_{\mathcal{K}}\{\mathrm{d}x\}, \\ \mathcal{H}_{k+1} & = & \operatorname{span}_{\mathcal{K}}\{\omega \in \mathcal{H}_k \mid \dot{\omega} \in \mathcal{H}_k\},\ k \geq 1. \end{array} \tag{3}$$

This sequence of subspaces was first introduced in [2, 3] to address the dynamic feedback linearization problem. It is clear that the sequence (3) is decreasing. Denote by k^* the least integer such that

$$\mathcal{H}_1 \supset \mathcal{H}_2 \supset \cdots \supset \mathcal{H}_{k^*} \supset \mathcal{H}_{k^*+1} = \mathcal{H}_{k^*+2} = \cdots =: \mathcal{H}_\infty. \tag{4}$$

Assume that $\mathcal{H}_\infty = 0$. We shall explain below the significance of this assumption. In [2] it was proven that there exists a set of one-forms $\mathcal{W} = \{\omega_1, \ldots, w_m\}$ and a list of integers $\{r_1, \ldots, r_m\}$ such that, for $1 \leq k \leq k^*$,

$$\mathcal{H}_k = \operatorname{span}_{\mathcal{K}}\{\omega_i^{(j)}, \mid r_i \geq k,\ 0 \leq j \leq r_i - k\}. \tag{5}$$

The integer r_i associated to the one-form w_i coincides with its relative degree. A set of one-forms satisfying (5) is called a system of *linearizing one-forms.*

According to (2), define

$$\begin{aligned} h^{(1)} &= h^{(1)}(x, u) \\ &= \frac{\partial h}{\partial x}[f(x) + g(x)u], \\ h^{(k+1)} &= h^{(k+1)}(x, u, \ldots, u^{(k)}) \\ &= \frac{\partial h^{(k)}}{\partial x}[f(x) + g(x)u] + \sum_{i=0}^{k-1} \frac{\partial h^{(k)}}{\partial u^{(i)}} u^{(i+1)}. \end{aligned}$$

Therefore, associated to the system Σ, we can define two sequences of subspaces $\{\mathcal{E}_k\}$ and $\{\mathcal{F}_k\}$ of $\mathcal{E}$ defined by

$$\begin{array}{rcl} \mathcal{E}_k & = & \operatorname{span}_{\mathcal{K}}\{\mathrm{d}x, \mathrm{d}h, \ldots, \mathrm{d}h^{(k)}\},\ k \geq 0, \\ \mathcal{F}_k & = & \operatorname{span}_{\mathcal{K}}\{\mathrm{d}h, \ldots, \mathrm{d}h^{(k)}\},\ k \geq 0. \end{array}$$

These two sequences of subspaces were first introduced in [6] to unify different notions of invertibility for nonlinear systems. The number $\rho^* = \dim \mathcal{E}_n - \dim \mathcal{E}_{n-1}$ is called the *rank of the system* Σ. It can be shown [6] that $\rho^* = \dim \mathcal{E}_n - \dim \mathcal{E}_{n-1} = \dim \mathcal{F}_n - \dim \mathcal{F}_{n-1}$. This characterization of rank was introduced in [6] and agrees with Fliess's definition [7]. Finally, for notational convenience, define $\mathcal{X} = \operatorname{span}_{\mathcal{K}}\{\mathrm{d}x\}$.

Remark 2.1. In paper [6] the notation $y^{(k+1)}(x, u, \ldots, u^{(k)})$ was used instead of $h^{(k+1)}(x, u, \ldots, u^{(k)})$. We use the latter notation because, in the next section, the $y_j^{(k)}$ will be used to denote the canonical system of coordinates of the extended output space. □

Remark 2.2. Throughout the paper we use the notion of pull-back of a differential form, as well as the differential forms version of Frobenius theorem. For details, the reader is referred to [4]. □

3. Main results.

3.1. Equivalence to prime systems. In this section we present new necessary and sufficient conditions for equivalence to prime system under state diffeomorphism, regular static state feedback, and output space diffeomorphism. In a sense, this result is a particular case of the more general notion of equivalence that we introduce below and provides a new linear algebraic characterization of the class of systems already identified in [10].

Definition 3.1. *A system* Π *is said to be a* prime system *if it is of the form*

$$\Pi \ : \left\{ \begin{array}{rcl} \dot{z}_{i1} & = & z_{i2}, \\ & \vdots & \\ \dot{z}_{i\kappa_i} & = & v_i, \\ \tilde{y}_i & = & z_{i1}, \ \ 1 \le i \le m, \end{array} \right. \tag{6}$$

where $z = (z_{11}, \dots, z_{1\kappa_1}, \dots, z_{m1}, \dots, z_{m\kappa_m}) \in \mathbb{R}^n$ *and* $n = \sum_{i=1}^{m} \kappa_i$ *for some integers* $\{\kappa_i\}_{i=1}^{m}$.

Definition 3.2. *The system* Σ *is said to be* equivalent to the prime system Π *if there exist*

(i) *A state diffeomorphism*

$$\begin{array}{rcl} \phi : M & \to & \mathbb{R}^n, \\ x & \mapsto & z = \phi(x); \end{array}$$

(ii) *a regular static state feedback* $u = \alpha(x) + \beta(x)v$*; that is,* $\beta(x)$ *is a square nonsingular matrix;*

(iii) *an output space diffeomorphism*

$$\begin{array}{rcl} \psi : Y & \to & \mathbb{R}^m, \\ y & \mapsto & \tilde{y} = \psi(y) \end{array}$$

such that the transformation of Σ *under* $(\phi, (\alpha, \beta), \psi)$ *equals* Π.

In order to state our first result, we need to introduce some notation. First define $\dot{\mathcal{H}}_k = \operatorname{span}_{\mathcal{K}}\{\dot{\omega} \mid \omega \in \mathcal{H}_k\}$. If $\{\omega_1, \dots, \omega_s\}$ is a basis of $\mathcal{H}_k$, it is easy to check that $\dot{\mathcal{H}}_k = \operatorname{span}_{\mathcal{K}}\{\omega_1, \dots, \omega_s, \dot{\omega}_1, \dots, \dot{\omega}_s\}$. However, in general, $\{\omega_1, \dots, \omega_s, \dot{\omega}_1, \dots, \dot{\omega}_s\}$ is not a basis of $\dot{\mathcal{H}}_k$, because it may happen that the latter set is not linearly independent. Also, recall that $\mathcal{F}_0 = \operatorname{span}_{\mathcal{K}}\{\mathrm{d}h(x)\}$.

Theorem 3.3. *Consider the square nonlinear system* Σ *and suppose that it satisfies* A1 *and* A2. *Then* Σ *is equivalent to prime system* Π *if and only if the following conditions are satisfied:*

(i) $\mathcal{H}_\infty = 0$;

(ii) *for* $k = 1, \dots, k^*$, $\mathcal{H}_k$ *is completely integrable;*

(iii) *for* $k = 1, \dots, k^*$, $\mathcal{H}_k = \dot{\mathcal{H}}_{k+1} \oplus \operatorname{span}_{\mathcal{K}}\{\mathcal{W}_k\}$, *where* $\mathcal{W}_k \subset \mathcal{F}_0$;

(iv) *for* $k = 1, \dots, k^*$, $\mathcal{H}_k \cap \mathcal{F}_0$ *is completely integrable.*

Remark 3.4. Theorem 3.3 can be seen as a dual version of Theorem 4 in [10]. In particular, conditions (i), (ii), and (iv) of Theorem 3.3 are equivalent, respectively, to conditions (ii), (i), and (iv) of Theorem 4 in [10]. Also, notice that our conditions require the construction of a single sequence of subspaces or codistributions. □

Remark 3.5. In [2, 3] it has been shown that conditions (i) and (ii) are the necessary and sufficient conditions for the system Σ without outputs to be linearizable by state diffeomorphism and regular static state feedback. In particular, condition (i) is a necessary and sufficient condition for strong accessibility. □

Remark 3.6. In the event that the conditions of Theorem 3.3 are satisfied, the *decoupling matrix* [9, p. 263], [13, p. 254] of the *transformed* output $\tilde{y} = \psi(y)$ has full rank m. Therefore, the problem of asymptotically tracking a desired output y_d is transformed into the problem of asymptotically tracking the desired output $\tilde{y}_d = \psi(y_d)$, which is a *linear* problem in the transformed coordinates. □

Proof of Theorem 3.3. *Necessity.* First notice that the subspaces $\mathcal{H}_k$ are invariant under state diffeomorphism and under regular static state feedback [2, 3]. Moreover, they are independent of the output map. Next we show that the subspace $\mathcal{F}_0 \subset \mathcal{X}$ is invariant under output space diffeomorphism. Suppose that $\tilde{y} = \tilde{h}(x) = \psi \circ h(x)$. By the chain rule, we have

$$\mathrm{d}\tilde{h}(x) = \left.\frac{\partial \psi}{\partial y}\right|_{y=h(x)} \mathrm{d}h(x).$$

This shows that $\operatorname{span}_{\mathcal{K}}\{\mathrm{d}\tilde{h}(x)\} \subset \operatorname{span}_{\mathcal{K}}\{\mathrm{d}h(x)\}$. Since $y \mapsto \psi(y)$ is a diffeomorphism, a similar argument shows that $\operatorname{span}_{\mathcal{K}}\{\mathrm{d}h(x)\} \subset \operatorname{span}_{\mathcal{K}}\{\mathrm{d}\tilde{h}(x)\}$, and hence $\operatorname{span}_{\mathcal{K}}\{\mathrm{d}\tilde{h}(x)\} = \operatorname{span}_{\mathcal{K}}\{\mathrm{d}h(x)\}$. In a similar manner, one can show that $\mathcal{F}_0$ is invariant under state space diffeomorphism. Finally, it is clear that $\mathcal{F}_0$ remains unchanged under state feedback. Thus conditions (i)–(iv) are invariant under all the considered transformations.

An easy computation shows that conditions (i)–(iv) are satisfied for a prime system Π. Therefore, if system Σ is equivalent to a prime system Π, conditions (i)–(iv) necessarily hold true.

Sufficiency. We proceed by induction, going down from $k = k^*$ to $k = 1$. First notice that $\mathcal{H}_{k^*+1} = \mathcal{H}_\infty = 0$, and hence $\dot{\mathcal{H}}_{k^*+1} = 0$. Therefore, when $k = k^*$, condition (iii) means that $\mathcal{H}_{k^*} = \operatorname{span}_{\mathcal{K}}\{\mathcal{W}_{k^*}\}$, where

$$\mathcal{W}_{k^*} = \{\omega_{k^*1}, \ldots, \omega_{k^*\rho_{k^*}}\} \subset \mathcal{F}_0. \tag{7}$$

Moreover, by condition (iv) we can assume without loss of generality that the forms ω_{k^*i} are exact, say $\omega_{k^*i} = \mathrm{d}\varphi_{k^*i}$. By (7), the forms ω_{k^*i} can also be written as follows:

$$\omega_{k^*i} = \mathrm{d}\varphi_{k^*i} = \sum_{j=1}^{m} a_{ij}\mathrm{d}h_j(x), \tag{8}$$

where $a_{ij} \in \mathcal{K}$.

Even though the forms ω_{k^*i} are linear combinations of the differentials $\mathrm{d}h_j(x)$, it is not possible to assert a priori that they are the pull-back of some forms on the output space. This assertion holds true if and only if the coefficients a_{ij} can be expressed as functions of the scalar outputs y_j. The following lemma states that this is the case indeed.

LEMMA 3.7. *The coefficients a_{ij}, for $1 \leq i \leq \rho_{k^*}$, $1 \leq j \leq m$, in* (8) *can be expressed as functions of $y_1, \ldots, y_m$ only; that is, $a_{ij} = a_{ij}(y) = a_{ij} \circ h(x)$.*

Proof. Since the forms ω_{k^*i} are exact, we have that

$$\mathrm{d}\omega_{k^*i} = \sum_{j=1}^{m} \mathrm{d}a_{ij} \wedge \mathrm{d}h_j(x) = 0, \; i = 1, \ldots, \rho_{k^*}. \tag{9}$$

Now, taking the exterior product of (9) with the $(m-1)$-form $\mathrm{d}h_1(x)\wedge\cdots\widehat{\mathrm{d}h_j(x)}\cdots\wedge \mathrm{d}h_m(x)$, where $\widehat{\mathrm{d}h_j(x)}$ means that that factor is omitted, yields

$$\text{(10)}\quad \mathrm{d}a_{ij}\wedge \mathrm{d}h_1(x)\wedge\cdots\wedge \mathrm{d}h_j(x)\wedge\cdots\wedge \mathrm{d}h_m(x)=0,\ 1\le i\le \rho_{k^*},\ 1\le j\le m.$$

Since $\operatorname{rank}\mathrm{d}h(x)=m$, it follows that the linear forms $\mathrm{d}h_j(x)$ are independent. Consequently, (10) implies that $\mathrm{d}a_{ij}\in \operatorname{span}_{\mathcal{K}}\{\mathrm{d}h(x)\}$. The latter means that a_{ij} are constant on each submanifold $h^{-1}(K)$, $K\in\mathbb{R}^m$. Again, since $\operatorname{rank}\mathrm{d}h(x)=m$, there are coordinates $(y_1,\ldots,y_m,q_1,\ldots,q_{n-m})$ of M such that $h:M\to Y$ becomes the canonical submersion $(y,q)\mapsto(y)$. Therefore,

$$a_{ij}(x)=a_{ij}(y,q)=a_{ij}(y,0)=a_{ij}(y).\qquad \square$$

By Lemma 3.7, we can define ρ_{k^*} forms on the output space Y by $\eta_{k^*i}=\sum_{j=1}^m a_{ij}(y)\mathrm{d}y_j$, $i=1,\ldots,\rho_{k^*}$. Then it is clear that the pull-back (see [4]) of the form η_{k^*i} under the map $h:M\to Y$ coincides precisely with the form ω_{k^*i}; that is,

$$\text{(11)}\qquad \omega_{k^*i}=h^*(\eta_{k^*i}),\ i=1,\ldots,\rho_{k^*}.$$

Now suppose that through steps $\ell=k^*$ to $\ell=k+1$ we have constructed sets of forms $\mathcal{W}_\ell=\{\omega_{\ell 1},\ldots,\omega_{\ell\rho_\ell}\}$ (some of them possibly empty) such that $\omega_{ij}=h^*(\eta_{ij})$ and

$$\mathcal{H}_{k+1}=\operatorname{span}_{\mathcal{K}}\{\mathcal{W}_i^{(j)},\ k+1\le i\le k^*,\ 0\le j\le i-k-1\},$$

where the notation $\mathcal{W}_i^{(j)}$ should be understood elementwise. Therefore, by condition (iii), we can choose a set of forms $\mathcal{W}_k=\{\omega_{k1},\ldots,\omega_{k\rho_k}\}\subset\mathcal{F}_0$ such that

$$\mathcal{H}_k=\operatorname{span}_{\mathcal{K}}\{\mathcal{W}_i^{(j)},\ k\le i\le k^*,\ 0\le j\le i-k\}.$$

Moreover, condition (iv) means that we can assume, without loss of generality, that ω_{ki} are exact, say $\omega_{ki}=\mathrm{d}\varphi_{ki}$, and that there are forms η_{ki} defined on the output space such that, for $i=1,\ldots,\rho_k$, $\omega_{ki}=h^*(\eta_{ki})$.

Repeat the above construction from $k=k^*$ to $k=1$, and let $\mathcal{W}=\cup_{i=1}^{k^*}\mathcal{W}_i=\{\omega_1,\ldots,\omega_s\}$ and $r_i=\{k\mid \omega_i\in\mathcal{W}_k\}$. Notice that, by construction, each ω_i belongs to one and only one set $\mathcal{W}_k$, whence the integer r_i is well-defined. Therefore, $\mathcal{W}=\{\omega_1,\ldots,\omega_s\}$ is a system of linearizing forms whose list of relative degrees is $\{r_1,\ldots,r_s\}$. As a matter of fact, it can be shown (see, e.g., [2, 3]) that $s=m$ and that $\sum_{i=1}^m r_i=n$. Recall that, by construction, the forms ω_i are exact, say $\omega_i=\mathrm{d}\varphi_i(x)$. Define

$$z_{ij}=\phi_{ij}(x)=\varphi_i^{(j-1)}(x),\ 1\le i\le m,\ 1\le j\le r_i.$$

It follows that the map $x\mapsto\phi(x)$ is a diffeomorphism. In coordinates z_{ij}, system Σ becomes

$$\begin{aligned}\dot z_{i1}&=z_{i2},\\ &\ \ \vdots\\ \dot z_{ir_i}&=a_i(z)+b_i(z)u,\ 1\le i\le m,\\ y&=h\circ\phi^{-1}(z).\end{aligned}$$

The fact that the forms $\omega_i^{(j)}$ are independent implies that the matrix $B(z)$, whose rows are $b_i(z)$, has full rank. Therefore, the static state feedback $u = [B(z)]^{-1}[v - a(z)]$ is well defined and yields

$$\begin{array}{rcl} \dot{z}_{i1} & = & z_{i2}, \\ & \vdots & \\ \dot{z}_{ir_i} & = & v_i, \ 1 \le i \le m, \\ y & = & h \circ \phi^{-1}(z). \end{array} \tag{12}$$

To conclude the proof, we just need to construct a suitable output space diffeomorphism. In order to do so, we need the following result.

LEMMA 3.8. *Let $\{\eta_1, \ldots, \eta_m\}$ be the collection of forms defined on the output space which satisfy $\omega_i = h^*(\eta_i)$. Then, for $i = 1, \ldots, m$, $\mathrm{d}\eta_i = 0$.*

Proof. As in the proof of Lemma 3.7, it is possible to choose a coordinates system $(y_1, \ldots, y_m, q_1, \ldots, q_{n-m})$, so that $h : M \to Y$ becomes the canonical projection $(y, q) \mapsto y$. Let $w_i = \sum_{j=1}^m a_{ij}\mathrm{d}h_j$. We have already shown that the a_{ij} can be expressed as functions of y_j only. Since $\omega_i = h^*(\eta_i)$, we have that necessarily $\eta_i = \sum_{j=1}^m a_{ij}\mathrm{d}y_j$. Define $\tau^i_{jk} = \frac{\partial a_{ik}}{\partial y_j} - \frac{\partial a_{ij}}{\partial y_k}$. Then, the two-form $\mathrm{d}\eta_i$ can be written as $\mathrm{d}\eta_i = \sum_{j<k} \tau^i_{jk}\mathrm{d}y_j \wedge \mathrm{d}y_k$. Now, recall that $\mathrm{d}\omega_i = \sum_{j=1}^m \mathrm{d}a_{ij} \wedge \mathrm{d}h_j$ so that, in coordinates (y, q), the two-form $\mathrm{d}\omega_i$ becomes $\mathrm{d}\omega_i = \sum_{j<k} \tau^i_{jk}\mathrm{d}y_j \wedge \mathrm{d}y_k$. By construction, the forms ω_i are exact, so that $\mathrm{d}\omega_i \equiv 0$. Therefore, the coefficients τ^i_{jk} must be identically zero. □

By virtue of Lemma 3.8, we can assume, without loss of generality, that, for $i = 1, \ldots, m$, $\eta_i = \mathrm{d}\psi_i(y)$. Finally, define the output space diffeomorphism $y \mapsto \psi(y)$. In coordinates $\tilde{y} = \psi(y)$, the system (12) is in prime form. □

Example 3.9 (see [10]). Consider the system

$$\begin{array}{rcl} \dot{x}_1 & = & u_1, \\ \dot{x}_2 & = & x_3, \\ \dot{x}_3 & = & u_2, \\ \dot{x}_4 & = & x_5, \\ \dot{x}_5 & = & x_6, \\ \dot{x}_6 & = & u_3, \end{array} \qquad \begin{array}{rcl} y_1 & = & x_1, \\ y_2 & = & x_2 + x_1x_5, \\ y_3 & = & x_4, \end{array} \tag{13}$$

defined on $M = \mathbb{R}^6$, $Y = \mathbb{R}^3$. For system (13) we easily compute

$$\begin{array}{rcl} \mathcal{H}_2 & = & \mathrm{span}_{\mathcal{K}}\{\mathrm{d}x_2, \mathrm{d}x_4, \mathrm{d}x_5\}, \\ \mathcal{H}_3 & = & \mathrm{span}_{\mathcal{K}}\{\mathrm{d}x_4\}, \\ \mathcal{H}_\infty & = & 0. \end{array}$$

Therefore, conditions (i) and (ii) of Theorem 3.3 are satisfied. Moreover, if we choose $\mathcal{W}_1 = \{\mathrm{d}x_1\}$, $\mathcal{W}_2 = \{\mathrm{d}x_2 + x_1\mathrm{d}x_5\}$, and $\mathcal{W}_3 = \{\mathrm{d}x_4\}$ we see also that condition (iii) is satisfied. However, $\mathrm{span}_{\mathcal{K}}\{\mathcal{W}_2\} = \mathcal{H}_2 \cap \mathcal{F}_0$ is not completely integrable. Therefore, system (13) is not equivalent to prime system. □

3.2. Generalized output transformations. Next we introduce the notion of GOT. As we point out below, the notion of equivalence studied in the previous section is a particular case of this new class of transformations.

DEFINITION 3.10. *Given two finite nonnegative integers d and d', a GOT consists of two smooth maps $\psi : Y \times (\mathbb{R}^m)^d \to \mathbb{R}^m$ and $\chi : \mathbb{R}^m \times (\mathbb{R}^m)^{d'} \to Y$ such that*

$$\tilde{y} = \psi(y, \dot{y}, \ldots, y^{(d)}), \tag{14}$$

$$y = \chi(\tilde{y}, \dot{\tilde{y}}, \ldots, \tilde{y}^{(d')}). \tag{15}$$

Even though $y = h(x)$, in general we have $\tilde{y} = \tilde{h}(x, u, \dots, u^{(d-1)})$. A GOT is called proper *if $\tilde{h}$ is a function of x only; i.e., $\tilde{y} = \tilde{h}(x)$.*

In the case when $d = d' = 0$, the GOT (14)–(15) reduces to an output space diffeomorphism. Notice, however, that in general $d \neq d'$.

Example 3.11. Let $Y = \mathbb{R}^3$, and let $y = (y_1, y_2, y_3)$ and $z = (z_1, z_2, z_3)$ be two systems of coordinates of $\mathbb{R}^3$. Consider the maps

$$\begin{array}{rcl} \psi : Y \times \mathbb{R}^3 & \rightarrow & \mathbb{R}^3, \\ (y, \dot{y}) & \mapsto & (y_1, y_2 + y_1\dot{y}_1, y_3 + \dot{y}_2), \\ \chi : \mathbb{R}^3 \times (\mathbb{R}^3)^2 & \rightarrow & Y, \\ (z, \dot{z}, \ddot{z}) & \mapsto & (z_1, z_2 - z_1\dot{z}_1, z_3 - \dot{z}_2 + z_1\ddot{z}_1 + \dot{z}_1^2). \end{array}$$

It is easy to verify that the pair (ψ, χ) is a GOT with $d = 1$ and $d' = 2$. □

Remark 3.12. Parallel to the algebraic definition of GOT given by Definition 3.10, a more geometric interpretation of this class of transformations can be given in terms of jet bundles (see, e.g., [15]). Consider the output $y \in Y$ as a function of time and assume that it is of class C^∞. Then, every trajectory $y(t)$ in the output space defines a section of the d-jet bundle $J^d(\mathbb{R}, Y)$ by $t \mapsto j_t^d = (t, y(t), \dot{y}, \dots, y^{(d)}(t))$. Similarly, every smooth curve $z(t)$ on $\mathbb{R}^m$ defines a section of the d'-jet bundle $J^{d'}(\mathbb{R}, \mathbb{R}^m)$. Therefore, the maps ψ and χ can be interpreted as bundle maps such that the following diagrams commute:

$$\begin{tikzcd} J^d(\mathbb{R}, Y) \arrow[r, "\psi"] \arrow[d, "\pi"'] & J^0(\mathbb{R}, \mathbb{R}^m) \arrow[d, "\pi"] \\ \mathbb{R} \arrow[r, "\mathrm{id}_{\mathbb{R}}"'] & \mathbb{R} \end{tikzcd} \qquad \begin{tikzcd} J^{d'}(\mathbb{R}, \mathbb{R}^m) \arrow[r, "\chi"] \arrow[d, "\pi"'] & J^0(\mathbb{R}, Y) \arrow[d, "\pi"] \\ \mathbb{R} \arrow[r, "\mathrm{id}_{\mathbb{R}}"'] & \mathbb{R} \end{tikzcd}$$

where $\pi : J^k(\mathbb{R}, N) \to \mathbb{R}$ is the source map and $\mathrm{id}_{\mathbb{R}}$ is the identity map. Roughly speaking, this means that to every smooth trajectory $y(t) \in Y$ corresponds one and only one trajectory $z(t) \in \mathbb{R}^m$. □

Several types of invariants have been associated with the input-output map of the system Σ, e.g., the *relative degrees* [9, 13], the *structure at infinity* [11], the *essential orders* [8], and the *rank of the system* [6, 7]. Among them, the most fundamental is, without doubt, the rank ρ^* (see section 2). Theorem 3.13 below states that the rank is invariant under proper GOTs.

THEOREM 3.13. *Let (ψ, χ) be a proper GOT, and let ρ^* and $\tilde{\rho}^*$ denote, respectively, the rank of the system Σ with respect to the output $y = h(x)$ and with respect to the new output $\tilde{y} = \psi(y, \dot{y}, \dots, y^{(d)}) = \tilde{h}(x)$. Then $\tilde{\rho}^* = \rho^*$.*

Proof. Let $\{\mathcal{F}_k\}$ and $\{\tilde{\mathcal{F}}_k\}$ denote, respectively, the sequences of subspaces associated to the system Σ with the output $y = h(x)$ and with the new output $\tilde{y} = \tilde{h}(x)$. Next note that, according to Definition 3.10, we have

$$\dot{\tilde{h}}_i = \sum_{k=1}^{m} \sum_{\ell=0}^{d} \left. \frac{\partial \psi_i}{\partial y_k^{(\ell)}} \right|_{h_k^{(\ell)}} h_k^{(\ell+1)},$$

$$\tilde{h}_i^{(j+1)} = \sum_{k=1}^{m} \sum_{\ell=0}^{d+j} \frac{\partial \tilde{h}_i^{(j)}}{\partial h_k^{(\ell)}} h_k^{(\ell+1)}, \qquad j \geq 1. \tag{16}$$

Pick an arbitrary form $\omega \in \tilde{\mathcal{F}}_k$, say $\omega = \sum_{i=1}^{m} \sum_{j=0}^{k} \tilde{a}_{ij} \mathrm{d}\tilde{h}_i^{(j)}$. By (16), it follows that there are coefficients a_{ij} such that $\omega = \sum_{i=1}^{m} \sum_{j=0}^{k+d} a_{ij} \mathrm{d}h_i^{(j)}$. This shows that, for every integer $k \geq 0$, we have $\tilde{\mathcal{F}}_k \subset \mathcal{F}_{k+d}$. Similarly, it is easily seen that, for $k \geq 0$, $\mathcal{F}_k \subset \tilde{\mathcal{F}}_{k+d'}$. Let $s = \max\{d, d'\}$. Then the subspaces $\{\mathcal{F}_k\}$ and $\{\tilde{\mathcal{F}}_k\}$ satisfy

$$\tilde{\mathcal{F}}_k \subset \mathcal{F}_{k+s} \subset \tilde{\mathcal{F}}_{k+2s} \subset \mathcal{F}_{k+3s} \subset \cdots . \tag{17}$$

Now suppose that $\tilde{\rho}^* < \rho^*$, say $\rho^* = \tilde{\rho}^* + r$, for some integer $r > 0$. We will show that this leads to contradiction. First note that, for every $k \geq n$, we have

$$\begin{array}{rcl} \dim \mathcal{F}_k - \dim \mathcal{F}_{k-1} & = & \rho^* = \tilde{\rho}^* + r, \\ \dim \tilde{\mathcal{F}}_k - \dim \tilde{\mathcal{F}}_{k-1} & = & \tilde{\rho}^*. \end{array} \tag{18}$$

Next let $\Theta_i = \{\theta_i^j,\ j = 1, \ldots, \ell_i\}$ denote some sets of linear forms such that Θ_0 is a basis of $\tilde{\mathcal{F}}_n$, $\Theta_0 \cup \Theta_1$ is a basis of $\mathcal{F}_{n+s}$, $\Theta_0 \cup \Theta_1 \cup \Theta_2$ is a basis of $\tilde{\mathcal{F}}_{n+2s}$, etc. Note that, by construction, $\ell_i \geq 0$ for every integer $i \geq 1$.

By definition of the sets Θ_i and by (18), we have that

$$\begin{array}{rclcl} \dim \mathcal{F}_{n+(2k+1)s} & = & \sum_{i=0}^{2k+1} \ell_i & = & \ell_0 + \ell_1 + 2ks\rho^* = \ell_0 + \ell_1 + 2ks\tilde{\rho}^* + 2ksr, \\ \dim \tilde{\mathcal{F}}_{n+(2k+2)s} & = & \sum_{i=0}^{2k+2} \ell_i & = & \ell_0 + (2k+2)s\tilde{\rho}^*. \end{array}$$

Solving for ℓ_{2k+2} we obtain

$$\ell_{2k+2} = 2s\tilde{\rho}^* - \ell_1 - 2ksr. \tag{19}$$

By (19), ℓ_{2k+2} becomes negative for k large enough, which is a contradiction. This concludes the proof. □

3.3. Equivalence under generalized transformations.

Definition 3.14. *The system Σ is said to be* equivalent to the prime system Π under proper GOT *if there exist*

(i) *a state diffeomorphism*

$$\begin{array}{rcl} \phi : M & \rightarrow & \mathbb{R}^n, \\ x & \mapsto & z = \phi(x); \end{array}$$

(ii) *a regular static state feedback $u = \alpha(x) + \beta(x)v$;*

(iii) *a proper GOT (ψ, χ)*

such that the transformation of Σ under $(\phi, (\alpha, \beta), (\psi, \chi))$ equals Π.

Remark 3.15. We restrict ourselves to proper GOTs because we are studying equivalence to prime systems, for which the output is a function of the state only. □

Theorem 3.16. *Consider the square nonlinear system Σ and suppose that it satisfies* A1 *and* A2. *Then Σ is equivalent to prime system Π under proper GOT if and only if the following conditions are satisfied:*

(i) $\mathcal{H}_\infty = 0$;

(ii) *for $k = 1, \ldots, k^*$, $\mathcal{H}_k$ is completely integrable;*

(iii) $\mathcal{X} \cap \mathcal{F}_{n-1} = \mathcal{X}$*;*

(iv) $\rho^* = m$.

Remark 3.17. A system which satisfies the hypothesis of Theorem 3.3 (and hence is equivalent to prime system Π) also satisfies conditions (i)–(iv) above. Clearly, conditions (i)–(ii) hold. Next, note that condition (iii) of Theorem 3.3 implies $\mathcal{H}_{k^*} \subset \mathcal{F}_0$, $\mathcal{H}_{k^*-1} \subset \mathcal{F}_1$, etc., and thus $\mathcal{X} = \mathcal{H}_1 \subset \mathcal{F}_{k^*-1} \subset \mathcal{F}_{n-1}$. Therefore condition (iii) holds. Finally, from Remark 3.6 we can see that condition (iv) holds as well. □

In order to prove Theorem 3.16, we need to introduce some notation. For a given nonnegative integer d, the extended state space, extended output space, and extended output map are defined, respectively, by $M^d = M \times (\mathbb{R}^m)^{d+1}$, $Y^d = Y \times (\mathbb{R}^m)^d$, and

$$\begin{array}{rcl} h_e : M \times (\mathbb{R}^m)^{(d+1)} & \to & Y \times (\mathbb{R}^m)^d, \\ (x, u, \dots, u^{(d)}) & \mapsto & (y, \dots, y^{(d)}) = (h(x), \dots, h^{(d)}(x, \dots, u^{(d-1)})). \end{array}$$

Also, we will need the following technical result.

PROPOSITION 3.18. *The rank ρ^* of the system Σ is equal to m if and only if, for every integer $N > 0$,* $\dim \mathcal{F}_N = mN$*.*

Proof. First suppose that $\rho^* = m$ but that for some integer N we have $\dim \mathcal{F}_N < mN$. Then, necessarily, there is a form $\mathrm{d}h_i^{(N)}$ and coefficients c_{ij}^k such that

$$\mathrm{d}h_i^{(N)} = \sum_{j=1}^{m} \sum_{k=0}^{(N-1)} c_{ij}^k \mathrm{d}h_j^{(k)} + \sum_{j \neq i} c_{ij}^N \mathrm{d}h_j^{(N)}.$$

The latter implies that for every integer $\bar{N} > N$ we have

$$\mathrm{d}h_i^{(\bar{N})} = \sum_{j=1}^{m} \sum_{k=0}^{(\bar{N}-1)} \bar{c}_{ij}^k \mathrm{d}h_j^{(k)} + \sum_{j \neq i} \bar{c}_{ij}^{\bar{N}} \mathrm{d}h_j^{(\bar{N})}. \tag{20}$$

Let us recall [6] that, for $N > n$, $\dim \mathcal{E}_N - \dim \mathcal{E}_{N-1} = \dim \mathcal{F}_N - \dim \mathcal{F}_{N-1}$. Thus, (20) implies that $\dim\, \mathcal{E}_{\bar{N}} - \dim\, \mathcal{E}_{\bar{N}-1} < m$. In particular, if we choose $\bar{N} > n$, we have that $\rho^* = \dim\, \mathcal{E}_{\bar{N}} - \dim\, \mathcal{E}_{\bar{N}-1} < m$, which is a contradiction. The converse is obvious. □

Proof of Theorem 3.16. *Necessity.* It is clear that conditions (i)–(iv) are satisfied for a system in prime form. Moreover, conditions (i)–(ii) are invariant under state diffeomorphism and regular static state feedback. On the other hand, Theorem 3.13 states that the rank ρ^* is invariant under proper GOTs. It remains to prove that condition (iii) is also invariant under proper GOTs. This part of the proof will be broken down into two lemmas.

LEMMA 3.19. *Let* $\mathcal{F}_\infty = \mathrm{span}_\mathcal{K}\{\mathrm{d}h_i^{(j)},\ 1 \le i \le m,\ j \ge 0\}$*. Then we have* $\mathcal{X} \cap \mathcal{F}_\infty = \mathcal{X} \cap \mathcal{F}_{n-1}$*.*

Proof. Pick an arbitrary linear form $\omega \in \mathcal{X} \cap \mathcal{F}_\infty$. Then there are coefficients b_{ij} and an integer $N \ge 0$ such that

$$\omega = \sum_{i=1}^{m} \sum_{j=0}^{N} b_{ij} \mathrm{d}h_i^{(j)}.$$

Next note that, by Proposition 3.18, the forms $\{\mathrm{d}h_i^{(j)},\ 1 \le i \le m,\ 0 \le j \le N\}$ are linearly independent. Then, necessarily, $\mathrm{d}h_i^{(j)} \notin \mathcal{X}$ for $j > n-1$. Since $\omega \in \mathcal{X}$, one concludes that $N < n-1$, whence $\omega \in \mathcal{X} \cap \mathcal{F}_{n-1}$. We have shown that $\mathcal{X} \cap \mathcal{F}_\infty \subset \mathcal{X} \cap \mathcal{F}_{n-1}$. On the other hand, it is obvious that $\mathcal{X} \cap \mathcal{F}_{n-1} \subset \mathcal{X} \cap \mathcal{F}_\infty$. □

LEMMA 3.20. *Let (ψ, χ) be a proper GOT, and let $\{\mathcal{F}_k\}$ and $\{\tilde{\mathcal{F}}_k\}$ denote, respectively, the sequences of subspaces associated with the system Σ with the output $y = h(x)$ and with the new output $\tilde{y} = \psi(y, \dot{y}, \dots, y^{(d)}) = \tilde{h}(x)$. Then $\tilde{\mathcal{F}}_{n-1} \cap \mathcal{X} = \mathcal{F}_{n-1} \cap \mathcal{X}$.*

Proof. In the proof of Theorem 3.13 we have shown that, for every integer $k \geq 0$, we have $\tilde{\mathcal{F}}_k \subset \mathcal{F}_{k+d}$. Pick an arbitrary form $\omega \in \tilde{\mathcal{F}}_{n-1} \cap \mathcal{X}$. Lemma 3.19 implies that

$$\omega \in \mathcal{F}_{n+d-1} \cap \mathcal{X} \subset \mathcal{F}_\infty \cap \mathcal{X} = \mathcal{F}_{n-1} \cap \mathcal{X}.$$

This proves that $\tilde{\mathcal{F}}_{n-1} \cap \mathcal{X} \subset \mathcal{F}_{n-1} \cap \mathcal{X}$. Since (ψ, χ) is a GOT, a symmetric argument shows that $\mathcal{F}_{n-1} \cap \mathcal{X} \subset \tilde{\mathcal{F}}_{n-1} \cap \mathcal{X}$. □

Sufficiency. Conditions (i)–(ii) imply that the system Σ is linearizable by regular static state feedback. Therefore, we can assume, without loss of generality, that Σ is in the form (12). Next notice that condition (iii) implies that, for $i = 1, \ldots, m$,

$$\mathrm{d}z_{i1} = \mathrm{d}\phi_{i1}(x) = \sum_{j=1}^{m} \sum_{k=0}^{d} a_{ij}^{k} \mathrm{d}h_j^{(k)}, \tag{21}$$

where $0 \leq d \leq n-1$.

LEMMA 3.21. *The coefficients a_{ij}^k appearing in* (21) *can be expressed as functions of* $\{y_i^{(j)},\ 1 \leq i \leq m,\ 0 \leq j \leq d\}$ *only.*

Proof. By Proposition 3.18, the forms $\{\mathrm{d}h_i^{(j)},\ 1 \leq i \leq m,\ 0 \leq j \leq d\}$ are independent. Therefore, a similar argument to that used in Lemma 3.7 can be used to conclude the proof. □

By Lemma 3.21, we can define m one-forms on the extended output space Y^d by

$$\eta_i = \sum_{j=1}^{m} \sum_{k=0}^{d} a_{ij}^{k} \mathrm{d}y_j^{(k)}.$$

Then it is clear that the pull-back of the forms η_i under the extended output map $h_e : M^d \to Y^d$ coincides precisely with the forms $\mathrm{d}z_{i1}$; that is, $\mathrm{d}z_{i1} = h_e^*(\eta_i)$.

LEMMA 3.22. *Let $\{\eta_1, \ldots, \eta_m\}$ be the collection of forms defined on the extended output space Y^d which satisfy $\mathrm{d}z_{i1} = h_e^*(\eta_i)$. Then, for $i = 1, \ldots, m$, $\mathrm{d}\eta_i = 0$.*

Proof. By Proposition 3.18, the extended output map $h_e : M^d \to Y^d$ has full rank equal to $m(d+1)$. Therefore, there exists a system of coordinates of the extended state space M^d such that $h_e : M^d \to Y^d$ becomes the canonical projection. A similar construction as in Lemma 3.8 shows then that $\mathrm{d}\eta_i = 0$. □

By virtue of Lemma 3.22, we can assume, without loss of generality, that, for $i = 1, \ldots, m$, $\eta_i = \mathrm{d}\psi_i(y, \dot{y}, \ldots, y^{(d)})$. Finally define a new output function $\tilde{y}$ by $\tilde{y}_i = \psi_i(y, \dot{y}, \ldots, y^{(d)})$. With this change of output variables, system (12) is in prime form. It remains to prove that there is an inverse map $y = \chi(\tilde{y}, \ldots, \tilde{y}^{(d')})$. Notice that $y = h(x) = h \circ \phi^{-1}(z)$ and that, by construction, $z_{ij} = \tilde{y}_i^{(j-1)}$. Therefore, it follows that necessarily

$$y = h \circ \phi^{-1}(z) = \chi(\tilde{y}, \ldots, \tilde{y}^{(d')})$$

for some integer $d' \geq 0$. □

Example 3.23 (Example 3.9, continued). We have shown that system (13) is not equivalent to a prime system under standard output space transformations, i.e., output space diffeomorphism. We shall show that system (13) is equivalent to prime system under proper GOTs. We have already shown that conditions (i)–(ii) are satisfied. It is easy to check that

$$\begin{aligned}
\mathcal{F}_0 &= \mathrm{span}_{\mathcal{K}}\{\mathrm{d}x_1, \mathrm{d}x_2 + x_1\mathrm{d}x_5, \mathrm{d}x_4\}, \\
\mathcal{F}_1 &= \mathrm{span}_{\mathcal{K}}\{\mathrm{d}x_1, \mathrm{d}x_2, \mathrm{d}x_3 + x_1\mathrm{d}x_6, \mathrm{d}x_4, \mathrm{d}x_5, \mathrm{d}u_1\}, \\
\mathcal{F}_2 &= \mathrm{span}_{\mathcal{K}}\{\mathrm{d}x_1, \mathrm{d}x_2, \mathrm{d}x_3, \mathrm{d}x_4, \mathrm{d}x_5, \mathrm{d}x_6, \mathrm{d}u_1, \mathrm{d}\dot{u}_1, \mathrm{d}u_2 + x_1\mathrm{d}u_3\}.
\end{aligned}$$

Note that $\mathcal{X} \subset \mathcal{F}_2$ and, since $\mathcal{F}_2 \subset \mathcal{F}_5$, we have $\mathcal{X} \cap \mathcal{F}_5 = \mathcal{X}$, and thus condition (iii) is also satisfied. Finally, lengthy but straightforward computations show that $\rho^* = \dim \mathcal{F}_6 - \dim \mathcal{F}_5 = 3$. Consequently, system (13) is equivalent to prime system.

Since the state equations are already in the form (1), in order to transform system (13) into prime form, we just need to find a suitable GOT. This can be accomplished as follows: first note that

$$\begin{array}{lclcl} \mathrm{d}x_1 & = & \mathrm{d}h_1 & = & h_e^*(\mathrm{d}y_1), \\ \mathrm{d}x_2 & = & \mathrm{d}h_2 - x_5\mathrm{d}h_1 - x_1\mathrm{d}\dot{h}_3 & = & h_e^*(\mathrm{d}y_2 - \dot{y}_3\mathrm{d}y_1 - y_1\mathrm{d}\dot{y}_3), \\ \mathrm{d}x_4 & = & \mathrm{d}h_3 & = & h_e^*(\mathrm{d}y_3). \end{array} \tag{22}$$

By integrating the right sides of (22), we find the GOT $(y, \dot{y}) \mapsto \tilde{y} = (y_1, y_2 - y_1\dot{y}_3, y_3)$. In coordinates $(\tilde{y}_1, \tilde{y}_2, \tilde{y}_3)$, system (13) is in prime form. The inverse output transformation is obviously given by $(\tilde{y}, \dot{\tilde{y}}) \mapsto y = (\tilde{y}_1, \tilde{y}_2 + \tilde{y}_1\dot{\tilde{y}}_3, \tilde{y}_3)$. □

As pointed out before, one immediate application of Theorems 3.3 and 3.16 is the possibility of (asymptotically or exactly) tracking a desired output trajectory $y_d(t)$. This is best illustrated by the following example.

Example 3.24 (Example 3.23, continued). The decoupling matrix [9, 13] associated with the original output $y = h(x)$ of system (13) is given by

$$B(x) = \begin{bmatrix} 1 & 0 & 0 \\ x_5 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Since $\operatorname{rank} B(x) = 2$, noninteracting control cannot be achieved by regular static state feedback.

On the other hand, the output functions which bring system (13) to prime form are given, as functions of x, by

$$\begin{array}{lcl} \tilde{y}_1 & = & y_1 = x_1, \\ \tilde{y}_2 & = & y_2 - y_1\dot{y}_3 = x_2, \\ \tilde{y}_3 & = & y_3 = x_4. \end{array}$$

The decoupling matrix $\tilde{B}(x)$ associated with the output $\tilde{y} = \tilde{h}(x)$ is simply $\tilde{B}(x) = I_3$. Therefore the *standard noninteracting feedback* [9, 13] can be used to decouple the scalar output components $\tilde{y}_i$, $i = 1, \dots, 3$. Now suppose that we want to asymptotically track a smooth output trajectory $y_d(t) = (y_{d1}(t), y_{d2}(t), y_{d3}(t))$ for system (13). Such trajectory is transformed in the new coordinates into $\tilde{y}_d(t) = (y_{d1}(t), y_{d2}(t) - y_{d1}(t)\dot{y}_{d3}(t), y_{d3}(t))$. Since (13) has been transformed into a prime system, the asymptotic output tracking problem is solved by *linear* state feedback in the transformed coordinates, namely

$$\begin{array}{lcl} u_1 & = & \dot{\tilde{y}}_{d1} - k_0^1(x_1 - \tilde{y}_{d1}), \\ u_2 & = & \ddot{\tilde{y}}_{d2} - k_1^2(x_3 - \dot{\tilde{y}}_{d2}) - k_0^2(x_2 - \tilde{y}_{d2}), \\ u_3 & = & \tilde{y}_{d3}^{(3)} - k_2^3(x_6 - \ddot{\tilde{y}}_{d3}) - k_1^3(x_5 - \dot{\tilde{y}}_{d3}) - k_0^3(x_4 - \tilde{y}_{d3}), \end{array}$$

where $s + k_0^1$, $s^2 + k_1^2 s + k_0^2$ and $s^3 + k_2^3 s^2 + k_1^3 s + k_0^3$ are Hurwitz polynomials. □

4. Conclusion and final remarks. We have introduced the notion of GOT for nonlinear systems and have shown that the linear algebraic framework introduced by Di Benedetto, Grizzle, and Moog [6] provides a rather convenient tool to study their properties. In particular, it has been shown that the rank of a system remains unchanged under such transformations.

It is worth mentioning that the class of GOT that we have introduced can be seen as the "dual" transformation of the class of generalized state feedbacks introduced in [14] and studied from the differential algebraic viewpoint in [5], where they were called *quasi–static* state feedbacks.

As an important application of this new class of transformations, a larger class of systems which are equivalent to prime systems has been identified. In turn, this result is applicable to control problems where output transformations are naturally allowed, such as output tracking and output regulation.

Of course, the conditions of Theorems 3.3 and 3.16 imply that the system Σ is *invertible* [6, 7], and hence noninteracting control can be achieved by dynamic state feedback. In that respect, Theorems 3.3 and 3.16 avoid the addition of extra dynamics to the system, as pointed out in [10].

An open issue for further research is the study of the notion of equivalence of nonlinear systems under GOTs, not necessarily proper.

REFERENCES

[1] E. Aranda-Bricaire and Ü. Kotta, *Equivalence of discrete-time nonlinear systems to prime systems*, J. Math. Systems Estim. Control, 8(1998).

[2] E. Aranda-Bricaire, C. H. Moog, and J.-B. Pomet, *A linear algebraic framework for dynamic feedback linearization*, IEEE Trans. Automat. Control, 40 (1995), pp. 127–132.

[3] E. Aranda-Bricaire, C. H. Moog, and J.-B. Pomet, *Infinitesimal Brunovsky form for nonlinear systems with applications to dynamic linearization*, in Geometry in Nonlinear Control and Differential Inclusions, B. Jakubczyk, W. Respondek, eds., Banach Center Publications Vol. 32, Institute of Mathematics, Polish Academy of Sciences, Warszawa, 1995, pp. 19–33.

[4] Y. Choquet-Bruhat, C. DeWitt-Morette, and M. Dillard-Bleick, *Analysis, Manifolds and Physics, Part* I: *Basics*, North-Holland, Amsterdam, 1989.

[5] E. Delaleau and M. Fliess, *Algorithme de structure, filtrations et découplage*, C. R. Acad. Sci. Paris Sér. I Math., 315 (1992), pp. 101–106.

[6] M. D. Di Benedetto, J. W. Grizzle, and C. H. Moog, *Rank invariants of nonlinear systems*, SIAM J. Control Optim., 27 (1989), pp. 658–672.

[7] M. Fliess, *Automatique et corps différentiels*, Forum Math., 1 (1989), pp. 227-238.

[8] A. Glumineau and C. H. Moog, *The essential orders and nonlinear decoupling*, Internat. J. Control, 50 (1989), pp. 1825–1834.

[9] A. Isidori, *Nonlinear Control Systems*, 2nd ed., Springer-Verlag, Heidelberg, 1989.

[10] R. Marino, W. Respondek, and A. J. van der Schaft, *Equivalence of nonlinear systems to input-output prime forms*, SIAM J. Control Optim., 32 (1994), pp. 387–407.

[11] C. H. Moog, *Nonlinear decoupling and structure at infinity*, Math. Control Signals Systems, 1 (1988), pp. 257–268.

[12] A. S. Morse, *Structural invariants of linear multivariable systems*, SIAM J. Control Optim., 11 (1973), pp. 446–465.

[13] H. Nijmeijer and A.J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.

[14] A. M. Perdon, G. Conte, and C. H. Moog, *Some canonical properties of nonlinear systems*, in Realization and Modelling in System Theory, M.A. Kaashoek, J. H. van Schuppen, A.C.M. Ran, eds., Progress in Systems and Control Theory, Vol. 3, Birkhäuser, Boston, 1990, pp. 89–96.

[15] C. Rogers and W. F. Shadwick, *Bäcklund Transformations and Their Applications*, Math. Sci. Engrg. 161, Academic Press, New York, 1982.

SIAM J. Control Optim.
Vol. 37, No. 1, pp. 131–164

QUADRATIC OPTIMAL CONTROL OF WELL-POSED LINEAR SYSTEMS*

OLOF J. STAFFANS†

Abstract. We study the infinite horizon quadratic cost minimization problem for a well-posed linear system in the sense of Salamon and Weiss. The quadratic cost function that we seek to minimize need not be positive, but it is convex and bounded from below. We assume the system to be jointly stabilizable and detectable and give a feedback solution to the cost minimization problem. Moreover, we connect this solution to the computation of either a (J, S)-inner or an S-normalized coprime factorization of the transfer function, depending on how the problem is formulated. We apply the general theory to get factorization versions of the bounded and positive real lemmas. In the case where the system is regular it is possible to show that the feedback operator can be expressed in terms of the Riccati operator and that the Riccati operator is a stabilizing self-adjoint solution of an algebraic Riccati equation. This Riccati equation is nonstandard in the sense that the weighting operator in the quadratic term differs from the expected one, and the computation of the correct weighting operator is a nontrivial task.

AMS subject classifications. 49J27, 47A68, 93A05

Key words. spectral factorization, (J, S)-inner-outer factorization, (J, S)-inner coprime factorization, S-normalized coprime factorization, algebraic Riccati equation

PII. S0363012996314257

1. Introduction.

This work treats the infinite horizon quadratic cost minimization problem for a time-invariant well-posed linear control system in the sense of Salamon and Weiss and extends the results presented in [23] to unstable systems. The approach is the same as in [22]: we first employ a preliminary state feedback to stabilize the system, and then we apply the theory developed in [23] to solve the quadratic cost minimization problem for the stable system. Working backwards we then obtain a solution to the original problem.

We consider two different types of cost functions. In the standard case both the control and the observation are equally penalized; we show that this leads to a problem that is equivalent to the computation of a normalized coprime factorization of the transfer function (see Corollary 4.9). It is possible to embed this type of problem into a more general class of problem where there is no cost on the control itself, only on the observation. In this setting the problem of quadratic cost minimization becomes equivalent to the computation of an inner coprime factorization of the transfer function, i.e., a coprime factorization with an inner numerator (see Theorem 4.4).

The infinite horizon quadratic cost minimization problem is also associated with an algebraic Riccati equation. Indeed, we show that in the case where the optimally controlled system and its adjoint are regular in the sense of Weiss, the Riccati operator satisfies an algebraic Riccati equation, and the feedback operator can be computed from the Riccati operator. However, in this connection we encounter a very interesting phenomenon: the weighting operator in the quadratic term of the Riccati equation differs from the expected one, and the computation of the correct weighting operator is a nontrivial task. The same operator is present in the formula that connects the

*Received by the editors May 3, 1995; accepted for publication (in revised form) July 10, 1997; published electronically September 25, 1998.
http://www.siam.org/journals/sicon/37-1/31425.html

†Åbo Akademi University, Department of Mathematics, FIN-20500 Åbo, Finland (Olof.Staffans@abo.fi, http://www.abo.fi/~staffans/).

Riccati operator to the feedback operator. This phenomenon was first reported in [18] in a stable setting for a more restricted class of transfer functions. Examples where this phenomenon occurs are given in [21], [22], [30], and [33].

We have based the discussion above on transfer functions rather than input/output maps since we believe that the former concept is more familiar to most readers. However, in the main body of the text we phrase our results in terms of input/output maps instead. In our opinion, this formulation is both easier and more intuitive than the transfer function formulation, and it has the advantage that generalizations to nonlinear and time-dependent systems are more immediate.

For a more detailed account of the existing Riccati equation theory for various classes of systems we refer the reader to [14] (and its forthcoming new version) and to the review [13]. However, we have to mention the very interesting paper by Flandoli, Lasiecka, and Triggiani [5]. In that paper the observation operator is bounded, but the authors have told us that the results of that paper can be extended to some classes of unbounded observation operators. Their approach is quite different from ours. They do not assume that the system is stabilizable and detectable. On the other hand, they also do not prove that the optimal system is well posed in our sense. They make no study of the input/output behavior of the closed loop system, and in particular, they do not mention the all-pass property of the optimal closed loop system (see Remark 2.8). In our opinion, this is the most characteristic property of the closed loop system.

The results presented here were originally obtained in the spring of 1995, and they were circulated in the form of two preprints [18, 19] with the titles "Coprime factorizations and optimal control of abstract linear systems" and "The nonstandard quadratic cost minimization problem for abstract linear systems." The former preprint treated the "standard" cost minimization problem and the latter a "nonstandard" cost minimization problem, where the cost function contains a possibly indefinite weighting operator but is still bounded from below. The latter was a straightforward modification of the former, and it was not included in the original submission to SIAM. However, later work on the H^∞ minimax problem has proved that the inclusion of the indefinite weighting operator would improve the future reference value of this work significantly.[1] This was one of the reasons for a major revision that was carried out in late 1996.[2] In the meantime we received preprints of [32] and [33], which overlap our section 2. The problem studied in [33] is essentially the same as in [23], summarized here in section 2, plus a Riccati equation theory for stable systems. However, neither paper fully contains the other.

This work is very closely related to [24]; in fact, they were both part of the same original submission to SIAM. We expect the reader to have access to [24] and refer freely to results in that paper. In particular, we send the reader to [24] for a short presentation of the basic theory of well-posed linear systems.

We use the following notation:

$\mathcal{L}(U;Y)$, $\mathcal{L}(U)$: The set of bounded linear operators from U into Y or from U into itself, respectively.

I: The identity operator.

A^*: The (Hilbert space) adjoint of the operator A.

[1] This is due to the fact that it makes the formulae look identical to those that are valid in the H^∞-case, although the underlying assumptions are different. See [25] and [26].

[2] At the same time [24] was separated into an independent paper.

$A \geq 0$: A is (self-adjoint and) positive definite.

$A \gg 0$: $A \geq \epsilon I$ for some $\epsilon > 0$, hence A is invertible.

$\mathrm{dom}(A)$: The domain of the (unbounded) operator A.

$\mathrm{range}(A)$: The range of the operator A.

$\mathbf{R}$, $\mathbf{R}^+$, $\mathbf{R}^-$: $\mathbf{R} = (-\infty, \infty)$, $\mathbf{R}^+ = [0, \infty)$, and $\mathbf{R}^- = (-\infty, 0]$.

$L^2(J;U)$: The set of U-valued L^2-functions on the interval J.

$L^2_\omega(J;U)$: $L^2_\omega(J;U) = \left\{ u \in L^2_{\mathrm{loc}}(J;U) \mid (t \mapsto \mathrm{e}^{-\omega t} u(t)) \in L^2(J;U) \right\}$.

$H^\infty_\omega(U;Y)$: The set of $\mathcal{L}(U;Y)$-valued H^∞ functions over the half-plane $\Re z > \omega$.

$TI_\omega(U;Y)$, $TI_\omega(U)$: The set of bounded linear time-invariant operators from $L^2_\omega(\mathbf{R};U)$ into $L^2_\omega(\mathbf{R};Y)$ or from $L^2_\omega(\mathbf{R};U)$ into itself.

$TIC_\omega(U;Y)$, $TIC_\omega(U)$: The set of causal operators in $TI_\omega(U;Y)$ or $TI_\omega(U)$.

$TIC(U;Y)$, $TIC(U)$: $TIC(U;Y) = TIC_0(U;Y)$ and $TIC(U) = TIC_0(U)$.

$\langle \cdot, \cdot \rangle_H$: The inner product in the Hilbert space H.

$\tau(t)$: The time-shift group $\tau(t)u(s) = u(t+s)$ (this is a left shift when $t > 0$ and a right shift when $t < 0$).

π_J: $(\pi_J u)(s) = u(s)$ if $s \in J$ and $(\pi_J u)(s) = 0$ if $s \notin J$. Here $J \subset \mathbf{R}$.

π_+, π_-: $\pi_+ = \pi_{\mathbf{R}^+}$ and $\pi_- = \pi_{\mathbf{R}^-}$.

We extend a L^2_ω-function u defined on a subinterval J of $\mathbf{R}$ to the whole real line by requiring u to be zero outside of J, and we denote the extended function by $\pi_J u$. We use the same symbol π_J both for the embedding operator $L^2_\omega(J) \to L^2_\omega(\mathbf{R})$ and for the corresponding projection operator $L^2_\omega(\mathbf{R}) \to L^2_\omega(J)$. With this interpretation, $\pi_J L^2_\omega(\mathbf{R};U) = L^2_\omega(J;U) \subset L^2_\omega(\mathbf{R};U)$ for each interval $J \subset \mathbf{R}$.

2. The stable quadratic cost minimization problem. Before looking at the general quadratic cost minimization problem for unstable systems, let us recall some basic results valid for stable systems.

Definition 2.1. *Let $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$ be a stable well-posed linear system on (U,H,Y)* [23, Definition 1], *and let $J = J^* \in \mathcal{L}(Y)$. The quadratic cost minimization problem for Ψ with cost operator J consists of finding, for each $x_0 \in H$, the infimum over all $u \in L^2(\mathbf{R}^+;U)$ of the cost*

$$Q(x_0, u) = \langle y, Jy \rangle_{L^2(\mathbf{R}^+;Y)}, \tag{2.1}$$

where $y = \mathcal{C}x_0 + \mathcal{D}\pi_+ u$ is the observation of Ψ with initial value $x_0 \in H$ and control $u \in L^2(\mathbf{R}^+;U)$. If there exists an operator $\Pi = \Pi^* \in \mathcal{L}(H)$ such that the optimal cost is given by

$$\inf_{u \in L^2(\mathbf{R}^+;U)} Q(x_0, u) = \langle x_0, \Pi x_0 \rangle_H,$$

then Π is called the Riccati operator of Ψ with cost operator J.

We have studied this problem in [23], but unfortunately, at that time we took the operator J to be the identity operator throughout. If J is positive definite, then it is possible to reduce J to the identity by a simple change of variable in the output space Y, but many applications, such as the positive (real) lemma and the bounded (real) lemma, require the use of a nondefinite J.[3] Fortunately, it turns out that the results

[3] We shall return elsewhere to the H^∞ theory which requires both a nondefinite cost operator J and a nondefinite sensitivity operator S.

presented in [23] remain valid with trivial modifications as long as the input/output map $\mathcal{D}$ of Ψ is J-coercive in the following sense:

DEFINITION 2.2. *Let $J = J^* \in \mathcal{L}(Y)$.*

(i) *The operator $\mathcal{D} \in TIC(U;Y)$ is J-coercive iff $\mathcal{D}^* J\mathcal{D} \gg 0$, that is, $\langle \mathcal{D}u, J\mathcal{D}u\rangle_{L^2(\mathbf{R};Y)} \geq \epsilon \|u\|^2_{L^2(\mathbf{R};U)}$ for all $u \in L^2(\mathbf{R};U)$ and some $\epsilon > 0$.*

(ii) *A stable well-posed linear system $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$ is J-coercive iff its input-output map $\mathcal{D}$ is J-coercive.*

Indeed, this is the case that is important in the applications to the bounded and positive (real) lemmas in section 8.

Since the solution to the cost minimization problem in the stable J-coercive case is almost identical to the one in [23] we simply present this solution below, leaving the proofs to the reader (it is done by inserting the operator J or S after each adjoint operator defined on Y or U, respectively).

DEFINITION 2.3 (see [23, Definitions 16 and 17]). *Let $J = J^* \in \mathcal{L}(Y)$, and let $S = S^* \in \mathcal{L}(U)$.*

(i) *The operator $\mathcal{N} \in TIC(U;Y)$ is (J,S)-inner iff $\mathcal{N}^* J \mathcal{N} = S$.*

(ii) *An operator $\mathcal{X} \in TIC(U;Y)$ is outer if the image of $L^2(\mathbf{R}^+;U)$ under $\mathcal{X}\pi_+$ is dense in $L^2(\mathbf{R}^+;Y)$.*

(iii) *An operator $\mathcal{X} \in TIC(U)$ is an (invertible) S-spectral factor of $\mathcal{D}^* J\mathcal{D} \in TI(U)$ iff $\mathcal{X}$ is invertible in $TIC(U)$ and $\mathcal{D}^* J \mathcal{D} = \mathcal{X}^* S \mathcal{X}$.*

(iv) *The factorization $\mathcal{D} = \mathcal{N}\mathcal{X}$ is a (J,S)-inner-outer factorization of $\mathcal{D} \in TIC(U;Y)$ if $\mathcal{N} \in TIC(U;Y)$ is (J,S)-inner and $\mathcal{X} \in TIC(U)$ is outer.*

(v) *In each case S is called the sensitivity operator of $\mathcal{N}$ or of the factorization.*

LEMMA 2.4 (see [23, Lemmas 13 and 18]). *Let $\mathcal{D} \in TIC(U;Y)$, $J = J^* \in \mathcal{L}(Y)$, $S \in \mathcal{L}(U)$, $S \gg 0$, $\widetilde{S} \in \mathcal{L}(U)$, and $\widetilde{S} \gg 0$.*

(i) *$\mathcal{D}^* J \mathcal{D}$ has an S-spectral factor $\mathcal{X}$ iff $\mathcal{D}$ is J-coercive.*

(ii) *If $\mathcal{X}$ is an S-spectral factor of $\mathcal{D}^* J\mathcal{D}$, then $\mathcal{N}\mathcal{X} = \left(\mathcal{D}\mathcal{X}^{-1}\right)\mathcal{X}$ is a (J,S)-inner-outer factorization of $\mathcal{D}$. Conversely, if $\mathcal{D}$ is J-coercive and $\mathcal{N}\mathcal{X}$ is a (J,S)-inner-outer factorization of $\mathcal{D}$, then $\mathcal{X}$ is an S-spectral factor of $\mathcal{D}^* J\mathcal{D}$.*

(iii) *The set of all possible S-spectral factors $\mathcal{X}$ of $\mathcal{D}^* J \mathcal{D}$ can be parameterized as $\mathcal{X} = E^{-1}\widetilde{\mathcal{X}}$ and $S = E^* \widetilde{S} E$, where $\widetilde{\mathcal{X}}$ is a fixed $\widetilde{S}$-spectral factor and $E \in \mathcal{L}(U)$ is an arbitrary invertible operator.*

(iv) *If $\mathcal{D}$ is J-coercive, then the Toeplitz operator $\pi_+\mathcal{D}^* J\mathcal{D}\pi_+$ is invertible, and its inverse can be written in the form $(\pi_+\mathcal{D}^* J \mathcal{D}\pi_+)^{-1} = \mathcal{X}^{-1}S^{-1}\pi_+(\mathcal{X}^*)^{-1}$. Here $\mathcal{X}$ is an arbitrary S-spectral factor of $\mathcal{D}^* J\mathcal{D}$. ($\mathcal{X}^{-1}S^{-1}\pi_+(\mathcal{X}^*)^{-1}$ does not depend on the particular factorization, only on $\mathcal{D}$ and J.)*

LEMMA 2.5 (see [23, Lemma 13 and Theorem 27]). *Let $J = J^* \in \mathcal{L}(Y)$, and let $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$ be a stable J-coercive well-posed linear system on (U,H,Y). Then, for each $x_0 \in H$, there is a unique control $u^{\mathrm{opt}}(x_0) \in L^2(\mathbf{R}^+;U)$ that minimizes the cost function $Q(x_0,u)$ in Definition 2.1. This control u^{opt} is given by*

$$u^{\mathrm{opt}}(x_0) = -\mathcal{X}^{-1}S^{-1}\pi_+\mathcal{N}^* J\mathcal{C}x_0,$$

where $\mathcal{N}\mathcal{X}$ is an arbitrary (J,S)-inner-outer factorization of $\mathcal{D}$ (cf. Lemma 2.4). The corresponding state $x^{\mathrm{opt}}(x_0)$, output $y^{\mathrm{opt}}(x_0)$, and the minimum $Q(x_0,u^{\mathrm{opt}}(x_0))$ of the cost function are given by

$$\begin{aligned} x^{\mathrm{opt}}(x_0) &= \mathcal{A}x_0 - \mathcal{B}\mathcal{X}^{-1}\tau S^{-1}\pi_+\mathcal{N}^* J\mathcal{C}x_0, \\ y^{\mathrm{opt}}(x_0) &= (I-P)\mathcal{C}x_0, \\ Q(x_0,u^{\mathrm{opt}}(x_0)) &= \langle x_0, \mathcal{C}^* J\,(I-P)\,\mathcal{C}x_0\rangle_H\,, \end{aligned}$$

where

$$P = \mathcal{D}\pi_+(\pi_+\mathcal{D}^* J\mathcal{D}\pi_+)^{-1}\pi_+\mathcal{D}^* J = I - \mathcal{N}S^{-1}\pi_+\mathcal{N}^* J$$

is the projection onto the range of $\mathcal{D}\pi_+$ *along the null space of* $\pi_+\mathcal{D}^* J$. *In particular,* Ψ *has a Riccati operator, namely*

$$\Pi = \mathcal{C}^* J\,(I - P)\,\mathcal{C},$$

and $y^{\mathrm{opt}}(x_0)$ *belongs to the null space of the projection* P, *i.e.,*

$$\pi_+\mathcal{D}^* J y^{\mathrm{opt}}(x_0) = \pi_+\mathcal{D}^* J\left(\mathcal{C}x_0 + \mathcal{D}\pi_+ u^{\mathrm{opt}}(x_0)\right) = 0.$$

We remark that, although the factorization $\mathcal{D} = \mathcal{N}\mathcal{X}$ and the operator S are not unique, the formulas given above produce the same result independently of how the factorization is chosen. This follows from Lemma 2.4 (see, in particular, part (iv)).

THEOREM 2.6 (see [23, Theorem 27]). *Let* $J = J^* \in \mathcal{L}(Y)$, *and let* $\Psi = \left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$ *be a stable J-coercive well-posed linear system on* (U, H, Y). *Let* $x_0 \in H$, *let* $x^{\mathrm{opt}}(x_0)$, $y^{\mathrm{opt}}(x_0)$, *and* $u^{\mathrm{opt}}(x_0)$ *be the optimal state, output, and control for the quadratic cost minimization problem, and let* Π *be the corresponding Riccati operator* (*see Lemma* 2.5).

(i) *Let* $\mathcal{D} = \mathcal{N}\mathcal{X}$ *be a* (J,S)*-inner-outer factorization of* $\mathcal{D}$, *and define* $\mathcal{M} = \mathcal{X}^{-1}$. *Then*

$$\begin{bmatrix}\mathcal{K} & \mathcal{F}\end{bmatrix} = \begin{bmatrix}-S^{-1}\pi_+\mathcal{N}^* J\mathcal{C} & (I - \mathcal{X})\end{bmatrix}$$

is a stable and stabilizing state feedback pair for Ψ [23, *Definition* 22] *and*

$$\begin{aligned}\begin{bmatrix}x^{\mathrm{opt}}(t,x_0)\\ y^{\mathrm{opt}}(x_0)\\ u^{\mathrm{opt}}(x_0)\end{bmatrix} &= \begin{bmatrix}\mathcal{A}_{\circlearrowleft}(t)\\ \mathcal{C}_{\circlearrowleft}\\ \mathcal{K}_{\circlearrowleft}\end{bmatrix} x_0 = \begin{bmatrix}\mathcal{A}(t) + \mathcal{B}\mathcal{M}\tau(t)\mathcal{K}\\ \mathcal{C} + \mathcal{N}\mathcal{K}\\ \mathcal{M}\mathcal{K}\end{bmatrix} x_0\\ &= \begin{bmatrix}\mathcal{A}(t)\\ \mathcal{C}\\ 0\end{bmatrix} x_0 - \begin{bmatrix}\mathcal{B}\mathcal{M}\tau(t)\\ \mathcal{N}\\ \mathcal{M}\end{bmatrix} S^{-1}\pi_+\mathcal{N}^* J\mathcal{C}x_0\end{aligned}$$

is equal to the state and output of the closed loop system $\Psi_{\circlearrowleft}$ *defined by*

$$\Psi_{\circlearrowleft} = \begin{bmatrix}\mathcal{A}_{\circlearrowleft} & \mathcal{B}_{\circlearrowleft}\\ \begin{bmatrix}\mathcal{C}_{\circlearrowleft}\\ \mathcal{K}_{\circlearrowleft}\end{bmatrix} & \begin{bmatrix}\mathcal{D}_{\circlearrowleft}\\ \mathcal{F}_{\circlearrowleft}\end{bmatrix}\end{bmatrix} = \begin{bmatrix}\mathcal{A} + \mathcal{B}\tau\mathcal{M}\mathcal{K} & \mathcal{B}\mathcal{M}\\ \begin{bmatrix}\mathcal{C} + \mathcal{N}\mathcal{K}\\ \mathcal{M}\mathcal{K}\end{bmatrix} & \begin{bmatrix}\mathcal{N}\\ \mathcal{M} - I\end{bmatrix}\end{bmatrix}$$

with initial value x_0, *initial time zero, and zero control* $u_{\circlearrowleft}$ (*see Figure* 2.1). *The Riccati operator* Π *of* Ψ *can be written in the following alternative forms*:

$$\Pi = \mathcal{C}^* J\mathcal{C} - \mathcal{K}^* S\mathcal{K} = \mathcal{C}^* J\mathcal{C}_{\circlearrowleft} = \mathcal{C}_{\circlearrowleft}^* J\mathcal{C}_{\circlearrowleft} = \mathcal{C}_{\circlearrowleft}^* J\mathcal{C}.$$

(ii) *Conversely, suppose that* $\left[\begin{smallmatrix}y^{\mathrm{opt}}(x_0)\\ u^{\mathrm{opt}}(x_0)\end{smallmatrix}\right]$ *is equal to the observation of some stable state feedback perturbation* $\Psi_{\circlearrowleft}$ *of* Ψ *with initial value* x_0, *initial time zero, zero control* $u_{\circlearrowleft}$, *and some admissible stable state feedback pair* $[\mathcal{K} \quad \mathcal{F}]$. *Then there exists an operator* $S \in \mathcal{L}(U)$, $S \gg 0$, *such that* $\mathcal{N}\mathcal{X}$ *is a* (J,S)*-inner-outer factorization of* $\mathcal{D}$, *where* $\mathcal{N} = \mathcal{D}\,(I - \mathcal{F})^{-1}$ *and* $\mathcal{X} = (I - \mathcal{F})$. *Moreover,* $\mathcal{K}$ *is given by* $\mathcal{K} = -S^{-1}\pi_+\mathcal{N}^* J\mathcal{C}$.

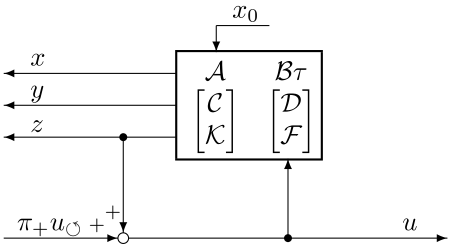

x_0
x
y
z
$\mathcal{A}$
$\mathcal{B}_\mathcal{T}$
$\begin{bmatrix}\mathcal{C}\\\mathcal{K}\end{bmatrix}$
$\begin{bmatrix}\mathcal{D}\\\mathcal{F}\end{bmatrix}$
$\pi_+u_\circlearrowright$ +
+
u

FIG. 2.1. *Optimal state feedback connection* $\Psi_\circlearrowright$ *in Theorem* 2.6.

(iii) *If* $y = \mathcal{C}_\circlearrowright x_0 + \mathcal{D}_\circlearrowright \pi_+ u_\circlearrowright$ *is the first output of the optimal closed loop system* $\Psi_\circlearrowright$ *with initial state* $x_0 \in H$ *and control* $u_\circlearrowright \in L^2(\mathbf{R}^+;U)$ (*see Figure* 2.1), *then the closed loop cost* $Q_\circlearrowright(x_0, u_\circlearrowright)$ *is given by*

$$Q_\circlearrowright(x_0,u_\circlearrowright) = \langle y, Jy\rangle_{L^2(\mathbf{R}^+;Y)} = \langle x_0, \Pi x_0\rangle_H + \langle u_\circlearrowright, S u_\circlearrowright\rangle_{L^2(\mathbf{R}^+;Y)}\,. \tag{2.2}$$

Proof. Only (iii) requires a proof, since this identity is not found in [23]. This proof goes as follows (the last equality follows from Lemma 2.5):

$$\begin{aligned}\langle y, Jy\rangle_{L^2(\mathbf{R}^+;Y)} &= \big\langle (y^{\mathrm{opt}}(x_0) + \mathcal{N}\pi_+u_\circlearrowright)(s), J(y^{\mathrm{opt}}(x_0) + \mathcal{N}\pi_+u_\circlearrowright)(s)\big\rangle_{L^2(\mathbf{R}^+;Y)}\\ &= \big\langle y^{\mathrm{opt}}(x_0), Jy^{\mathrm{opt}}(x_0)\big\rangle_{L^2(\mathbf{R}^+;Y)}\\ &\quad + 2\Re\big\langle \mathcal{X}\pi_+u_\circlearrowright(s), \mathcal{D}^*Jy^{\mathrm{opt}}(x_0)\big\rangle_{L^2(\mathbf{R}^+;U)}\\ &\quad + \langle u_\circlearrowright, \mathcal{N}^*J\mathcal{N}\pi_+u_\circlearrowright\rangle_{L^2(\mathbf{R}^+;U)}\\ &= \langle x_0, \Pi x_0\rangle_H + \langle u_\circlearrowright, Su_\circlearrowright\rangle_{L^2(\mathbf{R}^+;U)}\,. \qquad \square\end{aligned}$$

REMARK 2.7. *This theorem is actually true under weaker stability assumptions. It suffices if* $\mathcal{C}$ *and* $\mathcal{D}$ *are stable, i.e.,* $\mathcal{A}$ *and* $\mathcal{B}$ *need not be stable* [24, *Definition* 2.11]. *Of course, the corresponding closed loop* $\mathcal{A}_\circlearrowright$ *and* $\mathcal{B}_\circlearrowright$ *need not be stable in this case. Stability of* $\mathcal{A}$ *was not assumed in* [23], *and the stability of* $\mathcal{B}$ *was never used in a nontrivial way in the proofs* (*although it was assumed*). *See also* [33] *which requires no stability of* $\mathcal{A}$ *and* $\mathcal{B}$.

REMARK 2.8. *The conclusion of part* (iii) *of Theorem* 2.6 *says that the frequency response of the input/output map from the closed loop control* $u_\circlearrowright$ *in Figure* 2.1 *to the original output is completely flat, i.e., this input/output map is all-pass, with a power amplification level equal to* S. *Thus,* S *measures the sensitivity of the closed loop system with respect to deviations from the optimal strategy. This is the reason why we call* S *the* sensitivity operator *of the closed loop system.*

3. Quadratic cost minimization: Reduction to the stable case.

We are now ready to attack the unstable quadratic cost minimization problem. The definition of the problem is essentially the same as in the stable case.

DEFINITION 3.1. *Let* $\Psi = \left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$ *be a well-posed linear system on* (U,H,Y), *and let* $J = J^* \in \mathcal{L}(Y)$. *The* (*nonstandard*) *quadratic cost minimization problem for* Ψ *with cost operator* J *consists of finding, for each* $x_0 \in H$, *the infimum of the cost* $Q(x_0,u)$ *defined in* (2.1) *over all those* $u \in L^2(\mathbf{R}^+;U)$ *for which the corresponding observation* $y = \mathcal{C}x_0 + \mathcal{D}\pi_+u$ *of* Ψ *satisfies* $y \in L^2(\mathbf{R}^+;Y)$. *If there exists an operator* $\Pi = \Pi^* \in \mathcal{L}(H)$ *such that the optimal cost is given by*

$$\inf_{u\in L^2(\mathbf{R}^+;U)} Q(x_0,u) = \langle x_0, \Pi x_0\rangle_H\,,$$

then Π *is called the Riccati operator of* Ψ *with cost operator* J.

Clearly, Q is a quadratic, possibly unbounded, function of $u \in L^2(\mathbf{R}^+;U)$ due to the fact that $\mathcal{D}\pi_+$ is a linear, possibly unbounded, operator in $L^2(\mathbf{R}^+;U)$. The latter operator is not bounded on $L^2(\mathbf{R}^+;U)$ unless Ψ is input-output stable, but it is always closed.

LEMMA 3.2. *Let* $\mathcal{D} \in TIC_\alpha(U;Y)$ *for some* $\alpha \geq 0$. *Then the restriction* $\mathcal{D}_0$ *of the Toeplitz operator* $\mathcal{D}\pi_+$ *to the domain*

$$\operatorname{dom}(\mathcal{D}_0) = \left\{ u \in L^2(\mathbf{R}^+;U) \mid \mathcal{D}\pi_+ u \in L^2(\mathbf{R}^+;Y) \right\}$$

is a closed (possibly unbounded) linear operator from $\operatorname{dom}(\mathcal{D}_0) \subset L^2(\mathbf{R}^+;U)$ *into* $L^2(\mathbf{R}^+;Y)$.

Proof. This follows directly from the fact that $L^2(\mathbf{R}^+)$ is continuously imbedded in $L^2_\alpha(\mathbf{R}^+)$. □

We can say something more about how $\mathcal{D}_0$ maps $L^2(\mathbf{R}^+;U)$ into $L^2(\mathbf{R}^+;Y)$ in the case where $\mathcal{D}$ has a right coprime factorization.[4]

LEMMA 3.3. *Let* $\mathcal{D} \in TIC_\alpha(U;Y)$ *for some* $\alpha \geq 0$, *and suppose that* $\mathcal{D}$ *has a right coprime factorization* $(\mathcal{N},\mathcal{M})$ [24, *Definition* 4.2].

(i) *If* $u \in L^2_{\mathrm{loc}}(\mathbf{R}^+;U)$, $u_\flat \in L^2_{\mathrm{loc}}(\mathbf{R}^+;Y)$, *and* $y \in L^2_{\mathrm{loc}}(\mathbf{R}^+;Y)$ *satisfy*

$$u = \mathcal{M}\pi_+ u_\flat \text{ and } y = \mathcal{N}\pi_+ u_\flat,$$

then $u_\flat \in L^2(\mathbf{R}^+;U)$ *iff both* $u \in L^2(\mathbf{R}^+;U)$ *and* $y \in L^2(\mathbf{R}^+;Y)$. *Thus,* $\operatorname{dom}(\mathcal{D}_0)$ *is equal to the image of* $L^2(\mathbf{R}^+;U)$ *under* $\mathcal{M}\pi_+$, *and* $\operatorname{range}(\mathcal{D}_0)$ *is equal to the image of* $L^2(\mathbf{R}^+;U)$ *under* $\mathcal{N}\pi_+$. *In particular,* $\operatorname{dom}(\mathcal{D}_0)$ *is dense in* $L^2(\mathbf{R}^+;U)$ *iff* $\mathcal{M}$ *is outer.*

(ii) *With* u, $u_\flat$, *and* y *as above, there exist strictly positive constants* ϵ *and* M *such that*

$$\begin{aligned}\epsilon\left(\|u\|^2_{L^2(\mathbf{R}^+;U)} + \|y\|^2_{L^2(\mathbf{R}^+;Y)}\right) &\leq \|u_\flat\|^2_{L^2(\mathbf{R}^+;U)} \\ &\leq M\left(\|u\|^2_{L^2(\mathbf{R}^+;U)} + \|y\|^2_{L^2(\mathbf{R}^+;Y)}\right).\end{aligned}$$

Proof. Clearly, if $u_\flat \in L^2(\mathbf{R}^+;U)$, then both $u \in L^2(\mathbf{R}^+;U)$ and $y \in L^2(\mathbf{R}^+;Y)$. Conversely, if both $u \in L^2(\mathbf{R}^+;U)$ and $y \in L^2(\mathbf{R}^+;Y)$, then we can use the right coprimeness of $\mathcal{N}$ and $\mathcal{M}$ to write

$$u_\flat = \left(\widetilde{\mathcal{Y}}\mathcal{N} + \widetilde{\mathcal{X}}\mathcal{M}\right)\pi_+ u_\flat = \widetilde{\mathcal{Y}}y + \widetilde{\mathcal{X}}u,$$

and this implies that $u_\flat \in L^2(\mathbf{R}^+;U)$. The claims about the domain and range of $\mathcal{D}_0$ follow immediately, and so does claim (ii). □

As a special case of this result (take $\mathcal{N} = \mathcal{D}$ and $\mathcal{M} = I$) we get the following (trivial) estimate.

LEMMA 3.4. *For each* $\mathcal{D} \in TIC(U;Y)$ *there exist strictly positive constants* ϵ *and* M *such that, for all* $u \in L^2(\mathbf{R};U)$,

$$\epsilon\left(\|u\|^2_{L^2(\mathbf{R};U)} + \|\mathcal{D}u\|^2_{L^2(\mathbf{R};Y)}\right) \leq \|u\|^2_{L^2(\mathbf{R};U)} \leq \|u\|^2_{L^2(\mathbf{R};U)} + \|\mathcal{D}u\|^2_{L^2(\mathbf{R};Y)}.$$

A necessary and sufficient condition for the existence of a finite infimum for the nonstandard quadratic cost minimization problem is that the cost function Q is bounded from below as a function of u. This should be true for each fixed $x_0 \in H$. We shall actually impose a slightly stronger condition on Q which implies that not only does the infimum exist, but it is in fact a minimum.[5] However, before intro-

[4] This is true, e.g., when $\mathcal{D}$ is the input-output map of a jointly stabilizable and detectable well-posed linear system. See [24, Theorem 4.4].

[5] See Lemma 3.9.

ducing this condition, let us make the following simple observation about the stable case.

LEMMA 3.5. *Let $J = J^* \in \mathcal{L}(Y)$. The operator $\mathcal{D} \in TIC(U;Y)$ is J-coercive iff $\mathcal{D}^*J\mathcal{D} \geq \epsilon(\mathcal{D}^*\mathcal{D} + I)$ for some $\epsilon > 0$, i.e., $\langle \mathcal{D}u, J\mathcal{D}u\rangle_{L^2(\mathbf{R};Y)} \geq \epsilon\big(\|u\|^2_{L^2(\mathbf{R};U)} + \|\mathcal{D}u\|^2_{L^2(\mathbf{R};U)}\big)$ for all $u \in L^2(\mathbf{R};U)$.*

Proof. This follows from Definition 2.2 and Corollary 3.4. □

In the unstable case we turn the characterization of J-coercivity given in this lemma into a definition.

DEFINITION 3.6. *Let $J = J^* \in \mathcal{L}(Y)$, and let $\alpha \geq 0$.*

(i) *The operator $\mathcal{D} \in TIC_\alpha(U;Y)$ is J-coercive iff there exists a constant $\epsilon > 0$ such that*

$$\langle \mathcal{D}\pi_+u, J\mathcal{D}\pi_+u\rangle_{L^2(\mathbf{R}^+;Y)} \geq \epsilon\left(\|u\|^2_{L^2(\mathbf{R}^+;U)} + \|\mathcal{D}\pi_+u\|^2_{L^2(\mathbf{R}^+;Y)}\right)$$

for all those $u \in L^2(\mathbf{R}^+;U)$ for which $\mathcal{D}\pi_+u \in L^2(\mathbf{R}^+;Y)$.

(ii) *The system $\Psi = \left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$ on (U,H,Y) is J-coercive if there exist constants $M > 0$ and $\epsilon > 0$ such that the cost function Q defined in* (2.1) *satisfies*

$$Q(x_0,u) \geq \epsilon\left(\|u\|^2_{L^2(\mathbf{R}^+;U)} + \|y\|^2_{L^2(\mathbf{R}^+;Y)}\right) - M\,\|x_0\|^2_H \tag{3.1}$$

for all those $x_0 \in H$ and $u \in L^2(\mathbf{R}^+;U)$ for which $y = \mathcal{C}x_0 + \mathcal{D}\pi_+u \in L^2(\mathbf{R}^+;Y)$.

By Lemma 3.5, part (i) of Definition 2.2 is consistent with part (i) of Definition 3.6. That the second half of these definitions is also consistent follows from the next lemma.

LEMMA 3.7. *A stable system is J-coercive in the sense of Definition* 3.6 *iff its input-output map is J-coercive.*[6]

Proof. Trivially, the J-coercivity of an arbitrary system (stable or not) implies that its input-output map is J-coercive (take $x_0 = 0$).

Conversely, suppose that $\mathcal{D}$ is J-coercive, e.g., in the sense of Definition 2.2. For each $u \in L^2(\mathbf{R}^+;U)$ we have

$$\begin{aligned}&\langle \mathcal{D}\pi_+u + \mathcal{C}x_0, J(\mathcal{D}\pi_+u + \mathcal{C}x_0)\rangle_Y\\ &\qquad \geq \langle \mathcal{D}\pi_+u, J(\mathcal{D}\pi_+u)\rangle_Y - 2\,\|J\|\,\|\mathcal{D}\|\,\|\mathcal{C}\|\,\|u\|\,\|x_0\| - \|J\|\,\|\mathcal{C}\|^2\,\|x_0\|^2.\end{aligned}$$

Combining this with Lemma 3.5 and with the fact that for all positive constants a, b, and δ it is true that $2ab \leq \delta a^2 + (1/\delta)b^2$, we find that for some sufficiently large constant M, independent of u and x_0,

$$\langle \mathcal{D}\pi_+u + \mathcal{C}x_0, J(\mathcal{D}\pi_+u + \mathcal{C}x_0)\rangle_Y \geq \epsilon/2(\|u\|^2 + \|y\|^2) - M\,\|x_0\|^2.$$

Thus, the system is J-coercive in the sense of Definition 3.6. □

LEMMA 3.8. *Let $J = J^* \in \mathcal{L}(U;Y)$, and let $\mathcal{D} \in TIC_\alpha(U;Y)$ for some $\alpha \geq 0$. If $\mathcal{D}$ has a right coprime factorization $(\mathcal{N},\mathcal{M})$, then $\mathcal{D}$ is J-coercive iff $\mathcal{N}$ is J-coercive.*

Proof. This follows from Lemmas 3.3 and 3.5 and Definition 3.6. □

Our approach to the quadratic cost minimization problem is to first use a preliminary stabilizing feedback and to then minimize the stabilized problem. It is based on the following result.

[6] The same statement is actually true for all jointly stabilizable and detectable systems. See Lemma 3.9(iii).

LEMMA 3.9. *Let $J = J^* \in \mathcal{L}(Y)$, and let $\Psi = \left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$ be a well-posed linear system on (U, H, Y) with jointly stabilizing feedback and output injection pairs $[\mathcal{K}^1 \quad \mathcal{F}^1]$ and $\left[\begin{smallmatrix}\mathcal{H}\\ \mathcal{G}\end{smallmatrix}\right]$* [24, *Definition* 3.15]. *Let*

$$\begin{aligned}\Psi_\flat &= \begin{bmatrix} \mathcal{A}_\flat & \mathcal{B}_\flat \\ \begin{bmatrix}\mathcal{C}_\flat \\ \mathcal{K}^1_\flat\end{bmatrix} & \begin{bmatrix}\mathcal{D}_\flat \\ \mathcal{F}^1_\flat\end{bmatrix}\end{bmatrix} \\ &= \begin{bmatrix} \mathcal{A} + \mathcal{B}\tau\left(I - \mathcal{F}^1\right)^{-1}\mathcal{K}^1 & \mathcal{B}\left(I - \mathcal{F}^1\right)^{-1} \\ \begin{bmatrix}\mathcal{C} + \mathcal{D}\left(I - \mathcal{F}^1\right)^{-1}\mathcal{K}^1 \\ \left(I - \mathcal{F}^1\right)^{-1}\mathcal{K}^1\end{bmatrix} & \begin{bmatrix}\mathcal{D}\left(I - \mathcal{F}^1\right)^{-1} \\ \left(I - \mathcal{F}^1\right)^{-1} - I\end{bmatrix}\end{bmatrix}\end{aligned}$$

be the state feedback perturbed version of Ψ [24, *Lemma* 3.13] *with feedback pair $[\mathcal{K}^1 \quad \mathcal{F}^1]$.*

(i) *The output $y = \mathcal{C}x_0 + \mathcal{D}\pi_+ u$ of Ψ with initial value $x_0 \in H$ and control $u \in L^2_{\text{loc}}(\mathbf{R}^+; U)$ is equal to the first output $y = \mathcal{C}_\flat x_0 + \mathcal{D}_\flat \pi_+ u_\flat$ of $\Psi_\flat$ with the same initial value $x_0 \in H$ and control $u_\flat \in L^2_{\text{loc}}(\mathbf{R}^+; U)$ if we choose u and $u_\flat$ to satisfy*

$$u = \left(I - \mathcal{F}^1\right)^{-1}\left(\mathcal{K}^1 x_0 + \pi_+ u_\flat\right) = \mathcal{K}^1_\flat x_0 + \left(I + \mathcal{F}^1_\flat\right)\pi_+ u_\flat, \tag{3.2}$$

or equivalently,[7]

$$u_\flat = -\mathcal{K}^1 x_0 + \left(I - \mathcal{F}^1\right)\pi_+ u.$$

With this choice of u and $u_\flat$, the states $x(t) = \mathcal{A}(t)x_0 + \mathcal{B}\tau(t)\pi_+ u$ and $x(t) = \mathcal{A}_\flat(t)x_0 + \mathcal{B}_\flat\tau(t)\pi_+ u_\flat$ of the two systems are also equal for all $t \in \mathbf{R}^+$. Moreover, $u_\flat \in L^2(\mathbf{R}^+; U)$ iff both $y \in L^2(\mathbf{R}^+; Y)$ and $u \in L^2(\mathbf{R}^+; U)$, and there exists a constant M (independent of x_0, u, and $u_\flat$) such that

$$\begin{aligned}\|u\|^2_{L^2(\mathbf{R}^+;U)} &\le M\left(\|x_0\|^2_H + \|u_\flat\|^2_{L^2(\mathbf{R}^+;U)}\right), \\ \|u_\flat\|^2_{L^2(\mathbf{R}^+;U)} &\le M\left(\|x_0\|^2_H + \|y\|^2_{L^2(\mathbf{R}^+;Y)} + \|u\|^2_{L^2(\mathbf{R}^+;U)}\right).\end{aligned}$$

(ii) *The original system Ψ is J-coercive iff the feedback stabilized system $\Psi_\flat$ is so.*

(iii) *The original system Ψ is J-coercive iff its input/output map $\mathcal{D}$ is J-coercive.*

(iv) *If either (hence both) of the two systems is J-coercive, then the controls $u \in L^2(\mathbf{R}^+; U)$ of Ψ and $u_\flat \in L^2(\mathbf{R}^+; U)$ of $\Psi_\flat$ are uniquely determined by the initial state x_0 and the (first) output y. In particular, if the output $y = \mathcal{C}x_0 + \mathcal{D}\pi_+ u$ of Ψ with initial value x_0 and control $u \in L^2(\mathbf{R}^+; U)$ is equal to the first output $\mathcal{C}_\flat x_0 + \mathcal{D}_\flat \pi_+ u_\flat$ of $\Psi_\flat$ with initial value x_0 and control $u_\flat \in L^2(\mathbf{R}^+; U)$, then u and $u_\flat$ must satisfy* (3.2).

Proof. (i) The output of Ψ is given by $y = \mathcal{C}x_0 + \mathcal{D}\pi_+ u$ and the first output of $\Psi_\flat$ is given by $y = \mathcal{C}_\flat x_0 + \mathcal{D}_\flat u_\flat = (\mathcal{C} + \mathcal{D}(I - \mathcal{F}^1)^{-1}\mathcal{K}^1)x_0 + \mathcal{D}(I - \mathcal{F}^1)^{-1}\pi_+ u_\flat$, so we get the same output if we let u and $u_\flat$ satisfy (3.2). By [24, Theorem 4.4], $(\mathcal{D}_\flat, (I + \mathcal{F}^1_\flat))$ is a right coprime factorization of $\mathcal{D}$. Since it is possible to write the equations connecting u, $u_\flat$, and y in the form

$$\begin{aligned}u &= \left(I + \mathcal{F}^1_\flat\right)\pi_+ u_\flat + \mathcal{K}^1_\flat x_0, \\ y &= \mathcal{D}_\flat \pi_+ u_\flat + \mathcal{C}_\flat x_0,\end{aligned}$$

[7] See the equivalent [24, Figures 3.4 and 3.7].

and since (by the stability of $\Psi_\flat$) $\mathcal{K}_\flat x_0 \in L^2(\mathbf{R}^+;U)$ and $\mathcal{C}_\flat x_0 \in L^2(\mathbf{R}^+;Y)$, it follows from Lemma 3.3 that $u_\flat \in L^2(\mathbf{R}^+;U)$ iff both $y \in L^2(\mathbf{R}^+;Y)$ and $u \in L^2(\mathbf{R}^+;U)$. Moreover, the listed inequalities are true.

(ii) Suppose that Ψ is J-coercive. By the second of the two inequalities in part (i),

$$\epsilon/2 \left\|u\right\|^2_{L^2(\mathbf{R}^+;U)} + \epsilon/2 \left\|y\right\|^2_{L^2(\mathbf{R}^+;Y)} \geq -\epsilon/2 \left\|x_0\right\|^2_H + \epsilon/(2M) \left\|u_\flat\right\|^2_{L^2(\mathbf{R}^+;U)},$$

and this combined with (3.1) implies that $\Psi_\flat$ is J-coercive (replace M by $M + \epsilon/2$ and ϵ by $\min\{\epsilon/2, \epsilon/(2M)\}$). The proof of the converse part is similar but simpler.

(iii) This follows from part (ii) and Lemmas 3.7 and 3.8.

(iv) If the two controls $u^1_\flat$ and $u^2_\flat$ produce the same output $y = \mathcal{C}_\flat x_0 + \mathcal{D}_\flat u^1_\flat = \mathcal{C}_\flat x_0 + \mathcal{D}_\flat u^2_\flat$, then their difference $u^1_\flat - u^2_\flat$ satisfies $\mathcal{D}_\flat \pi_+(u^1_\flat - u^2_\flat) = 0$. As $\mathcal{D}_\flat$ is J-coercive, $\mathcal{D}_\flat \pi_+$ is one-to-one on $L^2(\mathbf{R}^+;U)$, and we find that $\pi_+(u^1_\flat - u^2_\flat) = 0$. Similarly, if the two controls u^1 and u^2 produce the same output $y = \mathcal{C}x_0 + \mathcal{D}\pi_+ u^1 = \mathcal{C}x_0 + \mathcal{D}\pi_+ u^2$, then their difference $u^1 - u^2$ satisfies $\mathcal{D}\pi_+(u^1 - u^2) = 0$. Define $z = (I + \mathcal{F}^1_\flat)^{-1}\pi_+(u^1 - u^2)$. Then $(I + \mathcal{F}^1_\flat)z = u^1 - u^2$ and $\mathcal{D}_\flat z = \mathcal{D}(u^1 - u^2) = 0$. Recall that $(\mathcal{D}_\flat, (I + \mathcal{F}^1_\flat))$ is a right coprime factorization of $\mathcal{D}$ [24, Theorem 4.4]. From Lemma 3.3 we conclude that $z \in L^2(\mathbf{R}^+;U)$, which combined with the J-coercivity of $\mathcal{D}_\flat$ implies that $z = 0$. Thus, $u^1 - u^2 = 0$ also. □

4. The solution to the unstable quadratic cost minimization problem.

Lemma 3.9 gives us the following preliminary solution to the general quadratic cost minimization problem.

LEMMA 4.1. *Let $J = J^* \in \mathcal{L}(Y)$, and let $\Psi = [\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}]$ be a jointly stabilizable and detectable J-coercive well-posed linear system on (U, H, Y). Then the quadratic cost minimization problem with cost operator J has a unique minimizing solution $u^{\mathrm{opt}}(x_0) \in L^2(\mathbf{R}^+;U)$. This solution can be computed as follows: We first feedback stabilize Ψ as described in Lemma* 3.9, *and we then apply Lemma* 2.5 *with Ψ replaced by the stabilized system $\Psi_\flat$ to get an optimal control $u^{\mathrm{opt}}_\flat(x_0)$, an optimal output $y^{\mathrm{opt}}(x_0)$, and an optimal state trajectory $x^{\mathrm{opt}}(x_0)$ for the stabilized system. The optimal control for the original system Ψ is then given by $u^{\mathrm{opt}}(x_0) = \mathcal{K}^1_\flat x_0 + (I + \mathcal{F}^1_\flat)\pi_+ u^{\mathrm{opt}}_\flat(x_0)$, and the optimal output and state for the original minimization problem is equal to the optimal output $y^{\mathrm{opt}}(x_0)$ and state $x^{\mathrm{opt}}(x_0)$ for the stabilized minimization problem. In particular, the original problem and the stabilized problem have the same Riccati operator.*

The solution given by Lemma 4.1 is not yet complete in the sense that it does not contain the same type of feedback description as Theorem 2.6 does for the stable case. Our next task will be to develop such a feedback description. This description will be given in terms of a right coprime factorization of the input/output map $\mathcal{D}$ with the special property that its numerator is (J, S)-inner. This notion is defined as follows.

DEFINITION 4.2. *Let $J = J^* \in \mathcal{L}(Y)$, let $S = S^* \in \mathcal{L}(U)$ be invertible, let $\mathcal{D} \in TIC_\alpha(U;Y)$ for some $\alpha \geq 0$, and let $(\mathcal{N}, \mathcal{M})$ be a right coprime factorization of $\mathcal{D}$ in TIC.*

(i) *If $\mathcal{N}$ is (J, S)-inner, then $(\mathcal{N}, \mathcal{M})$ is a (J, S)-inner right coprime factorization of $\mathcal{D}$.*

(ii) *If $[\begin{smallmatrix}\mathcal{N}\\ \mathcal{M}\end{smallmatrix}]$ is (I, S)-inner, i.e., if $\mathcal{N}^*\mathcal{N} + \mathcal{M}^*\mathcal{M} = S$, then $(\mathcal{N}, \mathcal{M})$ is an S-normalized right coprime factorization of $\mathcal{D}$.*

LEMMA 4.3. *Let $J = J^* \in \mathcal{L}(Y)$, and let $S \in \mathcal{L}(U)$, $S \gg 0$, $\widetilde{S} \in \mathcal{L}(U)$, and $\widetilde{S} \gg 0$. Let $\mathcal{D} \in TIC_\alpha(U;Y)$ for some $\alpha \geq 0$, and suppose that $\mathcal{D}$ has a right coprime factorization $(\mathcal{N}, \mathcal{M})$ in $TIC(U;Y)$.*

(i) *If $\mathcal{D}$ is stable, then $(\mathcal{N}, \mathcal{M})$ is a (J, S)-inner right coprime factorization of $\mathcal{D}$ iff $\mathcal{N}\mathcal{M}^{-1}$ is a (J, S)-inner-outer factorization of $\mathcal{D}$, or equivalently, iff $\mathcal{M}^{-1}$ is an S-spectral factor of $\mathcal{D}^*J\mathcal{D}$.*

(ii) *$\mathcal{D}$ has a (J, S)-inner right coprime factorization iff $\mathcal{D}$ is J-coercive.*

(iii) *The set of all possible (J, S)-inner right coprime factorizations $(\mathcal{N}, \mathcal{M})$ of $\mathcal{D}$ (where J and $\mathcal{D}$ are fixed while $\mathcal{N}$, $\mathcal{M}$ and S vary) can be parameterized as $\mathcal{N} = \widetilde{\mathcal{N}}E$, $\mathcal{M} = \widetilde{\mathcal{M}}E$, and $S = E^*\widetilde{S}E$, where $(\widetilde{\mathcal{N}}, \widetilde{\mathcal{M}})$ is a fixed $(J, \widetilde{S})$-inner right coprime factorization of $\mathcal{D}$ and $E \in \mathcal{L}(U)$ is an arbitrary invertible operator.*

Proof. (i) It is easy to see that if $\mathcal{X}$ is an S-spectral factor of $\mathcal{D}^*J\mathcal{D}$, and if we define $\mathcal{M} = \mathcal{X}^{-1}$ and $\mathcal{N} = \mathcal{D}\mathcal{X}$, then $(\mathcal{N}, \mathcal{M})$ is a (J, S)-inner right coprime factorization of $\mathcal{D}$ (it is coprime since $\mathcal{M}$ is invertible in $TIC(U)$). Conversely, if $(\mathcal{N}, \mathcal{M})$ is a (J, S)-inner right coprime factorization of $\mathcal{D}$, then $(\mathcal{D}, I)$ is another right coprime factorization of $\mathcal{D}$, and it follows from [24, Lemma 4.3(i)] that $\mathcal{M}$ has an inverse in $TIC(U)$. It is then obvious that $\mathcal{X} = \mathcal{M}^{-1}$ is an S-spectral factor of $\mathcal{D}^*J\mathcal{D}$.

(ii) If $\mathcal{D}$ is J-coercive, then by Lemmas 3.8 and 2.4(i), $\mathcal{N}$ is J-coercive and has a (J, S)-inner-outer factorization $\mathcal{N} = \widetilde{\mathcal{N}}\mathcal{X}$. According to Lemma 2.4(ii), $\mathcal{X}$ is invertible, and by [24, Lemma 4.3(i)], $(\widetilde{\mathcal{N}}, \widetilde{\mathcal{M}}) = (\mathcal{N}\mathcal{X}^{-1}, \mathcal{M}\mathcal{X}^{-1})$ is a (J, S)-inner right coprime factorization of $\mathcal{D}$.

On the other hand, if $\mathcal{D}$ has a (J, S)-inner right coprime factorization $(\mathcal{N}, \mathcal{M})$, then $\mathcal{N}$ is (J, S)-inner, hence J-coercive (since we assume that $S \gg 0$). By Lemma 3.8, $\mathcal{D}$ is J-coercive.

(iii) This follows from [24, Lemma 4.3(i)] and Lemma 2.4(iii). □

The following is our first main result.

THEOREM 4.4. *Let $J = J^* \in \mathcal{L}(Y)$, let $S \in \mathcal{L}(U)$, $S \gg 0$, and let $\Psi = \left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$ be a J-coercive jointly stabilizable and detectable well-posed linear system on (U, H, Y) [24, Definition 3.16]. Let $x^{\mathrm{opt}}(x_0)$, $y^{\mathrm{opt}}(x_0)$, and $u^{\mathrm{opt}}(x_0)$ be the optimal state, output, and control for the quadratic cost minimization problem for Ψ, and let Π be the corresponding Riccati operator (cf. Lemma 4.1).*

(i) *Let $(\mathcal{N}, \mathcal{M})$ be a (J, S)-inner right coprime factorization of $\mathcal{D}$. Then there is a unique feedback map $\mathcal{K}$ such that $[\mathcal{K} \quad \mathcal{F}] = [\mathcal{K} \quad (I - \mathcal{M}^{-1})]$ is an admissible stabilizing state feedback pair for Ψ and*

$$\begin{bmatrix} x^{\mathrm{opt}}(t, x_0) \\ y^{\mathrm{opt}}(x_0) \\ u^{\mathrm{opt}}(x_0) \end{bmatrix} = \begin{bmatrix} \mathcal{A}_{\circlearrowleft}(t) \\ \mathcal{C}_{\circlearrowleft} \\ \mathcal{K}_{\circlearrowleft} \end{bmatrix} x_0 = \begin{bmatrix} \mathcal{A}(t) + \mathcal{B}\mathcal{M}\tau(t)\mathcal{K} \\ \mathcal{C} + \mathcal{N}\mathcal{K} \\ \mathcal{M}\mathcal{K} \end{bmatrix} x_0$$

is equal to the state and output of the closed loop system $\Psi_{\circlearrowleft}$ defined by

$$\Psi_{\circlearrowleft} = \begin{bmatrix} \mathcal{A}_{\circlearrowleft} & \mathcal{B}_{\circlearrowleft} \\ \begin{bmatrix} \mathcal{C}_{\circlearrowleft} \\ \mathcal{K}_{\circlearrowleft} \end{bmatrix} & \begin{bmatrix} \mathcal{D}_{\circlearrowleft} \\ \mathcal{F}_{\circlearrowleft} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \mathcal{A} + \mathcal{B}\tau\mathcal{M}\mathcal{K} & \mathcal{B}\mathcal{M} \\ \begin{bmatrix} \mathcal{C} + \mathcal{N}\mathcal{K} \\ \mathcal{M}\mathcal{K} \end{bmatrix} & \begin{bmatrix} \mathcal{N} \\ \mathcal{M} - I \end{bmatrix} \end{bmatrix}$$

with initial value x_0, initial time zero, and zero control $u_{\circlearrowleft}$ (see Figure 2.1). The feedback map $\mathcal{K}$ is uniquely determined by the fact that $\mathcal{C}_{\circlearrowleft} = \mathcal{C} + \mathcal{N}\mathcal{K} \in \mathcal{L}(H; L^2(\mathbf{R}^+; Y))$, $\mathcal{K}_{\circlearrowleft} = \mathcal{M}\mathcal{K} \in \mathcal{L}(H; L^2(\mathbf{R}^+; U))$, and $\pi_+\mathcal{N}^*J\mathcal{C}_{\circlearrowleft} = 0$. Moreover, the Riccati operator of Ψ is given by

$$\Pi = \mathcal{C}_{\circlearrowleft}^* J \mathcal{C}_{\circlearrowleft} = (\mathcal{C} + \mathcal{N}\mathcal{K})^* J (\mathcal{C} + \mathcal{N}\mathcal{K}).$$

(ii) *If* $y = \mathcal{C}_{\circlearrowright}x_0 + \mathcal{D}_{\circlearrowright}\pi_+u_{\circlearrowright}$ *is the first output of the optimal closed loop system* $\Psi_{\circlearrowright}$ *in* (i) *with initial state* $x_0 \in H$ *and control* $u_{\circlearrowright} \in L^2(\mathbf{R}^+;U)$ (*see Figure* 2.1), *then the closed loop cost* $Q_{\circlearrowright}(x_0,u_{\circlearrowright})$ *is given by*

$$Q_{\circlearrowright}(x_0,u_{\circlearrowright}) = \langle y, Jy\rangle_{L^2(\mathbf{R}^+;Y)} = \langle x_0, \Pi x_0\rangle_H + \langle u_{\circlearrowright}, Su_{\circlearrowright}\rangle_{L^2(\mathbf{R}^+;Y)}\,. \tag{4.1}$$

(iii) *If* Ψ *is jointly* ω*-stabilizable and detectable for some* $\omega < 0$ [24, *Definition* 3.16], *and if* $\mathcal{N}$ *and* $\mathcal{M}$ *in* (i) *are right* ω*-coprime* [24, *Definition* 4.1], *then the closed loop system* $\Psi_{\circlearrowright}$ *is* ω*-stable.*

(iv) *If* $(\mathcal{N},\mathcal{M})$ *are given, then the feedback map* $\mathcal{K}$, *the Riccati operator* Π, *the closed loop semigroup* $\mathcal{A}_{\circlearrowright}$, *and the closed loop controllability and feedback maps* $\mathcal{C}_{\circlearrowright}$ *and* $\mathcal{K}_{\circlearrowright}$ *can be computed as follows*: *Choose some arbitrary jointly stabilizing feedback and output injection pairs* $[\mathcal{K}^1 \quad \mathcal{F}^1]$ *and* $\left[\begin{smallmatrix}\mathcal{H}\\ \mathcal{G}\end{smallmatrix}\right]$. *Then*

$$\begin{aligned}
\mathcal{K} &= \mathcal{M}^{-1}\mathcal{K}^1_\flat - S^{-1}\pi_+\mathcal{N}^*J\mathcal{C}_\flat,\\
\begin{bmatrix}\mathcal{A}_{\circlearrowright}\\ \mathcal{C}_{\circlearrowright}\\ \mathcal{K}_{\circlearrowright}\end{bmatrix} &= \begin{bmatrix}\mathcal{A}_\flat\\ \mathcal{C}_\flat\\ \mathcal{K}^1_\flat\end{bmatrix} - \begin{bmatrix}\mathcal{B}\mathcal{M}\tau\\ \mathcal{N}\\ \mathcal{M}\end{bmatrix} S^{-1}\pi_+\mathcal{N}^*J\mathcal{C}_\flat,\\
\Pi &= \mathcal{C}^*_\flat J\mathcal{C}_\flat - \left(\mathcal{K} - \mathcal{M}^{-1}\mathcal{K}^1_\flat\right)^* S\left(\mathcal{K} - \mathcal{M}^{-1}\mathcal{K}^1_\flat\right)\\
&= \mathcal{C}^*_\flat\left(J - J\mathcal{N}S^{-1}\pi_+\mathcal{N}^*J\right)\mathcal{C}_\flat = \mathcal{C}^*_\flat J\mathcal{C}_{\circlearrowright} = \mathcal{C}^*_{\circlearrowright}J\mathcal{C}_\flat,
\end{aligned}$$

where $\mathcal{A}_\flat = \mathcal{A} + \mathcal{B}\tau\mathcal{K}^1_\flat$, $\mathcal{C}_\flat = \mathcal{C} + \mathcal{D}\mathcal{K}^1_\flat$, *and* $\mathcal{K}^1_\flat = (I - \mathcal{F}^1)^{-1}\mathcal{K}^1$. (*If* Ψ *is stable, then we can can take* $\mathcal{K}^1_\flat = 0$, $\mathcal{A}_\flat = \mathcal{A}$, *and* $\mathcal{C}_\flat = \mathcal{C}$ *and get the same formulae as in Theorem* 2.6).

Proof. Let us first show that the conditions on $\mathcal{K}$ in (i) determine $\mathcal{K}$ uniquely. Suppose that we have two feedback maps $\mathcal{K}^1$ and $\mathcal{K}^2$ such that both $\mathcal{C}+\mathcal{N}\mathcal{K}^1$ and $\mathcal{C}+\mathcal{N}\mathcal{K}^2$ belong to $\mathcal{L}(H;L^2(\mathbf{R}^+;Y))$, both $\mathcal{M}\mathcal{K}^1$ and $\mathcal{M}\mathcal{K}^2$ belong to $\mathcal{L}(H;L^2(\mathbf{R}^+;U))$, and $\pi_+\mathcal{N}^*J(\mathcal{C}+\mathcal{N}\mathcal{K}^1) = \pi_+\mathcal{N}^*J(\mathcal{C}+\mathcal{N}\mathcal{K}^2)$. Then, for each $x \in H$, $\mathcal{N}(\mathcal{K}^1 - \mathcal{K}^2)x \in L^2(\mathbf{R}^+;Y)$, $\mathcal{M}(\mathcal{K}^1 - \mathcal{K}^2)x \in L^2(\mathbf{R}^+;U)$, and $\pi_+\mathcal{N}^*J(\mathcal{N}(\mathcal{K}^1 - \mathcal{K}^2)x) = 0$. By Lemma 3.3, $(\mathcal{K}^1 - \mathcal{K}^2)x \in L^2(\mathbf{R}^+;U)$, hence

$$0 = \pi_+\mathcal{N}^*J(\mathcal{N}(\mathcal{K}^1 - \mathcal{K}^2)x) = \pi_+(\mathcal{N}^*J\mathcal{N})(\mathcal{K}^1 - \mathcal{K}^2)x = S\pi_+(\mathcal{K}^1 - \mathcal{K}^2)x.$$

As $(\mathcal{K}^1 - \mathcal{K}^2)x$ is supported on $\mathbf{R}^+$ and S invertible, we must have $(\mathcal{K}^1 - \mathcal{K}^2)x = 0$ for all $x \in H$.

In order to prove the remainder of (i) we proceed as suggested by (iv); i.e., we choose preliminary jointly stabilizing feedback and output injection pairs $[\mathcal{K}^1 \quad \mathcal{F}^1]$ and $\left[\begin{smallmatrix}\mathcal{H}\\ \mathcal{G}\end{smallmatrix}\right]$ with interaction operator $\mathcal{E}_1$. The output injection pair and the interaction operator $\mathcal{E}_1$ play a very nonsignificant role below; they are only needed so that we can apply [24, Theorem 4.4] in order to show that $(\mathcal{D}(I-\mathcal{F}^1)^{-1}, (I-\mathcal{F}^1)^{-1})$ is a right coprime factorization of $\mathcal{D}$. We shall therefore ignore the output injection part of the system for the rest of this proof, but we return to this question at the end of the section.

We add the state feedback pair $[\mathcal{K}^1 \quad \mathcal{F}^1]$ to the system Ψ and close the state feedback loop as in Lemma 3.9 to get the stable system $\Psi_\flat$ given in that lemma. According to Lemma 4.1, the quadratic cost minimization problems for Ψ and $\Psi_\flat$ have the same optimal state $x^{\mathrm{opt}}(x_0)$ and output $y^{\mathrm{opt}}(x_0)$ and the optimal controls $u^{\mathrm{opt}}(x_0)$ and $u^{\mathrm{opt}}_\flat(x_0)$ are related to each other as in (3.2).

We want to apply Theorem 2.6 to solve the quadratic cost minimization problem for the closed loop system $\Psi_\flat$. By Lemmas 3.7 and 3.9, $\mathcal{D}_\flat$ is coercive. Since both

$(\mathcal{D}_\flat, (I + \mathcal{F}^1_\flat))$ and $(\mathcal{N}, \mathcal{M})$ are right coprime factorizations of $\mathcal{D}$, it follows from [24, Lemma 4.3] that the operator

$$\mathcal{X} = \mathcal{M}^{-1}\left(I + \mathcal{F}^1_\flat\right) = \left(\left(I - \mathcal{F}^1\right)\mathcal{M}\right)^{-1} \tag{4.2}$$

belongs to $TIC(U)$ and is invertible in $TIC(U)$. Thus, $\mathcal{N}\mathcal{X}$ is a (J, S)-inner-outer factorization of $\mathcal{D}_\flat$. By Theorem 2.6, the solution to the quadratic cost minimization problem for $\Psi_\flat$ is of state feedback type. More precisely, the pair

$$\begin{bmatrix} \mathcal{K}_\natural & \mathcal{F}_\natural \end{bmatrix} = \begin{bmatrix} -S^{-1}\pi_+\mathcal{N}^* J \mathcal{C}_\flat & (I - \mathcal{X}) \end{bmatrix}$$

is a stable stabilizing state feedback pair for $\Psi_\flat$, and if we further extended the system $\Psi_\flat$ into

$$\begin{bmatrix} \mathcal{A}_\flat & \mathcal{B}_\flat \\ \begin{bmatrix} \mathcal{C}_\flat \\ \mathcal{K}^1_\flat \\ \mathcal{K}_\natural \end{bmatrix} & \begin{bmatrix} \mathcal{D}_\flat \\ \mathcal{F}^1_\flat \\ \mathcal{F}_\natural \end{bmatrix} \end{bmatrix}$$

by adding the extra state feedback pair, and then close the new state feedback loop to get the stable closed loop system [24, Lemma 4.5]

$$\begin{aligned}
\Psi_{\flat\circlearrowleft} &= \begin{bmatrix} \mathcal{A}_{\flat\circlearrowleft} & \mathcal{B}_{\flat\circlearrowleft} \\ \begin{bmatrix} \mathcal{C}_{\flat\circlearrowleft} \\ \mathcal{K}^1_{\flat\circlearrowleft} \\ \mathcal{K}_{\natural\circlearrowleft} \end{bmatrix} & \begin{bmatrix} \mathcal{D}_{\flat\circlearrowleft} \\ \mathcal{F}^1_{\flat\circlearrowleft} \\ \mathcal{F}_{\natural\circlearrowleft} \end{bmatrix} \end{bmatrix} \\
&= \begin{bmatrix} \mathcal{A}_\flat + \mathcal{B}_\flat \tau \mathcal{X}^{-1}\mathcal{K}_\natural & \mathcal{B}_\flat \mathcal{X}^{-1} \\ \begin{bmatrix} \mathcal{C}_\flat + \mathcal{D}_\flat \mathcal{X}^{-1}\mathcal{K}_\natural \\ \mathcal{K}^1_\flat + \mathcal{F}^1_\flat \mathcal{X}^{-1}\mathcal{K}_\natural \\ \mathcal{X}^{-1}\mathcal{K}_\natural \end{bmatrix} & \begin{bmatrix} \mathcal{D}_\flat \mathcal{X}^{-1} \\ \mathcal{F}^1_\flat \mathcal{X}^{-1} \\ \mathcal{X}^{-1} - I \end{bmatrix} \end{bmatrix} \\
&= \begin{bmatrix} \mathcal{A} + \mathcal{B}\tau\left(\mathcal{K}^1_\flat + \mathcal{M}\mathcal{K}_\natural\right) & \mathcal{B}\mathcal{M} \\ \begin{bmatrix} \mathcal{C} + \mathcal{D}\left(\mathcal{K}^1_\flat + \mathcal{M}\mathcal{K}_\natural\right) \\ \mathcal{K}^1_\flat + \mathcal{F}^1\mathcal{M}\mathcal{K}_\natural \\ \left(I - \mathcal{F}^1\right)\mathcal{M}\mathcal{K}_\natural \end{bmatrix} & \begin{bmatrix} \mathcal{N} \\ \mathcal{F}^1\mathcal{M} \\ \left(I - \mathcal{F}^1\right)\mathcal{M} - I \end{bmatrix} \end{bmatrix},
\end{aligned}$$

then $x^{\mathrm{opt}}(x_0)$, $y^{\mathrm{opt}}(x_0)$, and $u_\flat^{\mathrm{opt}}(x_0)$ are given by

$$\begin{bmatrix} x^{\mathrm{opt}}(x_0) \\ y^{\mathrm{opt}}(x_0) \\ u_\flat^{\mathrm{opt}}(x_0) \end{bmatrix} = \begin{bmatrix} \mathcal{A}_{\flat\circlearrowleft} \\ \mathcal{C}_{\flat\circlearrowleft} \\ \mathcal{K}_{\natural\circlearrowleft} \end{bmatrix} x_0 = \begin{bmatrix} \mathcal{A} + \mathcal{B}\tau\left(\mathcal{K}^1_\flat + \mathcal{M}\mathcal{K}_\natural\right) \\ \mathcal{C} + \mathcal{D}\left(\mathcal{K}^1_\flat + \mathcal{M}\mathcal{K}_\natural\right) \\ \left(I - \mathcal{F}^1\right)\mathcal{M}\mathcal{K}_\natural \end{bmatrix} x_0$$

and $\mathcal{C}_{\flat\circlearrowleft}$ satisfies

$$\mathcal{N}^* J \mathcal{C}_{\flat\circlearrowleft} = 0.$$

From this result we are able to derive the conclusions listed in (i) and (iv). Most of the proof is ready. In particular, the formulae for the Riccati operator Π given in (iv) follow from Lemma 3.9 and the corresponding formulae in Theorem 2.6. It only remains to return to the original system Ψ and the original control $u^{\mathrm{opt}}(x_0)$.

The optimal control $u_\flat^{\mathrm{opt}}(x_0)$ for $\Psi_\flat$ corresponds to the optimal control

$$u^{\mathrm{opt}}(x_0) = \left(I - \mathcal{F}^1\right)^{-1}\left(\mathcal{K}^1 x_0 + u_\flat^{\mathrm{opt}}(x_0)\right) = \left(\mathcal{K}^1_\flat + \mathcal{M}\mathcal{K}_\natural\right) x_0$$

for the original system Ψ. We observe that $u^{\mathrm{opt}}(x_0)$ is equal to the sum of the two last outputs of $\Psi_{\flat\circlearrowright}$ with zero control. Let us add these two rows and combine them into one to get the system

$$\Psi_{\circlearrowright} = \begin{bmatrix} \mathcal{A}_{\circlearrowright} & \mathcal{B}_{\circlearrowright} \\ \begin{bmatrix} \mathcal{C}_{\circlearrowright} \\ \mathcal{K}_{\circlearrowright} \end{bmatrix} & \begin{bmatrix} \mathcal{D}_{\circlearrowright} \\ \mathcal{F}_{\circlearrowright} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \mathcal{A} + \mathcal{B}\tau\mathcal{M}\mathcal{K} & \mathcal{B}\mathcal{M} \\ \begin{bmatrix} \mathcal{C} + \mathcal{N}\mathcal{K} \\ \mathcal{M}\mathcal{K} \end{bmatrix} & \begin{bmatrix} \mathcal{N} \\ \mathcal{M} - I \end{bmatrix} \end{bmatrix},$$

where $\mathcal{K} = \mathcal{M}^{-1}\mathcal{K}_{\circlearrowright} = \mathcal{K}_{\natural} + \mathcal{M}^{-1}\mathcal{K}^1_{\flat}$. We then have

$$\begin{bmatrix} x^{\mathrm{opt}}(t, x_0) \\ y^{\mathrm{opt}}(x_0) \\ u^{\mathrm{opt}}(x_0) \end{bmatrix} = \begin{bmatrix} \mathcal{A}_{\circlearrowright}(t) \\ \mathcal{C}_{\circlearrowright} \\ \mathcal{K}_{\circlearrowright} \end{bmatrix} x_0 = \begin{bmatrix} \mathcal{A}(t) + \mathcal{B}\mathcal{M}\tau(t)\mathcal{K} \\ \mathcal{C} + \mathcal{N}\mathcal{K} \\ \mathcal{M}\mathcal{K} \end{bmatrix} x_0.$$

Moreover, since $\mathcal{C}_{\circlearrowright} = \mathcal{C}_{\flat\circlearrowright}$, we have $\mathcal{N}^* J \mathcal{C}_{\circlearrowright} = 0$, and $\Psi_{\circlearrowright}$ is the system that we get by closing the state feedback loop in the system

$$\begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \begin{bmatrix} \mathcal{C} \\ \mathcal{K} \end{bmatrix} & \begin{bmatrix} \mathcal{D} \\ \mathcal{F} \end{bmatrix} \end{bmatrix},$$

where $\mathcal{K}$ is the feedback map defined above and $\mathcal{F} = I - \mathcal{M}^{-1}$. This completes the proofs of both (i) and (iv).

The proof of (ii) is identical to the proof of Theorem 2.6(iii).

Finally, let us prove (iii). Under the assumption of (iii) we can throughout work with the notion of ω-stability instead of just plain stability (the latter notion is the same as ω-stability with $\omega = 0$). The only part of the extended optimal system $\Psi_{\flat\circlearrowright}$ whose ω-stability is not obvious is the state feedback map $\mathcal{K}_{\natural}$; all the other parts of the system are bounded linear operators on the correct spaces. Thus, we must show that $\mathcal{K}_{\natural} \in \mathcal{L}(H; L^2_{\omega}(\mathbf{R}^+; U))$. Recalling the definition of $\mathcal{K}_{\natural}$, we realize that it suffices to show that the anticausal operator $\mathcal{N}^*$ belongs to $TI_{\omega}(Y; U)$. Here the duality is with respect to the inner product in the unweighted L^2, so by standard duality theory $\mathcal{N}^* \in TI_{-\omega}(Y; U)$. However, since $\mathcal{N}^*$ is anticausal, this implies that $\mathcal{N}^*$ can be extended to an anticausal operator in $TI_{\beta}(Y; U)$ for all $\beta \leq -\omega$ [24, Lemmas 2.4 and 2.9]. In particular, since $\omega \leq 0$, $\mathcal{N}^* \in TI_{\omega}(Y; U)$. Thus, $\mathcal{K}_{\natural} \in \mathcal{L}(H; L^2_{\omega}(\mathbf{R}^+; U))$, and $\Psi_{\flat\circlearrowright}$ is stable. □

REMARK 4.5. *An inspection of the proof of Theorem* 2.6 *shows that if the system* Ψ *is jointly strongly stabilizable and detectable, then the optimal closed loop system* $\Psi_{\circlearrowright}$ *will be strongly stable, too* [24, *Lemma* 3.5].

Theorem 4.4 does not contain a converse part like the one found in Theorem 2.6(ii), since we have been able to prove only the following partial converse.

THEOREM 4.6. *Make the same hypothesis as in Theorem* 4.4. *Suppose that the solution to the quadratic cost minimization problem is of state feedback type in the sense that* $\left[\begin{smallmatrix} y^{\mathrm{opt}}(x_0) \\ u^{\mathrm{opt}}(x_0) \end{smallmatrix}\right]$ *is equal to the output of the closed loop system* $\Psi_{\circlearrowright}$ *with initial value* x_0, *initial time zero, zero input* $u_{\circlearrowright}$, *and some stabilizing state feedback pair* $[\mathcal{K} \quad \mathcal{F}]$. *Define* $\mathcal{M} = (I - \mathcal{F})^{-1}$ *and* $\mathcal{N} = \mathcal{D}\mathcal{M}$. *Then there exists a positive invertible operator* $S = S^* \in \mathcal{L}(U)$ *such that* $\mathcal{N}$ *is* (J, S)*-inner, and the claim* (ii) *in Theorem* 4.4 *is true for this closed loop system. If, moreover,* $\mathcal{N}$ *and* $\mathcal{M}$ *are right coprime, then* $(\mathcal{N}, \mathcal{M})$ *is a* (J, S)*-inner right coprime factorization of* $\mathcal{D}$. *This is true, in particular, whenever* Ψ *is exponentially stabilizable.*

Proof. We suppose that the solution to the quadratic cost minimization problem for Ψ is of state feedback type and claim that this implies that the solution to the quadratic cost minimization problem for the system $\Psi_\flat$ considered in the proof of Theorem 4.4 is also of state feedback type. The proof of this is based on [24, Lemma 4.5] and Lemma 3.9. We consider the combined system

$$\begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \begin{bmatrix} \mathcal{C} \\ \mathcal{K}^1 \\ \mathcal{K} \end{bmatrix} & \begin{bmatrix} \mathcal{D} \\ \mathcal{F}^1 \\ \mathcal{F} \end{bmatrix} \end{bmatrix},$$

where $(\mathcal{K}^1, \mathcal{F}^1)$ is the same preliminary feedback pair that we used in the proof of Theorem 4.4 and $[\mathcal{K} \quad \mathcal{F}]$ is the optimal feedback pair. By [24, Lemma 4.5], $[\mathcal{K} \quad \mathcal{F}]$ is a stabilizing feedback pair for this combined system (due to the coprimeness of $\mathcal{D}(I - \mathcal{F}^1)^{-1}$ and $(I - \mathcal{F}^1)^{-1}$). Moreover, the pair

$$\begin{bmatrix} \mathcal{K}_\natural & \mathcal{F}_\natural \end{bmatrix} = \begin{bmatrix} \mathcal{K} - (I - \mathcal{F})(I - \mathcal{F}^1)^{-1}\mathcal{K}^1 & I - (I - \mathcal{F})(I - \mathcal{F}^1)^{-1} \end{bmatrix}$$

is a stabilizing feedback pair for $\Psi_\flat$. By combining this fact with Lemma 3.9, we find that the optimal solution to the quadratic cost minimization problem for the system $\Psi_\flat$ is of state feedback type. However, in contrast to the situation covered by the converse part of Theorem 2.6, we do not know that the feedback pair $[\mathcal{K}_\natural \quad \mathcal{F}_\natural]$ itself is stable, and this causes some additional complications and prevents us from applying part (ii) of Theorem 2.6. Instead we argue directly, examining the proof of the converse part of Theorem 2.6 as presented in [23].

We know that $\mathcal{F}_\natural \in TIC_\alpha(U)$ for some $\alpha \geq 0$ (but not necessarily for $\alpha = 0$) and that $(I - \mathcal{F}_\natural)^{-1} \in TIC(U)$. Fortunately, it was the latter property that was important for a major part of the proof of Theorem 2.6(ii). By repeating the argument in [23] we find that if we define $\mathcal{M}_\natural = (I - \mathcal{F}_\natural)^{-1}$, then

$$\mathcal{M}_\natural^* \mathcal{D}_\flat^* J \mathcal{D}_\flat \mathcal{M}_\natural = S$$

for some nonnegative $S = S^* \in \mathcal{L}(U)$. However, the proof given there of the invertibility of S was based on the boundedness of $\mathcal{F}_\natural$, so it does not apply.

Since $\mathcal{M}_\natural$ is invertible in $TIC_\alpha(U)$, we know that $\mathcal{M}_\natural$ is one-to-one. This together with the invertibility of $\mathcal{D}_\flat^* J \mathcal{D}_\flat$ (which is a consequence of the J-coercivity) implies that S is one-to-one. Its inverse S^{-1} is a nonnegative, possibly unbounded, self-adjoint operator which has a nonnegative self-adjoint square root $S^{-1/2}$. Denote the domain of $S^{-1/2}$ by W. Then $\mathcal{M}_\natural S^{-1/2} \in TIC(W; U)$, and it can be extended to an operator on $TIC(U)$ since $S^{-1/2}\mathcal{M}_\natural^* \mathcal{D}_\flat^* J \mathcal{D}_\flat \mathcal{M}_\natural S^{1/2}$ can be extended to the identity operator on $TIC(U)$. We denote this extension of $\mathcal{M}_\natural S^{-1/2}$ by $\widetilde{\mathcal{M}}$. Then $S^{-1/2} = \mathcal{M}_\natural^{-1}\widetilde{\mathcal{M}} \in TIC_\alpha(W; U)$, and it can be extended to an operator in $TIC_\alpha(U)$ since the right-hand side of this equation belongs to $TIC_\alpha(U)$. But this means that $S^{-1/2}$ can be extended to an operator in $\mathcal{L}(U)$, hence $S^{1/2}$ and S must be invertible.

Since $\mathcal{N} = \mathcal{D}\mathcal{M} = \mathcal{D}(I - \mathcal{F})^{-1} = \mathcal{D}_\flat \mathcal{M}_\natural^{-1}$, and since $\mathcal{M}_\natural^* \mathcal{D}_\flat^* J \mathcal{D}_\flat \mathcal{M}_\natural = S$, we find that $\mathcal{N}$ is (J, S)-inner as claimed. The proof of the statement given in part (ii) of Theorem 4.4 remains the same as before.

We have not been able to prove that $\mathcal{N}$ and $\mathcal{M}$ must always be right coprime. This will be true if and only if the feedback pair $[\mathcal{K}^1 \quad \mathcal{F}^1]$ stabilizes the original system extended with the feedback pair $[\mathcal{K} \quad \mathcal{F}]$, cf. [24, Lemma 4.5]. In particular, it is true whenever $[\mathcal{K}^1 \quad \mathcal{F}^1]$ is exponentially stabilizing; see [24, Lemma 3.20]. □

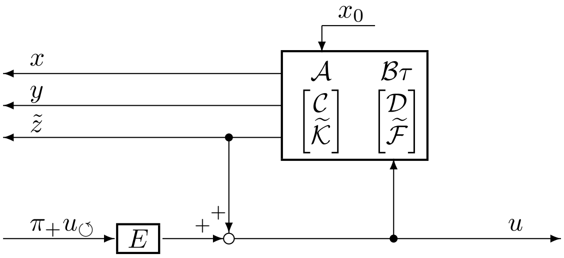

x_0
x
y
$\tilde z$
$\mathcal A$
$\mathcal B_\tau$
$\begin{bmatrix}\mathcal C\\ \widetilde{\mathcal K}\end{bmatrix}$
$\begin{bmatrix}\mathcal D\\ \widetilde{\mathcal F}\end{bmatrix}$
$\pi_+u_\circlearrowleft$
E
$+$
$+$
u

FIG. 4.1. *Externally parameterized optimal state feedback system.*

According to Lemmas 2.4 and 4.3, Theorems 2.6 and 4.4 contain a hidden free invertible parameter $E \in \mathcal L(U)$.[8] For example, the set of all possible (J,S)-inner coprime factorization of $\mathcal D$ in Theorem 4.4 can be parameterized as follows.

PROPOSITION 4.7. *Let $(\widetilde{\mathcal N},\widetilde{\mathcal M})$ be a particular $(J,\widetilde S)$-inner coprime factorization of $\mathcal D$. Then the set of all possible sensitivity operators S and all possible (J,S)-inner coprime factorizations $(\mathcal N,\mathcal M)$ of $\mathcal D$ (where J and $\mathcal D$ are fixed while $\mathcal N$, $\mathcal M$, and S vary) can be parameterized as*

$$S = E^*\widetilde S E, \qquad \mathcal N = \widetilde{\mathcal N}E, \qquad \mathcal M = \widetilde{\mathcal M}E,$$

where E varies over the set of all invertible operators in $\mathcal L(U)$. The corresponding feedback pair $[\mathcal K \quad \mathcal F]$ in Theorem 4.4 is given by

$$\mathcal K = E^{-1}\widetilde{\mathcal K}, \qquad (I-\mathcal F) = E^{-1}(I-\widetilde{\mathcal F}),$$

where $\widetilde{\mathcal K} = -\widetilde S\pi_+\widetilde{\mathcal N}^*J\mathcal C$ and $\widetilde{\mathcal F} = (I-\widetilde{\mathcal M}^{-1})$; i.e., $[\widetilde{\mathcal K} \quad \widetilde{\mathcal F}]$ is the feedback pair in Theorem 4.4 corresponding to the factorization $(\widetilde{\mathcal N},\widetilde{\mathcal M})$. The parameterized version of the formula for the closed loop system in Theorem 4.4 is

$$\Psi_\circlearrowleft = \begin{bmatrix}\mathcal A_\circlearrowleft & \mathcal B_\circlearrowleft\\ \begin{bmatrix}\mathcal C_\circlearrowleft\\ \mathcal K_\circlearrowleft\end{bmatrix} & \begin{bmatrix}\mathcal D_\circlearrowleft\\ \mathcal F_\circlearrowleft\end{bmatrix}\end{bmatrix} = \begin{bmatrix}\mathcal A+\mathcal B\widetilde{\mathcal M}\tau\widetilde{\mathcal K} & \mathcal B\widetilde{\mathcal M}E\\ \begin{bmatrix}\mathcal C+\widetilde{\mathcal N}\widetilde{\mathcal K}\\ \widetilde{\mathcal M}\widetilde{\mathcal K}\end{bmatrix} & \begin{bmatrix}\widetilde{\mathcal N}E\\ \widetilde{\mathcal M}E-I\end{bmatrix}\end{bmatrix}.$$

The first column is independent of E (but the second is not).

This follows from Lemma 4.3.

The operator E has a very simple interpretation: it represents a coordinate change in the input space for the closed loop system.

PROPOSITION 4.8. *Introduce the same notation as in Proposition 4.7. Then the two diagrams drawn in Figures 2.1 and 4.1 are equivalent in the sense that the relationships between all the signals with identical names are identical in the two diagrams (but z differs in general from $\tilde z$.)*

Proof. Clearly, if we can show that the relationships between $u_\circlearrowleft$ and u are the same in both the diagrams, then all the other relationships will be the same, too, since both diagrams say that

$$\begin{aligned} x &= \mathcal A x_0 + \mathcal B\tau u,\\ y &= \mathcal C x_0 + \mathcal D u. \end{aligned}$$

[8] In [23] this parameter was written out explicitly and it was explained why it cannot be avoided: it represents an undetermined feed-forward term inside the feedback loop.

In Figure 2.1 we have

$$u = \mathcal{K}x_0 + \mathcal{F}u + \pi_+ u_\flat,$$

from which we can solve u in the form

$$\begin{aligned} u &= (I - \mathcal{F})^{-1} (\mathcal{K}x_0 + \pi_+ u_\flat) \\ &= (I - \widetilde{\mathcal{F}})^{-1} E \left(E^{-1}\widetilde{\mathcal{K}}x_0 + \pi_+ u_\flat\right) \\ &= (I - \widetilde{\mathcal{F}})^{-1} (\widetilde{\mathcal{K}}x_0 + E\pi_+ u_\flat). \end{aligned}$$

On the other hand, Figure 4.1 says that

$$u = \widetilde{\mathcal{K}}x_0 + \widetilde{\mathcal{F}}u + E\pi_+ u_\flat,$$

and this equation is equivalent to the one above. □

The minimization problem considered in Theorem 4.4 leads to an inner coprime factorization. If instead we use the different cost function

$$Q_1(x_0, u) = \|y\|^2_{L^2(\mathbf{R}^+;Y)} + \|u\|^2_{L^2(\mathbf{R}^+;U)}, \tag{4.3}$$

then we get a normalized coprime factorization.

COROLLARY 4.9. *Let* $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$ *be a jointly stabilizable and detectable well-posed linear system on* (U, H, Y). *Let* $x^{\mathrm{opt}}(x_0)$, $y^{\mathrm{opt}}(x_0)$, *and* $u^{\mathrm{opt}}(x_0)$ *be the optimal state, output, and control for the quadratic cost minimization problem described in Definition* 3.1, *but with the cost function* $Q(x_0, u)$ *replaced by the cost function* $Q_1(x_0, u)$ *in* (4.3). *If* $S = S^* \in \mathcal{L}(U)$ *and* $(\mathcal{N}, \mathcal{M})$ *is an* S*-normalized right coprime factorization of* $\mathcal{D}$ (*in the sense of Definition* 4.2), *then there is a unique feedback map* $\mathcal{K}$ *such that* $[\mathcal{K} \quad \mathcal{F}] = [\mathcal{K} \quad (I - \mathcal{M}^{-1})]$ *is an admissible stabilizing state feedback pair for* Ψ *and*

$$\begin{bmatrix} x^{\mathrm{opt}}(t, x_0) \\ y^{\mathrm{opt}}(x_0) \\ u^{\mathrm{opt}}(x_0) \end{bmatrix} = \begin{bmatrix} \mathcal{A}_\circlearrowleft(t) \\ \mathcal{C}_\circlearrowleft \\ \mathcal{K}_\circlearrowleft \end{bmatrix} x_0 = \begin{bmatrix} \mathcal{A}(t) + \mathcal{B}\mathcal{M}\tau(t)\mathcal{K} \\ \mathcal{C} + \mathcal{N}\mathcal{K} \\ \mathcal{M}\mathcal{K} \end{bmatrix} x_0$$

is equal to the state and output of the closed loop system $\Psi_\circlearrowleft$ *defined by*

$$\Psi_\circlearrowleft = \begin{bmatrix} \mathcal{A}_\circlearrowleft & \mathcal{B}_\circlearrowleft \\ \begin{bmatrix} \mathcal{C}_\circlearrowleft \\ \mathcal{K}_\circlearrowleft \end{bmatrix} & \begin{bmatrix} \mathcal{D}_\circlearrowleft \\ \mathcal{F}_\circlearrowleft \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \mathcal{A} + \mathcal{B}\tau\mathcal{M}\mathcal{K} & \mathcal{B}\mathcal{M} \\ \begin{bmatrix} \mathcal{C} + \mathcal{N}\mathcal{K} \\ \mathcal{M}\mathcal{K} \end{bmatrix} & \begin{bmatrix} \mathcal{N} \\ \mathcal{M} - I \end{bmatrix} \end{bmatrix}$$

with initial value x_0, *initial time zero*, *and zero input* $u_\circlearrowleft$ (*see Figure* 2.1). *The feedback map* $\mathcal{K}$ *is uniquely determined by the fact that* $\mathcal{C}_\circlearrowleft = \mathcal{C} + \mathcal{N}\mathcal{K} \in \mathcal{L}(H; L^2(\mathbf{R}^+; Y))$, $\mathcal{K}_\circlearrowleft = \mathcal{M}\mathcal{K} \in \mathcal{L}(H; L^2(\mathbf{R}^+; U))$, *and*

$$\pi_+ (\mathcal{N}^*\mathcal{C}_\circlearrowleft + \mathcal{M}^*\mathcal{K}_\circlearrowleft) = 0.$$

Moreover, *the Riccati operator of* Ψ *is given by*

$$\Pi = \mathcal{C}^*_\circlearrowleft\mathcal{C}_\circlearrowleft + \mathcal{K}^*_\circlearrowleft\mathcal{K}_\circlearrowleft = (\mathcal{C} + \mathcal{N}\mathcal{K})^*(\mathcal{C} + \mathcal{N}\mathcal{K}) + (\mathcal{M}\mathcal{K})^*(\mathcal{M}\mathcal{K}).$$

Proof. Apply Theorem 4.4 with Ψ replaced by the augmented system

$$\Psi_{\mathrm{aug}} = \begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \begin{bmatrix} \mathcal{C} \\ 0 \end{bmatrix} & \begin{bmatrix} \mathcal{D} \\ I \end{bmatrix} \end{bmatrix},$$

and use the fact that $(\mathcal{N}, \mathcal{M})$ is an S-normalized right coprime factorization of $\mathcal{D}$ iff $([\begin{smallmatrix}\mathcal{N}\\ \mathcal{M}\end{smallmatrix}], \mathcal{M})$ is an (I, S)-inner right coprime factorization of $[\begin{smallmatrix}\mathcal{D}\\ I\end{smallmatrix}]$. Also observe that the augmented system is always coercive. The net effect is that throughout one replaces J by I, $\mathcal{D}$ by $[\begin{smallmatrix}\mathcal{D}\\ I\end{smallmatrix}]$, $\mathcal{N}$ by $[\begin{smallmatrix}\mathcal{N}\\ \mathcal{M}\end{smallmatrix}]$, $\mathcal{C}$ by $[\begin{smallmatrix}\mathcal{C}\\ 0\end{smallmatrix}]$, $\mathcal{C}_\flat$ by $[\begin{smallmatrix}\mathcal{C}_\flat\\ \mathcal{K}^1_\flat\end{smallmatrix}]$, and $\mathcal{C}_\circlearrowleft$ by $[\begin{smallmatrix}\mathcal{C}_\circlearrowleft\\ \mathcal{K}_\circlearrowleft\end{smallmatrix}]$ in Theorem 4.4. $\square$

Let us end this section with a discussion of the joint stabilizability and detectability assumption in Theorem 4.4. This assumption was needed so that we could apply [24, Theorem 4.4] and conclude that the preliminary feedback gives us a coprime factorization of the input/output map. The optimal feedback given by Theorem 4.4 gives us another stabilizing feedback pair. However, we are not able to prove that the optimal feedback pair and the original stabilizing output injection pair $[\begin{smallmatrix}\mathcal{H}\\ \mathcal{G}\end{smallmatrix}]$ are jointly stabilizing (and we do not even expect this to be true in full generality). The problem is that these two pairs need not have a well-defined interaction operator $\mathcal{E}$.

By using the fact that the interaction operator $\mathcal{E}$ is determined (modulo a static part) by its Hankel operator $\mathcal{K}\mathcal{H}$, it is possible to construct $\mathcal{E}$ (whenever such an interaction operator exists). To do this we have to take a closer look at the proof of Theorem 4.4. The critical step in the proof is the addition of the state feedback row to the preliminary stabilized system $\Psi_\flat$. Let us redo this part of the proof, restoring the omitted output injection column

$$\begin{bmatrix} \mathcal{H} \\ \mathcal{G} \\ \mathcal{E}_1 \end{bmatrix}$$

to the system Ψ_1; i.e., let us start with the full system

$$\Psi_{\text{ext}} = \begin{bmatrix} \mathcal{A} & \begin{bmatrix} \mathcal{H} & \mathcal{B} \end{bmatrix} \\ \begin{bmatrix} \mathcal{C} \\ \mathcal{K}^1 \end{bmatrix} & \begin{bmatrix} \mathcal{G} & \mathcal{D} \\ \mathcal{E}_1 & \mathcal{F}^1 \end{bmatrix} \end{bmatrix}$$

and close the state feedback loop to get the stable system

$$\begin{aligned}\Psi_\flat &= \begin{bmatrix} \mathcal{A}_\flat & \begin{bmatrix} \mathcal{H}_\flat & \mathcal{B}_\flat \end{bmatrix} \\ \begin{bmatrix} \mathcal{C}_\flat \\ \mathcal{K}^1_\flat \end{bmatrix} & \begin{bmatrix} \mathcal{G}_\flat & \mathcal{D}_\flat \\ \mathcal{E}_{1\flat} & \mathcal{F}^1_\flat \end{bmatrix} \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{A} + \mathcal{B}\tau \left(I - \mathcal{F}^1\right)^{-1} \mathcal{K}^1 & \begin{bmatrix} \mathcal{H} + \mathcal{B}\left(I - \mathcal{F}^1\right)^{-1} \mathcal{E}_1 & \mathcal{B}\left(I - \mathcal{F}^1\right)^{-1} \end{bmatrix} \\ \begin{bmatrix} \mathcal{C} + \mathcal{D}\left(I - \mathcal{F}^1\right)^{-1} \mathcal{K}^1 \\ \left(I - \mathcal{F}^1\right)^{-1} \mathcal{K}^1 \end{bmatrix} & \begin{bmatrix} \mathcal{G} + \mathcal{D}\left(I - \mathcal{F}^1\right)^{-1} \mathcal{E}_1 & \mathcal{D}\left(I - \mathcal{F}^1\right)^{-1} \\ \left(I - \mathcal{F}^1\right)^{-1} \mathcal{E}_1 & \left(I - \mathcal{F}^1\right)^{-1} - I \end{bmatrix} \end{bmatrix}.\end{aligned}$$

To this system we want to add a new state feedback row $[\mathcal{K}_\natural \quad \mathcal{E}_\natural \quad \mathcal{F}_\natural]$. To see how this row should be constructed we examine the feedback pair

$$\begin{bmatrix} \mathcal{K}_\natural & \mathcal{F}_\natural \end{bmatrix} = \begin{bmatrix} -S^{-1}\pi_+\mathcal{N}^* J \mathcal{C}_\flat & (I - \mathcal{X}) \end{bmatrix}$$

that we used in the proof of Theorem 4.4. Since

$$\begin{aligned} \mathcal{X} &= \mathcal{M}^{-1}(I - \mathcal{F}^1)^{-1} = S^{-1}\mathcal{N}^* J \mathcal{N}\mathcal{M}^{-1}(I - \mathcal{F}^1)^{-1} \\ &= S^{-1}\mathcal{N}^* J \mathcal{D}(I - \mathcal{F}^1)^{-1} = S^{-1}\mathcal{N}^* J \mathcal{D}_\flat, \end{aligned}$$

this pair can be rewritten in the alternative form

$$\begin{bmatrix} \mathcal{K}_\natural & \mathcal{F}_\natural \end{bmatrix} = \begin{bmatrix} 0 & I \end{bmatrix} - S^{-1}\begin{bmatrix} \pi_+\mathcal{N}^* J \mathcal{C}_\flat & \mathcal{N}^* J \mathcal{D}_\flat \end{bmatrix},$$

which gives us a clue to the correct definition of $\mathcal{E}_\natural$.

LEMMA 4.10. *Let $\Psi = \left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$ be a stable well-posed linear system on (U, H, Y), and let $\mathcal{N}^* \in TI(Y; Z)$ be anticausal. Define*

$$\mathcal{K} = \pi_+\mathcal{N}^*\mathcal{C}, \qquad \mathcal{E} = \mathcal{N}^*\mathcal{D}.$$

(i) *The map $\mathcal{K}$ satisfies*

$$\mathcal{K}\mathcal{A}(t) = \pi_+\tau(t)\mathcal{K}, \qquad t \in \mathbf{R}^+,$$

and the Hankel operator of $\mathcal{E}$ is given by

$$\pi_+\mathcal{E}\pi_- = \mathcal{K}\mathcal{B}.$$

(ii) *If $\mathcal{E}$ can be written as a sum $\mathcal{E} = \mathcal{E}_- + \mathcal{E}_+$, where both $\mathcal{E}_-$ and $\mathcal{E}_+$ belong to $TI(Y; Z)$ and $\mathcal{E}_+$ is causal and $\mathcal{E}_-$ anticausal, then*

$$\begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \begin{bmatrix}\mathcal{C}\\ \mathcal{K}\end{bmatrix} & \begin{bmatrix}\mathcal{D}\\ \mathcal{E}_+\end{bmatrix}\end{bmatrix}$$

is a stable well-posed linear system on $(U, H, Y \times Z)$.

(iii) *Conversely, if there exists some $\mathcal{E}_+$ for which the system above is a stable well-posed linear system on $(U, H, Y \times Z)$, then we get a splitting $\mathcal{E} = \mathcal{E}_- + \mathcal{E}_+$ of the type described above by defining $\mathcal{E}_- = \mathcal{E} - \mathcal{E}_+$. This splitting is unique modulo a static operator E.*

Proof. (i) To compute $\mathcal{K}\mathcal{A}(t)$ we use the anticausality and time invariance of $\mathcal{N}^*$ and part (iii) of [24, Definition 2.1] to get

$$\begin{aligned}\mathcal{K}\mathcal{A}(t) &= \pi_+\mathcal{N}^*\mathcal{C}\mathcal{A}(t)\\ &= \pi_+\mathcal{N}^*\pi_+\tau(t)\mathcal{C}\\ &= \pi_+\mathcal{N}^*\tau(t)\mathcal{C}\\ &= \pi_+\tau(t)\mathcal{N}^*\mathcal{C}\\ &= \pi_+\tau(t)\pi_+\mathcal{N}^*\mathcal{C}\\ &= \pi_+\tau(t)\mathcal{K}.\end{aligned}$$

Almost the same argument, but with part (iii) of [24, Definition 2.1] replaced by part (iv), gives

$$\begin{aligned}\pi_+\mathcal{E}\pi_- &= \pi_+\mathcal{N}^*\mathcal{D}\pi_-\\ &= \pi_+\mathcal{N}^*\pi_+\mathcal{D}\pi_-\\ &= \pi_+\mathcal{N}^*\mathcal{C}\mathcal{B}\\ &= \mathcal{K}\mathcal{B}.\end{aligned}$$

(ii) This follows immediately from part (i) and [24, Definition 2.1] (the Hankel operator of $\mathcal{E}_-$ is zero).

(iii) Clearly $\mathcal{E}_-$ is anticausal, since $\mathcal{E}$ and $\mathcal{E}_+$ have the same Hankel operator. The uniqueness statement follows from [23, Lemma 6]. □

By applying this lemma to the crucial step in the proof of Theorem 4.4 (and using [24, Lemma 3.5]) we get the following addition to Theorem 4.4.

Corollary 4.11. *Let* $\left[\begin{smallmatrix}\mathcal{H}\\ \mathcal{G}\end{smallmatrix}\right]$ *and* $\mathcal{E}_1$ *be the output injection pair and the interaction operator used in the proof of Theorem* 4.4. *Then the optimal state feedback pair* $[\mathcal{K}\quad\mathcal{F}]$ *and the output injection pair* $\left[\begin{smallmatrix}\mathcal{H}\\ \mathcal{G}\end{smallmatrix}\right]$ *are jointly stabilizing iff* $\mathcal{N}^*J\mathcal{G}_\flat = \mathcal{N}^*J(\mathcal{G}+(I-\mathcal{F}^1)^{-1}\mathcal{E}_1)$ *can be split into a causal and an anticausal part that both belong to* $TI(Y;U)$.

We shall not need this result here and leave the proof to the reader.

For completeness, let us also mention the following "dual" result, where one uses an anticausal time-invariant operator to construct a new output injection pair for a well-posed linear system. This result is needed in the solution to the "dual" quadratic optimal filtering problem.

Lemma 4.12. *Let* $\Psi = \left[\begin{smallmatrix}\mathcal{A} & \mathcal{B}\\ \mathcal{C} & \mathcal{D}\end{smallmatrix}\right]$ *be a stable well-posed linear system on* (U,H,Y), *and let* $\widetilde{\mathcal{N}}^* \in TI(Z;U)$ *be anticausal. Define*

$$\mathcal{H} = \mathcal{B}\widetilde{\mathcal{N}}^*\pi_-, \qquad \mathcal{E} = \mathcal{D}\widetilde{\mathcal{N}}^*.$$

(i) *The map* $\mathcal{H}$ *satisfies*

$$\mathcal{A}(t)\mathcal{H} = \mathcal{H}\tau(t)\pi_-, \qquad t\in\mathbf{R}^+,$$

and the Hankel operator of $\mathcal{E}$ *is given by*

$$\pi_+\mathcal{E}\pi_- = \mathcal{C}\mathcal{H}.$$

(ii) *If* $\mathcal{E}$ *can be written as a sum* $\mathcal{E} = \mathcal{E}_- + \mathcal{E}_+$, *where both* $\mathcal{E}_-$ *and* $\mathcal{E}_+$ *belong to* $TI(Z;U)$ *and* $\mathcal{E}_+$ *is causal and* $\mathcal{E}_-$ *anticausal, then*

$$\begin{bmatrix}\mathcal{A} & [\mathcal{H} \quad \mathcal{B}]\\ \mathcal{C} & [\mathcal{E}_+ \quad \mathcal{D}]\end{bmatrix}$$

is a stable well-posed linear system on $(Z\times U,H,Y)$.

(iii) *Conversely, if there exists some* $\mathcal{E}_+$ *for which the system above is a stable well-posed linear system on* $(Z\times U,H,Y)$, *then we get a splitting* $\mathcal{E} = \mathcal{E}_- + \mathcal{E}_+$ *of the type described above by defining* $\mathcal{E}_- = \mathcal{E} - \mathcal{E}_+$. *This splitting is unique modulo a static operator* E.

The proof of this lemma is very similar to the proof of Lemma 4.10, and we leave it to the reader.

Remark 4.13. *Theorem* 4.4 *and Corollary* 4.9 *remain true in the case where we minimize* $\langle y, Jy\rangle_{L^2_\beta(\mathbf{R}^+;Y)}$ *instead, where* $\beta\in\mathbf{R}$ *is arbitrary. We must then throughout replace the unweighted space* L^2 *by the weighted space* L^2_β. *In particular, the notion of an inner operator should be redefined so that it refers to the weighted space* L^2_β, *and the adjoints should be computed with respect to the inner product in the weighted space* L^2_β.

Remark 4.14. *The results of this section remain valid if throughout we replace the algebra of time-invariant bounded linear operators from* $L^2(\mathbf{R};U)$ *into* $L^2(\mathbf{R};Y)$ *by various subalgebras, for example, the algebra of convolution operators induced by measures with finite total variation. The main exception is that spectral factorizations and inner-outer factorizations need not exist in all subalgebras. In particular, Theorem* 4.4 *remains valid. In the algebra studied in* [18] *spectral factorizations and inner-outer factorizations do exist and the input/output maps can always be decomposed into causal and anticausal parts, as required by Lemmas* 4.10 *and* 4.12.

5. The optimal problem on a finite time interval. Our next goal is to show that the Riccati operator in Theorem 4.4 satisfies an algebraic Riccati equation involving the generating operators of the system Ψ. Two such algebraic Riccati equations were given in [23], namely the *open loop* and the *closed loop* algebraic Riccati equations. The derivation of the closed loop Riccati equation was based entirely on the properties of the optimal closed loop system, and that argument remains valid since the optimal closed loop system is stable. However, the open loop system can be unstable, and in order to derive the open loop Riccati equation we have to study the behavior of the optimal system on a finite time interval.

LEMMA 5.1. *Make the same assumption and introduce the same notations as in Theorem* 4.4. *Then, for all $x_0 \in H$ and all $t \geq 0$,*

$$\pi_{[0,t]}\left(\mathcal{D}^* J\pi_{[0,t]}\mathcal{C}_{\circlearrowleft} + \tau(-t)\mathcal{B}^*\Pi\mathcal{A}_{\circlearrowleft}(t)\right) = 0, \tag{5.1}$$

$$\pi_{[0,t]}\left(\mathcal{N}^* J\pi_{[0,t]}\mathcal{C}_{\circlearrowleft} + \tau(-t)\mathcal{M}^*\mathcal{B}^*\Pi\mathcal{A}_{\circlearrowleft}(t)\right) = 0, \tag{5.2}$$

$$\mathcal{C}_{\circlearrowleft}^* J\pi_{[0,t]}\mathcal{C}_{\circlearrowleft} + \mathcal{A}_{\circlearrowleft}^*(t)\Pi\mathcal{A}_{\circlearrowleft}(t) = \Pi, \tag{5.3}$$

$$\mathcal{C}^* J\pi_{[0,t]}\mathcal{C}_{\circlearrowleft} + \mathcal{A}^*(t)\Pi\mathcal{A}_{\circlearrowleft}(t) = \Pi. \tag{5.4}$$

Proof. Fix $t \geq 0$. Let us perform the minimization of the cost function $Q(x_0, u)$ separately with respect to $u_1 = \pi_{[0,t]}u$ and $u_2 = \pi_{[t,\infty)}u$. To do this we write $Q(x_0, u)$ in the form

$$Q(x_0, u) = \int_0^t \langle y(s), Jy(s)\rangle_Y \, ds + \int_t^\infty \langle y(s), Jy(s)\rangle_Y \, ds,$$

where y is the output of Ψ with initial time zero, initial value x_0, and control u. Let x be the corresponding state of Ψ. Since

$$\pi_{[0,t]}y = \pi_{[0,t]}\mathcal{C}x_0 + \pi_{[0,t]}\mathcal{D}\pi_{[0,t]}u \text{ and } \pi_{[t,\infty)}y = \tau(-t)\mathcal{C}x(t) + \mathcal{D}\pi_{[t,\infty)}u,$$

we observe that the first term depends only on x_0 and u_1 and the second term only on $x(t)$ and u_2. If we fix u_1 and minimize with respect to u_2, then, because of the time invariance of the problem, the minimum is equal to $\langle x(t), \Pi x(t)\rangle_H$. Thus, we are left with the problem of minimizing the cost function

$$Q(x_0, u) = \int_0^t \langle y(s), Jy(s)\rangle_Y \, ds + \langle x(t), \Pi x(t)\rangle_H \tag{5.5}$$

with respect to u_1, where $\pi_{[0,t]}y$ is the function given above and $x(t) = \mathcal{A}(t)x_0 + \mathcal{B}\tau(t)u_1$. Differentiating (5.5) with respect to u_1 and setting the result to be zero we get (5.1). That also (5.2) holds follows from the time invariance and anticausality of $\mathcal{M}^*$ and the fact that $\mathcal{M}^*\mathcal{D}^* = \mathcal{N}^*$.

By replacing y and x in (5.5) by $y^{\text{opt}}(x_0)$ and $x^{\text{opt}}(x_0)$ we find that

$$\langle x_0, \Pi x_0\rangle_H = \int_0^t \left\langle y^{\text{opt}}(x_0, s), Jy^{\text{opt}}(x_0, s)\right\rangle_Y \, ds + \left\langle x^{\text{opt}}(t, x_0), \Pi x^{\text{opt}}(t, x_0)\right\rangle_Y,$$

from which (5.3) follows since $y^{\text{opt}}(x_0) = \mathcal{C}_{\circlearrowleft}x_0$ and $x^{\text{opt}}(t, x_0) = \mathcal{A}_{\circlearrowleft}(t)x_0$. This

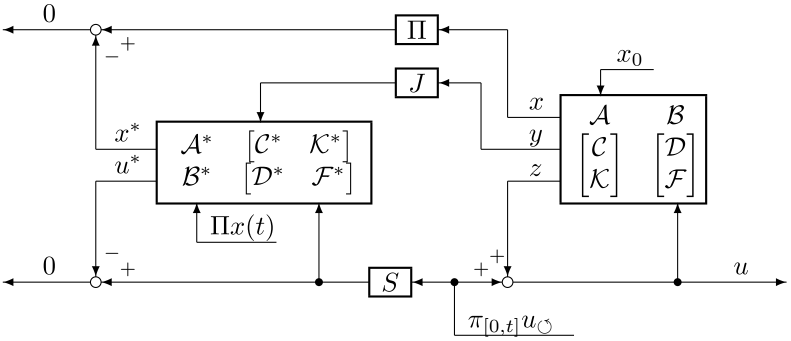

0
Π
$-$
$+$
J
x_0
x
$\mathcal{A}$
$\mathcal{B}$
x^*
$\mathcal{A}^*$
$\begin{bmatrix}\mathcal{C}^* & \mathcal{K}^*\end{bmatrix}$
$\mathcal{B}^*$
$\begin{bmatrix}\mathcal{D}^* & \mathcal{F}^*\end{bmatrix}$
y
$\begin{bmatrix}\mathcal{C}\\ \mathcal{K}\end{bmatrix}$
$\begin{bmatrix}\mathcal{D}\\ \mathcal{F}\end{bmatrix}$
u^*
z
$\Pi x(t)$
0
$-$
$+$
S
$+$
$+$
u
$\pi_{[0,t]}u_\circlearrowleft$

FIG. 5.1. *Primal-dual connection with primal feedback.*

combined with (5.1) implies that for all x_0 and x_1 in H,

$$\begin{aligned}\langle x_1, \Pi x_0\rangle_H &= \int_0^t \langle (\mathcal{C}_\circlearrowleft x_1)(s), J(\mathcal{C}_\circlearrowleft x_0)(s)\rangle_Y \, ds + \langle \mathcal{A}_\circlearrowleft(t)x_1, \Pi\mathcal{A}_\circlearrowleft(t)x_0\rangle_H \\ &= \int_0^t \left\langle (\mathcal{C}x_1 + \mathcal{D}u^{\text{opt}}(x_1))(s), Jy^{\text{opt}}(x_0,s)\right\rangle_Y \, ds \\ &\quad + \left\langle \mathcal{A}(t)x_1 + \mathcal{B}\tau(t)u^{\text{opt}}(x_1), \Pi x^{\text{opt}}(x_0,t)\right\rangle_H \\ &= \int_0^t \langle (\mathcal{C}x_1)(s), J\mathcal{C}_\circlearrowleft x_0(s)\rangle_Y \, ds + \langle \mathcal{A}(t)x_1, \mathcal{A}_\circlearrowleft(t)x_0\rangle_H \,,\end{aligned}$$

which gives us (5.4). □

We remark that Lemma 5.1 has been proved independently by Hans Zwart [35] under weaker assumptions.

The preceding lemma can be interpreted as a result concerning the state and output of the adjoint system Ψ^* if we use the state x multiplied by Π and the output y multiplied by J of the original system as initial value and control for Ψ.

COROLLARY 5.2. *Make the same assumption and introduce the same notations as in Theorem* 4.4. *Let* $x^{\text{opt}}(x_0)$ *and* $y^{\text{opt}}(x_0)$ *denote the optimal state and optimal output in the quadratic cost minimization problem for* Ψ, *and let* x^* *and* u^* *denote the state and output of the adjoint system* Ψ^* *with initial time* $t > 0$, *initial value* $\Pi x^{\text{opt}}(t, x_0)$, *and control* $y^{\text{opt}}(x_0)$ (*see Figure* 5.1 *with* $u_\circlearrowleft = 0$). *Then* $x^*(s) = \Pi x^{\text{opt}}(x_0, s)$ *for all* $s \in [0,t]$ *and* $\pi_{[0,t]}u^* = 0$. *The same formulae are true if instead* x^* *and* u^* *denote the state and output of the optimal adjoint system* $\Psi^*_\circlearrowleft$ *with initial time* $t > 0$, *initial value* $\Pi x^{\text{opt}}(t, x_0)$, *and control* $y^{\text{opt}}(x_0)$.

Proof. Fix $0 \le s \le t$. By the definition of the state of the adjoint system Ψ^*, we have

$$\begin{aligned}x^*(s) &= \mathcal{A}^*(t-s)\Pi x^{\text{opt}}(t,x_0) + \mathcal{C}^*J\tau(s)\pi_{[s,t]}y^{\text{opt}}(x_0) \\ &= \left(\mathcal{A}^*(t-s)\Pi\mathcal{A}_\circlearrowleft(t-s) + \mathcal{C}^*J\tau(s)\pi_{[s,t]}\tau(-s)\mathcal{C}_\circlearrowleft\right) x^{\text{opt}}(x_0,s) \\ &= \left(\mathcal{A}^*(t-s)\Pi\mathcal{A}_\circlearrowleft(t-s) + \mathcal{C}^*J\pi_{[0,t-s]}\mathcal{C}_\circlearrowleft\right) x^{\text{opt}}(x_0,s) \\ &= \Pi x^{\text{opt}}(x_0,s),\end{aligned}$$

where the last equality follows from (5.4). The same computation is valid if we replace Ψ^* by $\Psi^*_\circlearrowleft$.

The restriction of the output of Ψ^* to $[0,t]$ is given by

$$\pi_{[0,t]}u^* = \pi_{[0,t]}\left(\tau(-t)\mathcal{B}^*\Pi x^{\mathrm{opt}}(t,x_0) + \mathcal{D}^* J\pi_{[0,t]}y^{\mathrm{opt}}(x_0)\right),$$

and this is zero according to (5.1). To prove the same result with Ψ^* replaced by the optimal closed loop adjoint system $\Psi^*_{\circlearrowleft}$ we argue in the same way, but replace (5.1) by (5.2). □

REMARK 5.3. *A similar result is true for nonzero inputs $u_{\circlearrowleft}$ to the primal and $Su_{\circlearrowleft}$ to the dual system* (*see Figure* 5.1). *This follows from Corollary* 5.7 *below, since the connection in Figure* 5.3 *becomes identical to the one in Figure* 5.1 *if we replace u in Figure* 5.3 *by $u_{\circlearrowleft} + z$.*

Up to now we have in this section made only marginal use of Theorem 4.4, but the remaining results depend heavily on the characterization of the optimal feedback pair given in that theorem. We begin with the following key lemma.

LEMMA 5.4. *Make the same assumption and introduce the same notations as in Theorem* 4.4. *Then, for all $t \ge 0$,*

$$\pi_+\tau(-t)\mathcal{B}^*_{\circlearrowleft}\Pi\mathcal{B}_{\circlearrowleft}\tau(t)\pi_+ + \pi_{[0,t]}\mathcal{D}^*_{\circlearrowleft}J\pi_{[0,t]}\mathcal{D}_{\circlearrowleft}\pi_{[0,t]} = S\pi_{[0,t]}.$$

Proof. To prove this we compute, using the facts that $\Pi = \mathcal{C}^*_{\circlearrowleft}J\mathcal{C}_{\circlearrowleft}$ and $\pi_+\mathcal{D}_{\circlearrowleft}\pi_- = \pi_+\mathcal{N}\pi_- = \mathcal{C}_{\circlearrowleft}\mathcal{B}_{\circlearrowleft}$,

$$\begin{aligned}
&\pi_+\tau(-t)\mathcal{B}^*_{\circlearrowleft}\Pi\mathcal{B}_{\circlearrowleft}\tau(t)\pi_+ + \pi_{[0,t]}\mathcal{D}^*_{\circlearrowleft}J\pi_{[0,t]}\mathcal{D}_{\circlearrowleft}\pi_{[0,t]}\\
&\quad= \pi_+\tau(-t)\mathcal{B}^*_{\circlearrowleft}\mathcal{C}^*_{\circlearrowleft}J\mathcal{C}_{\circlearrowleft}\mathcal{B}_{\circlearrowleft}\tau(t)\pi_+ + \pi_{[0,t]}\mathcal{D}^*_{\circlearrowleft}J\pi_{[0,t]}\mathcal{D}_{\circlearrowleft}\pi_{[0,t]}\\
&\quad= \pi_{[0,t]}\tau(-t)\pi_-\mathcal{N}^*J\pi_+\mathcal{N}\pi_-\tau(t)\pi_+ + \pi_{[0,t]}\mathcal{D}^*_{\circlearrowleft}J\pi_{[0,t]}\mathcal{D}_{\circlearrowleft}\pi_{[0,t]}.
\end{aligned}$$

The combination of operators in the first term on the last row satisfies

$$\begin{aligned}
\pi_+\mathcal{N}\pi_-\tau(t)\pi_+ &= \pi_+\mathcal{N}\tau(t)\pi_{[0,t]}\\
&= \pi_+\tau(t)\mathcal{N}\pi_{[0,t]}\\
&= \tau(t)\pi_{[t,\infty)}\mathcal{N}\pi_{[0,t]},
\end{aligned}$$

so we can continue the computation above as (recalling that $\mathcal{N}^*J\mathcal{N} = S$)

$$\begin{aligned}
&\pi_+\tau(-t)\mathcal{B}^*_{\circlearrowleft}\Pi\mathcal{B}_{\circlearrowleft}\tau(t)\pi_+ + \pi_{[0,t]}\mathcal{D}^*_{\circlearrowleft}J\pi_{[0,t]}\mathcal{D}_{\circlearrowleft}\pi_{[0,t]}\\
&\quad= \pi_{[0,t]}\mathcal{N}^*J\pi_{[t,\infty)}\tau(-t)\tau(t)\pi_{[t,\infty)}\mathcal{N}\pi_{[0,t]} + \pi_{[0,t]}\mathcal{D}^*_{\circlearrowleft}J\pi_{[0,t]}\mathcal{D}_{\circlearrowleft}\pi_{[0,t]}\\
&\quad= \pi_{[0,t]}\mathcal{N}^*J\pi_{[t,\infty)}\mathcal{N}\pi_{[0,t]} + \pi_{[0,t]}\mathcal{N}^*J\pi_{[0,t]}\mathcal{N}\pi_{[0,t]}\\
&\quad= \pi_{[0,t]}\mathcal{N}^*J\mathcal{N}\pi_{[0,t]}\\
&\quad= S\pi_{[0,t]},
\end{aligned}$$

from which the claim follows. □

LEMMA 5.5. *Make the same assumption and introduce the same notations as in Theorem* 4.4. *Then, for all $t \ge 0$,*

$$\begin{aligned}
&(5.6) & S\pi_{[0,t]}\mathcal{K} &= -\pi_{[0,t]}\left(\mathcal{N}^*J\pi_{[0,t]}\mathcal{C} + \tau(-t)\mathcal{M}^*\mathcal{B}^*\Pi\mathcal{A}(t)\right)\\
&& &= -\pi_{[0,t]}\left(\mathcal{D}^*_{\circlearrowleft}J\pi_{[0,t]}\mathcal{C} + \tau(-t)\mathcal{B}^*_{\circlearrowleft}\Pi\mathcal{A}(t)\right),\\
&(5.7) & S\pi_{[0,t]}\mathcal{M}^{-1}\pi_+ &= \pi_{[0,t]}\left(\mathcal{N}^*J\pi_{[0,t]}\mathcal{D} + \tau(-t)\mathcal{M}^*\mathcal{B}^*\Pi\mathcal{B}\tau(t)\right)\pi_{[0,t]}\\
&& &= \pi_{[0,t]}\left(\mathcal{D}^*_{\circlearrowleft}J\pi_{[0,t]}\mathcal{D} + \tau(-t)\mathcal{B}^*_{\circlearrowleft}\Pi\mathcal{B}\tau(t)\right)\pi_{[0,t]},\\
&(5.8) & \Pi &= \mathcal{A}^*(t)\Pi\mathcal{A}(t) + \mathcal{C}^*J\pi_{[0,t]}\mathcal{C} - \mathcal{K}^*S\pi_{[0,t]}\mathcal{K}.
\end{aligned}$$

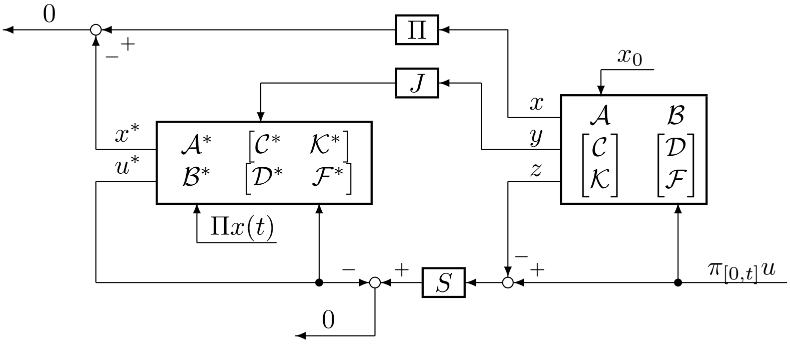

0
−
+
Π
J
x_0
x
y
z
x^*
u^*
$\mathcal{A}^*$ $\begin{bmatrix}\mathcal{C}^* & \mathcal{K}^*\end{bmatrix}$
$\mathcal{B}^*$ $\begin{bmatrix}\mathcal{D}^* & \mathcal{F}^*\end{bmatrix}$
$\Pi x(t)$
$\mathcal{A}$ $\mathcal{B}$
$\begin{bmatrix}\mathcal{C}\\ \mathcal{K}\end{bmatrix}$ $\begin{bmatrix}\mathcal{D}\\ \mathcal{F}\end{bmatrix}$
−
+
S
−
+
$\pi_{[0,t]}u$
0

FIG. 5.2. *Primal-dual connection with dual feedback.*

Proof. Since $y^{\mathrm{opt}}(x_0) = (\mathcal{C} + \mathcal{D}_{\circlearrowleft}\mathcal{K})x_0$ and $x^{\mathrm{opt}}(x_0, t) = (\mathcal{A}(t) + \mathcal{B}_{\circlearrowleft}\tau(t)\mathcal{K})$, we get from (5.2) and Lemma 5.4

$$\begin{aligned}
0 &= \pi_{[0,t]}\mathcal{D}^*_{\circlearrowleft}J\pi_{[0,t]}y^{\mathrm{opt}}(x_0) + \pi_{[0,t]}\tau(-t)\mathcal{B}^*_{\circlearrowleft}\Pi x^{\mathrm{opt}}(x_0, t)\\
&= \pi_{[0,t]}\mathcal{D}^*_{\circlearrowleft}J\pi_{[0,t]}\left(\mathcal{C} + \mathcal{D}_{\circlearrowleft}\mathcal{K}\right)x_0\\
&\quad + \pi_{[0,t]}\tau(-t)\mathcal{B}^*_{\circlearrowleft}\Pi\left(\mathcal{A}(t) + \mathcal{B}_{\circlearrowleft}\tau(t)\mathcal{K}\right)x_0\\
&= \pi_{[0,t]}\mathcal{D}^*_{\circlearrowleft}J\pi_{[0,t]}\mathcal{C}x_0 + \pi_{[0,t]}\tau(-t)\mathcal{B}^*_{\circlearrowleft}\Pi\mathcal{A}(t)x_0\\
&\quad + \pi_{[0,t]}\mathcal{D}^*_{\circlearrowleft}J\pi_{[0,t]}\mathcal{D}_{\circlearrowleft}\mathcal{K}x_0\\
&\quad + \pi_{[0,t]}\tau(-t)\mathcal{B}^*_{\circlearrowleft}\Pi\mathcal{B}_{\circlearrowleft}\tau(t)\mathcal{K}x_0\\
&= \pi_{[0,t]}\mathcal{D}^*_{\circlearrowleft}J\pi_{[0,t]}\mathcal{C}x_0 + \pi_{[0,t]}\tau(-t)\mathcal{B}^*_{\circlearrowleft}\Pi\mathcal{A}(t)x_0\\
&\quad + \pi_{[0,t]}S\mathcal{K}x_0.
\end{aligned}$$

This is (5.6).

The proof of (5.7) is very similar, and we leave it to the reader.

The identity (5.8) follows from (5.4) and (5.6) since they give

$$\begin{aligned}
\Pi &= \mathcal{A}^*_{\circlearrowleft}(t)\Pi\mathcal{A}(t) + \mathcal{C}^*_{\circlearrowleft}J\pi_{[0,t]}\mathcal{C}\\
&= \left(\mathcal{A}^*(t) + \mathcal{K}^*\mathcal{M}^*\tau(-t)\mathcal{B}^*\right)\Pi\mathcal{A}(t) + \left(\mathcal{C}^*J + \mathcal{K}^*\mathcal{N}^*J\right)\pi_{[0,t]}\mathcal{C}\\
&= \mathcal{A}^*(t)\Pi\mathcal{A}(t) + \mathcal{C}^*J\pi_{[0,t]}\mathcal{C} - \mathcal{K}^*S\pi_{[0,t]}\mathcal{K}. \qquad \square
\end{aligned}$$

Lemma 5.5 can be used to derive the following result.

COROLLARY 5.6. *Make the same assumption and introduce the same notations as in Theorem* 4.4. *Let x and y denote the state and output of Ψ with initial time zero, initial state x_0, and control u, and let x^* and u^* denote the state and output of the closed loop optimal adjoint system $\Psi^*_{\circlearrowleft}$ with initial time $t > 0$, initial value $\Pi x(t)$, and control Jy* (*see Figure* 5.2). *Then $x^*(s) = \Pi x(s)$ for all $s \in [0, t]$ and u^* is given by*

$$\pi_{[0,t]}u^* = -S\pi_{[0,t]}\left(\mathcal{K}x_0 - (I - \mathcal{F})\pi_{[0,t]}u\right).$$

Thus, apart from the factor $-S$ and the different feed-through term, this is the same signal that is produced by the optimal state feedback output of Ψ.

We leave the proof of this corollary to the reader since it is essentially the same as the proof of Corollary 5.2, with Lemma 5.1 replaced by Lemma 5.5.

It is possible to reformulate the preceding result in a way that does not involve any feedback, only feed-forward.

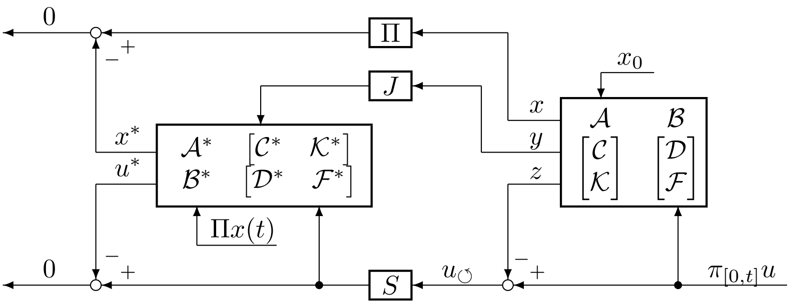

0
$-$
$+$
Π
J
x_0
x
y
z
$\mathcal{A}$
$\mathcal{B}$
$\begin{bmatrix}\mathcal{C}\\\mathcal{K}\end{bmatrix}$
$\begin{bmatrix}\mathcal{D}\\\mathcal{F}\end{bmatrix}$
x^*
u^*
$\mathcal{A}^*$
$\begin{bmatrix}\mathcal{C}^* & \mathcal{K}^*\end{bmatrix}$
$\mathcal{B}^*$
$\begin{bmatrix}\mathcal{D}^* & \mathcal{F}^*\end{bmatrix}$
$\Pi x(t)$
0
$-$
$+$
S
$u_{\circlearrowleft}$
$-$
$+$
$\pi_{[0,t]}u$

FIG. 5.3. *Feed-forward primal-dual connection.*

COROLLARY 5.7. *Make the same assumption and introduce the same notations as in Theorem* 4.4. *Let x, y, and z denote the state, the output, and the state feedback output of Ψ with initial time zero, initial state x_0, and control u, and let x^* and u^* denote the state and output of the adjoint system Ψ^* with initial time $t > 0$, initial value $\Pi x(t)$, control Jy, and output injection signal $S(u - z)$ (see Figure* 5.3). *Then $x^*(s) = \Pi x(s)$ for all $s \in [0, t]$ and u^* is given by*

$$\pi_{[0,t]}u^* = -S\pi_{[0,t]}\left(\mathcal{K}x_0 - (I - \mathcal{F})\pi_{[0,t]}u\right).$$

Proof. This follows from Corollary 5.6, which tells us that all the input signals (and initial states) in Figures 5.2 and 5.3 are identical; hence the outputs are also identical. □

6. The algebraic Riccati equation. With the aid of the formulae in the preceding section we can repeat the computations in [23, Sections 9 and 10] with the following results. (We refer the reader to [2], [23, Sections 7 and 8], and [29] for discussions on the generating operators of well-posed linear systems.)

THEOREM 6.1. *Make the same assumptions and introduce the same notations as in Theorem* 4.4. *Extend the system Ψ into*

$$\Psi = \begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \begin{bmatrix}\mathcal{C}\\\mathcal{K}\end{bmatrix} & \begin{bmatrix}\mathcal{D}\\\mathcal{F}\end{bmatrix} \end{bmatrix}$$

by adding the optimal state feedback pair $(\mathcal{K}, \mathcal{F})$. Let

$$\Psi_{\circlearrowleft} = \begin{bmatrix} \mathcal{A}_{\circlearrowleft} & \mathcal{B}_{\circlearrowleft} \\ \begin{bmatrix}\mathcal{C}_{\circlearrowleft}\\\mathcal{K}_{\circlearrowleft}\end{bmatrix} & \begin{bmatrix}\mathcal{D}_{\circlearrowleft}\\\mathcal{F}_{\circlearrowleft}\end{bmatrix} \end{bmatrix} = \begin{bmatrix} \mathcal{A} + \mathcal{B}\tau\mathcal{M}\mathcal{K} & \mathcal{B}\mathcal{M} \\ \begin{bmatrix}\mathcal{C} + \mathcal{N}\mathcal{K}\\\mathcal{M}\mathcal{K}\end{bmatrix} & \begin{bmatrix}\mathcal{N}\\\mathcal{M} - I\end{bmatrix} \end{bmatrix}$$

be the optimal closed loop system given by Theorem 4.4. *Denote the generating operators of Ψ and $\Psi_{\circlearrowleft}$ by the same letters as the corresponding operators* [23, *Sections* 7 *and* 8].

(i) *The Riccati operator* Π *of* Ψ *satisfies the Lyapunov equations*

$$\begin{aligned}\langle Ax_0, \Pi x_1\rangle_H + \langle x_0, \Pi A x_1\rangle_H &= -\langle Cx_0, JCx_1\rangle_Y + \langle Kx_0, SKx_1\rangle_U\,,\\ &\qquad x_0, x_1 \in \operatorname{dom}(A),\\ \langle Ax_0, \Pi x_1\rangle_H + \langle x_0, \Pi A_{\circlearrowleft} x_1\rangle_H &= -\langle Cx_0, JC_{\circlearrowleft}x_1\rangle_Y\,,\\ &\qquad x_0 \in \operatorname{dom}(A), \quad x_1 \in \operatorname{dom}(A_{\circlearrowleft}),\\ \langle A_{\circlearrowleft}x_0, \Pi x_1\rangle_H + \langle x_0, \Pi A_{\circlearrowleft} x_1\rangle_H &= -\langle C_{\circlearrowleft}x_0, JC_{\circlearrowleft}x_1\rangle_Y\,,\\ &\qquad x_0, x_1 \in \operatorname{dom}(A_{\circlearrowleft}).\end{aligned}$$

(ii) *The Lyapunov equations in* (i) *can be rewritten in the form*

$$\begin{aligned}\Pi Ax &= -\left(A^*\Pi + C^*JC - K^*SK\right)x\\ &= -\left(A_{\circlearrowleft}^*\Pi + C_{\circlearrowleft}^*JC\right)x, \qquad x \in \operatorname{dom}(A),\\ \Pi A_{\circlearrowleft}x &= -\left(A^*\Pi + C^*JC_{\circlearrowleft}\right)x\\ &= -\left(A_{\circlearrowleft}^*\Pi + C_{\circlearrowleft}^*JC_{\circlearrowleft}\right)x, \qquad x \in \operatorname{dom}(A_{\circlearrowleft}).\end{aligned}$$

(iii) *In addition, suppose that the extended system* Ψ *is regular together with its adjoint* [28, *Theorem* 5.8]. *Denote the feed-through operators with the same letters as their corresponding input/output maps* [23, *Sections* 7 *and* 8], *and let an over-line denote the strong Weiss extension of an observation map* (*see* [23, *Proposition* 36], *for example,* $\overline{C}x = \lim_{\lambda\to\infty} C_\lambda x$, *where* $C_\lambda = \lambda C(\lambda I - A)^{-1}$ *is the* "*Yosida approximation*" *of* C). *Then*

$$\begin{bmatrix} A_{\circlearrowleft} & B_{\circlearrowleft}\\ \begin{bmatrix} C_{\circlearrowleft}\\ K_{\circlearrowleft}\end{bmatrix} & \begin{bmatrix} D_{\circlearrowleft}\\ F_{\circlearrowleft}\end{bmatrix}\end{bmatrix} = \begin{bmatrix} A + BM\overline{K} & BM\\ \begin{bmatrix} \overline{C} + N\overline{K}\\ M\overline{K}\end{bmatrix} & \begin{bmatrix} N\\ M - I\end{bmatrix}\end{bmatrix},$$

where the equation for $B_{\circlearrowleft}$ *should be interpreted as* $\overline{B}_{\circlearrowleft}^* = M^*\overline{B}^*$.

(iv) *In the regular case* (iii) *above, the operator* $\overline{B}^*\Pi$ *satisfies the equations*

$$\begin{aligned}SKx &= -M^*\left(\overline{B}^*\Pi + D^*JC\right)x, \quad x \in \operatorname{dom}(A),\\ 0 &= \left(\overline{B}^*\Pi + D^*JC_{\circlearrowleft}\right)x, \quad x \in \operatorname{dom}(A_{\circlearrowleft}).\end{aligned}$$

(v) *In the regular case* (iii) *above, the Riccati operator* Π *satisfies the algebraic Riccati equation*

$$\begin{aligned}&\langle Ax_0, \Pi x_1\rangle_H + \langle x_0, \Pi Ax_1\rangle_H + \langle Cx_0, JCx_1\rangle_Y\\ &\quad = \left\langle M^*\left(\overline{B}^*\Pi + D^*JC\right)x_0, S^{-1}M^*\left(\overline{B}^*\Pi + D^*JC\right)x_1\right\rangle_U,\\ &\qquad x_0, x_1 \in \operatorname{dom}(A).\end{aligned}$$

In particular, if $M = I$,[9] *then*

$$\begin{aligned}&\langle Ax_0, \Pi x_1\rangle_H + \langle x_0, \Pi Ax_1\rangle_H + \langle Cx_0, JCx_1\rangle_Y\\ &\quad = \left\langle\left(\overline{B}^*\Pi + D^*JC\right)x_0, S^{-1}\left(\overline{B}^*\Pi + D^*JC\right)x_1\right\rangle_U,\\ &\qquad x_0, x_1 \in \operatorname{dom}(A).\end{aligned}$$

[9] This means that the feed-through operator of $\mathcal{F}$ is taken to be zero; i.e., there is "no feed-forward term inside the feedback loop." See the discussion after [23, Theorem 27].

Proof. (i) Take x_0 and x_1 in the indicated domains, apply (5.3), (5.4), or (5.8) to x_1, take the inner product with x_0, differentiate with respect to t, and substitute $t = 0$.

(ii) This follows from (i).

(iii) These formulae are found in [23, Sections 7 and 8] and proved in [29].

(iv) Let $x = \mathcal{A}x_0$ and $y = \mathcal{C}x_0$ denote the state and output of Ψ with initial time zero, initial state x_0, and zero control u. Referring to Corollary 5.6, we let $x^* = \Pi x$ and $u^* = -S\mathcal{K}x_0$ be the state and output of the optimal adjoint system $\Psi^*_{\circlearrowleft}$ with initial time $t > 0$, initial value $\Pi x(t)$, and control Jy (restricted to the interval $[0,t]$). Take $x_0 \in \operatorname{dom}(A)$. Then all the inputs and outputs belong to $W^{1,2}([0,t])$, $x(t) \in \operatorname{dom}(A)$, and, for all $s \in [0,t]$,

$$y(s) = Cx(s), \qquad u^*(s) = -SKx(s)$$

(the proofs of these claims are analogous to the proofs of [23, Propositions 29 and 36]). By part (ii), the initial values of the state $x^*(t) = \Pi x(t)$ and control $Jy(t) = JCx(t)$ satisfy

$$A^*_{\circlearrowleft}x^*(t) + C^*_{\circlearrowleft}Jy(t) = (A^*_{\circlearrowleft}\Pi + C^*_{\circlearrowleft}JC)x(t) = \Pi Ax(t) \in H.$$

Thus, by [23, Proposition 36(ii)], the output u^* of $\Psi^*_{\circlearrowleft}$ is related to the input y and the state x^* through the formula

$$u^*(s) = \overline{B}^*_{\circlearrowleft}x^*(s) + D^*_{\circlearrowleft}Jy(s),$$

which combined with (iii) gives us

$$-SKx(s) = u^*(s) = M^*\overline{B}^*\Pi x(s) + N^*JCx(s) = M^*(\overline{B}^*\Pi + D^*JC)x(s).$$

Taking $s = 0$ we get the first formula in (iv). We get the second formula by replacing Corollary 5.6 by Corollary 5.2.

(v) Combine (i) and (iv). □

7. Computation of the sensitivity operator. Looking at the different formulae involving $\overline{B}^*\Pi$ in part (iv) of Theorem 6.1, a natural question to ask is whether it is possible to compute $\overline{B}^*\Pi x$ for all x in the Hilbert space W_B defined in [23, Section 7]. This is the space of all possible initial values x_0 satisfying the equation $Ax_0 + Bu_0 \in H$ for some $u_0 \in U$ [23, Lemma 32]. It is invariant in the sense that the controlled state $x(t)$ of Ψ stays in W_B under the action of a control u in $W^{1,2}(\mathbf{R}^+;U)$, provided the initial values x_0 and $u_0 = u(0)$ satisfy $Ax_0 + Bu_0 \in H$ [23, Remark 34]. Moreover, it contains both the domains $\operatorname{dom}(A)$ and $\operatorname{dom}(A_{\circlearrowleft})$; in fact, it contains all the domains of the generators of any state feedback perturbed version of $\mathcal{A}$ [23, Proposition 37]. Another related question is whether it is possible to write all the different Lyapunov equations given in part (ii) of Theorem 6.1 into one common form. A third related question concerns the crucial sensitivity operator S appearing in Theorem 6.1: is it possible to give a formula for this operator in terms of the original data and the Riccati operator Π? The answers to all these questions are affirmative, as can be shown with the aid of Corollary 5.7.

Theorem 7.1. *Make the same assumptions and introduce the same notations as in Theorem* 4.4. *Denote the generating operators of* Ψ *by the same letters as the corresponding operators* [23, *Section* 7], *and let* $\widehat{\mathcal{D}}$ *and* $\widehat{\mathcal{F}}$ *be the transfer functions of* $\mathcal{D}$ *and* $\mathcal{F}$ [24, *Lemma* 2.9]. *Let* $x_0 \in H$ *and* $u_0 \in U$ *satisfy* $Ax_0 + Bu_0 \in H$.

(i) *If $\alpha \in \mathbf{C}$ has real part bigger than the growth rate of Ψ, then the vectors $y_0 \in Y$ and $w_0 \in U$ defined by*[10]

$$y_0 = C(\alpha I - A)^{-1}(\alpha x_0 - Ax_0 - Bu_0) + \widehat{\mathcal{D}}(\alpha)u_0, \tag{7.1}$$

$$w_0 = -K(\alpha I - A)^{-1}(\alpha x_0 - Ax_0 - Bu_0) + (I - \widehat{\mathcal{F}}(\alpha))u_0 \tag{7.2}$$

are independent of α. Moreover,

$$A^*\Pi x_0 + C^*Jy_0 + K^*Sw_0 = -\Pi\,(Ax_0 + Bu_0) \in H, \tag{7.3}$$

and, for all $\beta \in \mathbf{C}$ with real part bigger than the growth rate of Ψ,

$$(I - \widehat{\mathcal{F}}(\beta))^*Sw_0 = B^*(\overline{\beta}I - A^*)^{-1}\Pi(\overline{\beta}x_0 + Ax_0 + Bu_0) + (\widehat{\mathcal{D}}(\beta))^*Jy_0. \tag{7.4}$$

In particular, Π maps the space W_B defined in [23, *Section* 7] *continuously into the space $V^*_{(C,K)}$ defined in* [23, *Proposition* 39].

(ii) *If Ψ is regular, then* (7.1) *and* (7.2) *can be written in the alternative forms*

$$y_0 = \overline{C}x_0 + Du_0, \tag{7.5}$$

$$w_0 = -\overline{K}x_0 + (I - F)u_0 = -\overline{K}x_0 + Xu_0, \tag{7.6}$$

which substituted into (7.3) *gives*

$$\begin{aligned}(A^*\Pi + C^*J\overline{C} - K^*S\overline{K})x_0 + \Pi(Ax_0 + Bu_0)\\ = -(C^*JD + K^*SX)u_0.\end{aligned} \tag{7.7}$$

(iii) *If Ψ^* is regular, then* (7.4) *can be written in the alternative form*

$$(I - F^*)Sw_0 = X^*Sw_0 = \overline{B}^*\Pi x_0 + D^*Jy_0. \tag{7.8}$$

(iv) *If both Ψ and Ψ^* are regular, then* (7.5) *and* (7.6) *combined with* (7.8) *give*

$$(\overline{B}^*\Pi + D^*J\overline{C} + X^*S\overline{K})\,x_0 = (X^*SX - D^*JD)u_0. \tag{7.9}$$

In particular, if $X = I$ (i.e., $F = 0$ and there is "no feed-forward term inside the feedback loop"), then

$$(\overline{B}^*\Pi + D^*J\overline{C} + S\overline{K})\,x_0 = (S - D^*JD)u_0.$$

A special case of the last formula is found in [18, Formula (60)]. (The setting in [18] is different, and that formula is actually valid in a much larger (Banach) space than W_B.)

Proof. (i) Take some $x_0 \in H$ and $u_0 \in U$ satisfying $Ax_0 + Bu_0 \in H$. Choose some $u \in W^{1,2}([0,t];U)$ with $u(0) = u_0$. Consider the connection described in Corollary 5.7. As in that corollary, we let $x = \mathcal{A}x_0 + \mathcal{B}\tau\pi_+u$, $y = \mathcal{C}x_0 + \mathcal{D}\pi_+u$, and $z = \mathcal{K}x_0 + \mathcal{F}\pi_+u$ denote the state, the output, and the state feedback output of Ψ with initial time zero, initial state x_0, and control u. Furthermore, let $w = (u - z)$. Then, by [23, Proposition 29], all the inputs and and outputs of the extended primal system Ψ

[10] For notational simplicity we have in this theorem replaced $u_\circlearrowleft$ in Figure 5.3 by w. It represents the input to the optimal closed loop system; cf. Figure 5.1.

belong to $W^{1,2}([0,t])$ and the state x is continuously differentiable in H. It follows from, for example, [16, Formula (2.1;2)] combined with [23, Proposition 29] that (7.1) and (7.2) hold with x_0, u_0, y_0, and w_0 replaced by $x(t)$, $u(t)$, $u(t)$, and $w(t)$ for all $t \geq 0$. In particular, defining $y_0 = y(0)$ and $w_0 = w(0)$ we get (7.1) and (7.2).

Let us continue with a discussion of the dual system Ψ^* in Corollary 5.7. We fix some $t > 0$ and let $x^*(s) = \mathcal{A}^*(t-s)\Pi x(t) + \mathcal{C}^* J\tau(s)\pi_{[0,t]}y + \mathcal{K}^* S\tau(s)\pi_{[0,t]}w$, $0 \leq s \leq t$, be the state and $u^* = \tau(-t)\mathcal{B}^*\Pi x(t) + \mathcal{D}^* J\pi_{[0,t]}y + \mathcal{F}^* S\pi_{[0,t]}w$ be the output of Ψ^* with initial time $t > 0$, initial value $\Pi x(t)$, control Jy, and output injection signal Sw (throughout restricting all the functions to the interval $[0,t]$). According to Corollary 5.7, $x^* = \Pi x$ and $u^* = Sw$. The former equation implies that x^* is continuously differentiable in H (since x is continuously differentiable in H). The derivative of x is $x' = Ax + Bu$ (see [23, Proposition 29]), and the derivative of x^* is $(x^*)' = -A^*x^* - C^*Jy - K^*Sw$ (this equation is always true in the larger space W^* defined in [23, Section 7], but this time we know that the derivative actually belongs to the smaller space H). Equating the derivative of x^* with the derivative of Πx we get (7.3). Equation (7.4) is derived in the same way as equations (7.1) and (7.2) were derived above, except that we also have to use the additional fact that $u^* = Sw$ (and we have used (7.3) to slightly simplify the result).

(ii) If Ψ is regular, then (7.5) and (7.6) are the limits of (7.1) and (7.2) as $\alpha \to \infty$.

(iii) If Ψ^* is regular, then (7.8) is the limit of (7.4) as $\beta \to \infty$.

(iv) This is immediate. □

The preceding theorem provides us with the following formula, among others, for the sensitivity operator S.

Corollary 7.2. *In the case where both the extended system Ψ and its adjoint are regular and $X = I$*[11] *the following additional claims are true:*

(i) *For all $u_0 \in U$, we have*

$$Su_0 = D^*JDu_0 + \lim_{\alpha\to\infty} \overline{B}^*\Pi(\alpha I - A)^{-1}Bu_0.$$

In particular, $S = D^*JD$ iff the limit above is zero for all $u_0 \in U$.

(ii) *If for some $u_0 \in U$ it is true that $Bu_0 \in H$, then*

$$Su_0 = D^*JDu_0.$$

(iii) *If $S = D^*JD$, then, for all $x_0 \in W_B$,*

$$\overline{K}x_0 = -S^{-1}\left(\overline{B}^*\Pi + D^*J\overline{C}\right)x_0.$$

Proof. To prove part (i) it suffices to apply (7.9) with $x_0 = (\alpha I - A)^{-1}Bu_0$, let $\alpha \to \infty$, and use the regularity of the system. Part (ii) follows directly from (7.9) with $x_0 = 0$. The final claim is obvious (see (7.9)). □

Remark 7.3. *As we shall prove elsewhere* [25], *the difference $S - D^*JD$ is positive* (*negative*) *definite whenever Π is positive* (*negative*) *definite on the reachable subspace.* (*The proof is a fairly straightforward application of Lemma* 5.4.) *This is related to the fact that the factorization is* (J,S)*-lossless iff Π is positive on the reachable subspace.*

8. Applications: The bounded and positive real lemmas.

By applying the preceding theory we can derive the first available versions of the strict bounded and positive (real) lemmas for general well-posed linear systems.

[11] We lose no generality by assuming that $X = I$; see Proposition 4.7.

In the positive real lemma and the bounded real lemma we need a cost function containing both the output y and the control u. For this reason we do in the same way as in Corollary 4.9 and adjoin a copy of the control to the output; i.e., we study the augmented system

$$\Psi_{\text{aug}} = \begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \begin{bmatrix} \mathcal{C} \\ 0 \end{bmatrix} & \begin{bmatrix} \mathcal{D} \\ I \end{bmatrix} \end{bmatrix}. \tag{8.1}$$

By replacing the identity cost operator in Corollary 4.9 by a more general cost operator J defined on $U \times Y$, we get the following result.

COROLLARY 8.1. *Let $J = J^* = \left[\begin{smallmatrix} Q & L^* \\ L & R \end{smallmatrix}\right] \in \mathcal{L}(Y \times U)$, let $S = S^* \in \mathcal{L}(U)$, $S \gg 0$, and let $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$ be a jointly stabilizable and detectable well-posed linear system on (U, H, Y).*

(i) *The extended system Ψ_{aug} in* (8.1) *is J-coercive iff $\mathcal{D}$ has a right coprime factorization $(\mathcal{N}, \mathcal{M})$ for which $\left[\begin{smallmatrix} \mathcal{N} \\ \mathcal{M} \end{smallmatrix}\right]$ is (J, S)-inner.*

(ii) *Assuming J-coercivity, let $x^{\text{opt}}(x_0)$, $y^{\text{opt}}(x_0)$, and $u^{\text{opt}}(x_0)$ be the optimal state, output, and control for the quadratic cost minimization problem described in Definition* 3.1, *but with the original system Ψ replaced by the extended system Ψ_{aug}. Let $(\mathcal{N}, \mathcal{M})$ be a right coprime factorization of $\mathcal{D}$ of the type described in* (i). *Then there is a unique feedback map $\mathcal{K}$ such that $[\mathcal{K} \quad \mathcal{F}] = [\mathcal{K} \quad (I - \mathcal{M}^{-1})]$ is an admissible stabilizing state feedback pair for Ψ and*

$$\begin{bmatrix} x^{\text{opt}}(t, x_0) \\ y^{\text{opt}}(x_0) \\ u^{\text{opt}}(x_0) \end{bmatrix} = \begin{bmatrix} \mathcal{A}_{\circlearrowleft}(t) \\ \mathcal{C}_{\circlearrowleft} \\ \mathcal{K}_{\circlearrowleft} \end{bmatrix} x_0 = \begin{bmatrix} \mathcal{A}(t) + \mathcal{B}\mathcal{M}\tau(t)\mathcal{K} \\ \mathcal{C} + \mathcal{N}\mathcal{K} \\ \mathcal{M}\mathcal{K} \end{bmatrix} x_0$$

is equal to the state and output of the closed loop system $\Psi_{\circlearrowleft}$ defined by

$$\Psi_{\circlearrowleft} = \begin{bmatrix} \mathcal{A}_{\circlearrowleft} & \mathcal{B}_{\circlearrowleft} \\ \begin{bmatrix} \mathcal{C}_{\circlearrowleft} \\ \mathcal{K}_{\circlearrowleft} \end{bmatrix} & \begin{bmatrix} \mathcal{D}_{\circlearrowleft} \\ \mathcal{F}_{\circlearrowleft} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \mathcal{A} + \mathcal{B}\tau\mathcal{M}\mathcal{K} & \mathcal{B}\mathcal{M} \\ \begin{bmatrix} \mathcal{C} + \mathcal{N}\mathcal{K} \\ \mathcal{M}\mathcal{K} \end{bmatrix} & \begin{bmatrix} \mathcal{N} \\ \mathcal{M} - I \end{bmatrix} \end{bmatrix}$$

with initial value x_0, initial time zero, and zero input $u_{\circlearrowleft}$ (see Figure 2.1). *The feedback map $\mathcal{K}$ is uniquely determined by the fact that $\mathcal{C}_{\circlearrowleft} = \mathcal{C} + \mathcal{N}\mathcal{K} \in \mathcal{L}(H; L^2(\mathbf{R}^+; Y))$, $\mathcal{K}_{\circlearrowleft} = \mathcal{M}\mathcal{K} \in \mathcal{L}(H; L^2(\mathbf{R}^+; U))$, and*

$$\pi_+ \left[\mathcal{N}^* \quad \mathcal{M}^*\right] J \begin{bmatrix} \mathcal{C}_{\circlearrowleft} \\ \mathcal{K}_{\circlearrowleft} \end{bmatrix} = 0.$$

Moreover, the Riccati operator of Ψ is given by

$$\Pi = \left[\mathcal{C}_{\circlearrowleft}^* \quad \mathcal{K}_{\circlearrowleft}^*\right] J \begin{bmatrix} \mathcal{C}_{\circlearrowleft} \\ \mathcal{K}_{\circlearrowleft} \end{bmatrix}.$$

(iii) *If Ψ is stable, then*

$$\mathcal{K} = -S^{-1}\pi_+ \left[\mathcal{N}^* \quad \mathcal{M}^*\right] J \begin{bmatrix} \mathcal{C} \\ 0 \end{bmatrix} = -S^{-1}\pi_+ \left(\mathcal{N}^* Q + \mathcal{M}^* L\right) \mathcal{C}$$

and

$$\Pi = \mathcal{C}^* Q \mathcal{C} - \mathcal{K}^* S \mathcal{K}.$$

(iv) *In the case where the extended system* Ψ *is regular together with its adjoint the formulae connecting* K *and* Π *in Theorem* 6.1(iv)–(v) *become* (*with the normalization* $M = I$)

$$
\begin{gathered}
Kx_0 = -S^{-1}\big(\overline{B}^*\Pi + (D^*Q + L^*)C\big)x_0, \\
\langle Ax_0, \Pi x_1\rangle_H + \langle x_0, \Pi A x_1\rangle_H + \langle Cx_0, QCx_1\rangle_Y = \langle Kx_0, SKx_1\rangle_U\,, \\
x_0, x_1 \in \operatorname{dom}(A)
\end{gathered}
$$

and the sensitivity operator S *is given by the strong limit* (*for each fixed* $u_0 \in U$)

$$
Su_0 = \begin{bmatrix} D^* & I \end{bmatrix} J \begin{bmatrix} D \\ I \end{bmatrix} u_0 + \lim_{\alpha\to\infty} \overline{B}^*\Pi(\alpha I - A)^{-1}Bu_0.
$$

Proof. Part (i) follows from Lemma 4.3(ii). Part (ii) follows from Theorem 4.4 in the same way as Corollary 4.9 does. Part (iii) follows from Theorem 2.6. Part (iv) follows from Theorem 6.1 and Corollary 7.2(i). □

From this result we can obtain a factorization version of the *strict bounded (real) lemma* as follows: We let Ψ be stable and choose J to be

$$
J = \begin{bmatrix} -I & 0 \\ 0 & \gamma^2 I \end{bmatrix},
$$

where γ is a real constant. Then the extended system is J-coercive iff the input/output map $\mathcal{D}$ satisfies

$$
\|\mathcal{D}\|_{TIC(U;Y)} < \gamma. \tag{8.2}
$$

Thus, Corollary 8.1 applies iff (8.2) holds. In this case the formulae in Corollary 8.1(ii)–(iii) become

$$
\begin{gathered}
\mathcal{D} = \mathcal{N}\mathcal{M}^{-1}, \qquad \gamma^2\mathcal{M}^*\mathcal{M} - \mathcal{N}^*\mathcal{N} = S, \\
\mathcal{K} = S^{-1}\pi_+\mathcal{N}^*\mathcal{C}, \qquad \gamma^2\pi_+\mathcal{M}^*\mathcal{K}_{\circlearrowleft} = \pi_+\mathcal{N}^*\mathcal{C}_{\circlearrowleft}, \\
\begin{bmatrix} \mathcal{C}_{\circlearrowleft} \\ \mathcal{K}_{\circlearrowleft} \end{bmatrix} = \begin{bmatrix} \mathcal{C} \\ 0 \end{bmatrix} + \begin{bmatrix} \mathcal{N} \\ \mathcal{M} \end{bmatrix} \mathcal{K} = \begin{bmatrix} \mathcal{C} \\ 0 \end{bmatrix} + \begin{bmatrix} \mathcal{N} \\ \mathcal{M} \end{bmatrix} S^{-1}\pi_+\mathcal{N}^*\mathcal{C}, \\
\Pi = \gamma^2\mathcal{K}_{\circlearrowleft}^*\mathcal{K}_{\circlearrowleft} - \mathcal{C}_{\circlearrowleft}^*\mathcal{C}_{\circlearrowleft} = -\mathcal{C}^*\left(I + \mathcal{N}S^{-1}\pi_+\mathcal{N}^*\right)\mathcal{C}.
\end{gathered}
$$

The connecting and Lyapunov equations in Corollary 8.1(iv) become (for x_0 and $x_1 \in \operatorname{dom}(A)$)

$$
\begin{gathered}
Kx_0 = -S^{-1}\left(\overline{B}^*\Pi - D^*C\right)x_0, \\
\langle Ax_0, \Pi x_1\rangle_H + \langle x_0, \Pi A x_1\rangle_H = \langle Cx_0, Cx_1\rangle_Y + \langle Kx_0, SKx_1\rangle_U\,.
\end{gathered}
$$

Observe that the parameter γ enters these equations only through the sensitivity operator S, which is given by the strong limit (for each fixed $u_0 \in U$)

$$
Su_0 = \left(\gamma^2 I - D^*D\right)u_0 + \lim_{\alpha\to\infty} \overline{B}^*\Pi(\alpha I - A)^{-1}Bu_0.
$$

We remark that in our setting Π is negative definite; to get the standard setting where Π is positive [12, Theorem 3.7.1], we must replace J by $-J$ and maximize instead of minimize. This will replace S by $-S$ and Π by $-\Pi$.

The *strictly positive (real) lemma* is a statement about a stable system $\Psi = \left[\begin{smallmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right]$ on (U, H, U) (i.e., the output space of this system is equal to its input space). The input/output map $\mathcal{D}$ of Ψ is strictly positive iff

$$\int_{\mathbf{R}^+} \left(\langle (\mathcal{D}\pi_+ u)(s), u(s) \rangle_U + \langle u(s), (\mathcal{D}\pi_+ u)(s) \rangle_U \right) \, ds \geq \epsilon \, \|u\|^2_{L^2(\mathbf{R}^+;U)}$$

for all $u \in L^2(\mathbf{R}^+; U)$ and some $\epsilon > 0$. Clearly, $\mathcal{D}$ is strictly positive iff the extended system Ψ_{aug} is J-coercive with respect to the operator

$$J = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}.$$

Thus, Corollary 8.1 applies with this J iff $\mathcal{D}$ is strictly positive. The formulae of Corollary 8.1(ii)–(iii) become in this case

$$\begin{aligned} &\mathcal{D} = \mathcal{N}\mathcal{M}^{-1}, \qquad && \mathcal{M}^*\mathcal{N} + \mathcal{N}^*\mathcal{M} = S, \\ &\mathcal{K} = -S^{-1}\pi_+\mathcal{M}^*\mathcal{C}, \qquad && \pi_+ \left(\mathcal{M}^*\mathcal{C}_{\circlearrowleft} + \mathcal{N}^*\mathcal{K}_{\circlearrowleft} \right) = 0, \end{aligned}$$

$$\begin{bmatrix} \mathcal{C}_{\circlearrowleft} \\ \mathcal{K}_{\circlearrowleft} \end{bmatrix} = \begin{bmatrix} \mathcal{C} \\ 0 \end{bmatrix} + \begin{bmatrix} \mathcal{N} \\ \mathcal{M} \end{bmatrix} \mathcal{K} = \begin{bmatrix} \mathcal{C} \\ 0 \end{bmatrix} - \begin{bmatrix} \mathcal{N} \\ \mathcal{M} \end{bmatrix} S^{-1}\pi_+\mathcal{M}^*\mathcal{C},$$

$$\Pi = \mathcal{K}^*_{\circlearrowleft}\mathcal{C}_{\circlearrowleft} + \mathcal{C}^*_{\circlearrowleft}\mathcal{K}_{\circlearrowleft} = -\mathcal{K}^* S \mathcal{K} = -\mathcal{C}^*\mathcal{M}S^{-1}\pi_+\mathcal{M}^*\mathcal{C}.$$

The connecting and Lyapunov equations in Corollary 8.1(iv) become (for x_0 and $x_1 \in \operatorname{dom}(A)$)

$$\begin{aligned} K x_0 &= -S^{-1} \left(\overline{B}^* \Pi + C \right) x_0, \\ \langle A x_0, \Pi x_1 \rangle_H + \langle x_0, \Pi A x_1 \rangle_H &= \langle K x_0, S K x_1 \rangle_U \end{aligned}$$

and the sensitivity operator S is given by the strong limit (for each fixed $u_0 \in U$)

$$S u_0 = (D + D^*) \, u_0 + \lim_{\alpha \to \infty} \overline{B}^* \Pi (\alpha I - A)^{-1} B u_0.$$

Again Π is negative; to get a positive Π we should change the sign of J and maximize instead of minimize [12, Problem 3.25].

In the Pritchard–Salamon case the applications to the bounded and positive (real) lemmas that we have presented above are found in [31, Remark 4.34].

REFERENCES

[1] J. Bontsema and R. F. Curtain, *Perturbation properties of a class of infinite-dimensional systems with unbounded control and observation*, IMA J. Math. Control Inform., 5 (1988), pp. 333–352.

[2] R. F. Curtain, *Representations of infinite-dimensional systems*, in Three Decades of Mathematical Systems Theory, Lecture Notes in Control and Information Sciences 135, H. Nijmeijer and J. M. Schumacher, eds., Springer-Verlag, Berlin, 1989, pp. 101–128.

[3] R. F. Curtain and G. Weiss, *Well posedness of triples of operators (in the sense of linear systems theory)*, in Control and Optimization of Distributed Parameter Systems, Basel, Birkhäuser-Verlag, 1989, pp. 401–416.

[4] R. F. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.

[5] F. Flandoli, I. Lasiecka, and R. Triggiani, *Algebraic Riccati equations with non-smoothing observation arising in hyperbolic and Euler-Bernoulli boundary control problems*, Ann. Mat. Pura Appl., 153 (1988), pp. 307–382.

[6] B. A. Francis, *A Course in H_∞ Control Theory*, Lecture Notes in Control and Information Sciences 88, Springer-Verlag, Berlin, 1987.

[7] T. T. Georgiou and M. C. Smith, *Optimal robustness in the gap metric*, IEEE Trans. Automat. Control, 35 (1990), pp. 673–686.

[8] T. T. Georgiou and M. C. Smith, *Robust stabilization in the gap metric: Controller design for distributed plants*, IEEE Trans. Automat. Control, 37 (1992), pp. 1133–1143.

[9] P. Grabowski, *The LQ controller synthesis problem*, IMA J. Math. Control Inform., 10 (1993), pp. 131–148.

[10] P. Grabowski,*The LQ controller problem: An example*, IMA J. Math. Control Inform., 11 (1994), pp. 355–368.

[11] P. Grabowski and F. M. Callier, *Admissible Observation Operators. Duality of Observation and Control Using Factorizations*, Research Report 96–12, Facultés Universitaires de Namur. Publications du Département de Mathématique, FUNDP, Namur, Belgium, 1996.

[12] M. Green and D. L. L. Limebeer, *Linear Robust Control*, Prentice–Hall, Englewood Cliffs, NJ, 1995.

[13] I. Lasiecka, *Riccati equations arising from boundary and point control problems*, in Analysis and Optimization of Systems: State and Frequency Domain Approaches for Infinite-Dimensional Systems, Lecture Notes in Control and Information Sciences 185, R. F. Curtain, ed., Springer-Verlag, Berlin, New York, 1993, pp. 23–45.

[14] I. Lasiecka and R. Triggiani, *Differential and Algebraic Riccati Equations with Applications to Boundary/Point Control Problems: Continuous Theory and Approximation Theory*, Lecture Notes in Control and Information Sciences 164, Springer-Verlag, Berlin, 1991.

[15] K. A. Morris, *State feedback and estimation of well-posed systems*, Math. Control Signals Systems, 7 (1994), pp. 351–388.

[16] D. Salamon, *Infinite dimensional linear systems with unbounded control and observation: A functional analytic approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.

[17] D. Salamon, *Realization theory in Hilbert space*, Math. Systems Theory, 21 (1989), pp. 147–164.

[18] O. J. Staffans, *Coprime Factorizations and Optimal Control of Abstract Linear Systems*, Research Report A348, Helsinki University of Technology, Institute of Mathematics, April 1995.

[19] O. J. Staffans, *The Nonstandard Quadratic Cost Minimization Problem for Abstract Linear Systems*, Research Report A351, Helsinki University of Technology, Institute of Mathematics, May 1995.

[20] O. J. Staffans, *Quadratic optimal control of stable systems through spectral factorization*, Math. Control Signals Systems, 8 (1995), pp. 167–197.

[21] O. J. Staffans, *On the discrete and continuous time infinite-dimensional algebraic Riccati equations*, Systems Control Lett., 29 (1996), pp. 131–138.

[22] O. J. Staffans, *Quadratic Optimal Control through Coprime and Spectral Factorizations*, Reports on Computer Science & Mathematics, Series A 178, Åbo Akademi University, Åbo, Finland, 1996.

[23] O. J. Staffans, *Quadratic optimal control of stable well-posed linear systems*, Trans. Amer. Math. Soc., 349(1997), pp. 3679–3715.

[24] O. J. Staffans, *Coprime factorizations and well-posed linear systems*, SIAM J. Control Optim., 36(1997), pp. 1268–1292.

[25] O. J. Staffans, *Feedback representations of critical controls for well-posed linear systems*, International J. Robust Nonlinear Control, to appear.

[26] O. J. Staffans, *On the distributed stable full information H^∞ minimax problem*, International J. Robust Nonlinear Control, to appear.

[27] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.

[28] G. Weiss, *Transfer functions of regular linear systems. Part* I: *Characterizations of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.

[29] G. Weiss, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.

[30] G. Weiss and H. Zwart, *An example in linear quadratic optimal control*, Systems Control Lett., 33(1998), pp. 339–349.

[31] M. Weiss, *Riccati Equations in Hilbert Space*: *A Popov Function Approach*, Ph.D. thesis, Rijksuniversiteit Groningen, Groningen, the Netherlands, 1994.

[32] M. Weiss and G. Weiss, *The spectral factorization approach to the LQ problem for regular linear systems*, in Proc. of the Third European Control Conference, Rome, Italy, September 1995, Vol. 3, pp. 2247–2250.

[33] M. Weiss and G. Weiss, *Optimal control of stable weakly regular linear systems*, Math. Control Signals Systems, 10(1997), pp. 287–330.

[34] S. Q. Zhu, *On normalized Bezout fractions of distributed LTI systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 489–491.

[35] H. Zwart, *Linear quadratic optimal control for abstract linear systems*, in Modelling and Optimization of Distributed Parameter Systems, Chapman and Hall, New York, 1996, pp. 175–182.

SIAM J. Control Optim.
Vol. 37, No. 1, pp. 165–176

SUCCESSIVE APPROXIMATIONS OF LINEAR CONTROL MODELS[*]

JOHN R. BIRGE[†] AND SAMER TAKRITI[‡]

Abstract. We present a technique for approximating sequences of linear programs with varying right-hand sides and study the geometric properties of this approximation. Our approximation has an efficiency advantage over optimal solutions. When applied to deterministic control problems, the suggested technique outperforms the linear feedback model and provides accurate results (error of 5.8% in our numerical example). Numerical experience with stochastic models indicates that this approach may outperform the limited lookahead policies while maintaining low computational requirements.

Key words. stochastic programming, linear programming, optimal control, linear feedback control, certainty equivalent control, limited lookahead policies

AMS subject classifications. 90C15, 90C05, 93E20, 90C90

PII. S0363012995293864

1. Introduction.

In most stochastic optimal control applications, dynamic programming fails to solve the given model due to the lack of a closed form solution for the cost-to-go problem at each stage (see, for example, chapter 4 of [2]). These models are usually simplified by discretizing the continuous variables and applying the dynamic programming algorithm to the resulting finite-dimensional state and control spaces. Even then, a solution can be obtained only in the cases in which the dimension of the state-space is small. When an on-line solution is required, things become more complicated and one has to settle for an approximation that provides a suboptimal control for the given problem. Many of these approximations, such as certainty equivalent control and open-loop feedback control, require repeated solutions of an optimization model with varying parameters. When the parameters change quickly and the computational cost is large, the computed solution may become obsolete. In such situations, great emphasis is placed on reducing the calculation time. Unfortunately, this may reduce the quality of the solution obtained.

An example of the need for a quick solution is that of controlling the movements of a surface ship in restrained waters [6]. A manual control system places severe demands on the crew's skills, making the use of a more automated control highly desirable. A feedback control continuously measures the state of the vessel (location, velocity, and acceleration) and then makes a decision that maintains optimal or near-optimal performance. The performance of such systems is evaluated using a function of the state variables. In general, the goal is to minimize or maximize that function subject to linear system dynamics. Quantities to be minimized can be time, fuel, cost, etc., while those to be maximized include speed, efficiency, and profit.

With time discretized to simplify the calculations, the model becomes a discrete-time linear system [1, 5]. In this case, the state of the controlled system at stage $i+1$ is assumed to be a linear function of its state at stage i and of the decision made

[*] Received by the editors October 27, 1995; accepted for publication (in revised form) January 19, 1998; published electronically September 29, 1998.

http://www.siam.org/journals/sicon/37-1/29386.html

[†] Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109-2117 (jrbirge@umich.edu).

[‡] Mathematical Sciences Department, IBM Thomas J. Watson Research Center, P. O. Box 218, Yorktown Heights, NY 10598 (takriti@watson.ibm.com).

at that stage. If the performance measure is a linear function of the states and the controls, then a linear program is necessary to find an optimal control. In this linear case, the states can be described using the linear equations

$$x^{i+1} = \mathcal{A}x^i + \mathcal{B}u^i,\ i = 0,\ \ldots,\ T-1.$$

Here, $\mathcal{A}$ is the state-transition matrix, u^i is the control input vector of the system at stage i, and $\mathcal{B}u^i$ represents the contribution of the control variables to the change in the state vector. We compute an optimal control by minimizing a linear function of the states and the decisions over a horizon of T periods. That is,

$$\begin{array}{lrcl} \inf_{x,u} & \sum_{i=0}^{T} c_x^i x^i & + & \sum_{i=0}^{T-1} c_u^i u^i \\ \text{subject to} & x^0 & = & t, \\ & x^{i+1} & = & \mathcal{A}x^i + \mathcal{B}u^i,\ i = 0,\ \ldots,\ T-1, \\ & x & \geq & 0, \end{array} \tag{1}$$

where c_x and c_u are row vectors representing the cost associated with x and u. The vector t is the initial state of the system. One may also impose other linear constraints on the control vector u. Note that when the initial conditions vary, the only difference between the linear programs solved is in the first m' entries of the right-hand side, where m' is the dimension of t. Our intent is to develop fast methods that take advantage of this structure.

Using the right-hand side of (1), we partition the equality constraints into two sets. The first is the set of constraints for which the right-hand side vector varies. In (1), for example, this set is $x^0 = t$. The second is the set of constraints for which the right-hand side is known in advance. In our model, this set is $\mathcal{A}x^i + \mathcal{B}u^i - x^{i+1} = 0,\ i = 0,\ \ldots,\ T-1$. We denote the constraints of the first set by A, the constraints of the second set by B, the right-hand side vectors by t and b, and the objective function by c. Our discussion is general in that we do not make any assumptions regarding the structure of A and B.

To simplify notation, we denote the decision variables—state and control—by x. Therefore, the control problem is to evaluate the mathematical program

$$\begin{array}{rcl} \psi(t) = \inf_x c^T x & & \\ \text{subject to } Ax & = & t, \\ Bx & = & b, \\ x & \geq & 0 \end{array}$$

for different values of t. Generally speaking, the size of these programs is quite large. The time needed to compute an exact solution using classical linear programming techniques is lengthy. The solution obtained may become obsolete, since the state vector that was fed to the control algorithm is very different from the real state vector to which the control is applied. In practice, a linear relation is often assumed between the state vector and the control, so that one can make a decision quickly for any given initial state. This form of linear feedback control may, however, be suboptimal.

This paper presents a technique for quickly evaluating sequences of linear programs with varying right-hand sides. Our method combines features of linear feedback control with the complete optimization procedure to obtain an approach that is both fast and accurate. It is a generalization of the approach developed by Birge and Wets [3, 4]. While Birge and Wets deal with quickly evaluating a function $\phi(t) = \{\inf c^T x : Ax = t, x \geq 0\}$ for a varying $t \in \Re^m$, we allow our right-hand side vector to have a varying component, $t \in \Re^{m'}$, and a known one, $b \in \Re^{m-m'}$. Due to their assumption that all entries in the right-hand side can vary, Birge and Wets deal with a sublinear function. To approximate it, they evaluate $\phi(t)$ at m different points, $t^1, \dots, t^m$, and at $-t^1, \dots, -t^m$. These vectors are chosen so that the square matrix $D = [t^1, \dots, t^m]$ forms a basis for A. When a right-hand side, t, is provided, an upper bound on $\phi(t)$ can be computed as $\sum_{k=1}^m (D_{k.}^{-1}t)^+\phi(t^k) + (D_{k.}^{-1}t)^-\phi(-t^k)$, where $D_{k.}^{-1}$ denotes the kth row of the inverse of D. If $D_{k.}^{-1}t \geq 0$, then $(D_{k.}^{-1}t)^+ = D_{k.}^{-1}t$ and $(D_{k.}^{-1}t)^- = 0$. On the other hand, if $D_{k.}^{-1}t < 0$, then $(D_{k.}^{-1}t)^+ = 0$ and $(D_{k.}^{-1}t)^- = -D_{k.}^{-1}t$.

The rest of this paper is organized as follows. Section 2 describes the mathematical model, its geometric properties, and the polyhedral approximation. We discuss our approximation and related implementation issues in section 3 and then present some numerical results in section 4.

2. Geometric properties.

Mathematically, the linear control problem is equivalent to evaluating

$$\begin{array}{rrcl} \psi(t) = \inf_x c^T x & & & \\ \text{subject to } & Ax & = & t, \\ & Bx & = & b, \\ & x & \geq & 0, \end{array} \tag{2}$$

where $t \in \Re^{m'}$ and $b \in \Re^{m-m'}$ are the random (varying) and deterministic (fixed) parts of the right-hand side vector, respectively, and $A \in \Re^{m' \times n}$ and $B \in \Re^{m-m' \times n}$ are the constraint matrices. We make the assumptions that $[A^T|B^T]^T$ is of full row rank, b is in the nonnegative hull of the column vectors of B, and the set of the dual solutions of (2) is not empty.

PROPOSITION 1. *Under the above assumptions, the function $\psi(t)$ is proper and convex. Furthermore, the set of feasible right-hand sides, t, of the program* (2) *and the epigraph of $\psi(t)$ are convex polyhedra.*

Proof. To show that $\psi(t)$ is proper, consider any feasible solution, x^*, for the linear system $Bx = b$, $x \geq 0$, and the corresponding $t^* = Ax^*$. Then, $\psi(t^*) \leq c^T x^* < \infty$. The set of feasible right-hand sides is a convex polyhedron since it is the intersection of the convex polyhedral cone $\{[y^T, z^T]^T : y = Ax, z = Bx, x \geq 0\}$ (see, for example, [3] for a proof) and the hyperplanes $\{[y^T, z^T]^T : z_i = b_i\}$, $i = m' + 1, \dots, m$. That is, it is the intersection of a finite number of half-spaces. To prove that the set $\text{epi}(\psi)$ is a polyhedron, note that

$$\begin{aligned} \text{epi}(\psi) &= \{[v, t^T]^T : v \geq \psi(t)\} \\ &= \{[v, t^T]^T : v \geq [t^T, b^T][\pi^{k^T}, \mu^{k^T}]^T, k = 1, \dots, K\}, \end{aligned}$$

where $[\pi^{k^T}, \mu^{k^T}]^T$, $k = 1, \dots, K$, is the set of dual solutions corresponding to all dual feasible bases of (2). So, $\text{epi}(\psi)$ is the intersection of a finite number of closed half-spaces; hence, it is a convex polyhedron. □

Since $\text{epi}(\psi)$ is a convex polyhedron, it can be represented as a convex combination of its extreme points and a nonnegative combination of its extreme directions (see

Theorem 18.5 in [10]). Therefore,

$$\text{epi}(\psi) = \left\{ \begin{bmatrix} v \\ t \end{bmatrix} : \begin{bmatrix} v \\ t \end{bmatrix} = \sum_{k=1}^{k'} \mu_k \begin{bmatrix} v_k \\ t^k \end{bmatrix} + \sum_{k=k'+1}^{K} \mu_k \begin{bmatrix} v_k \\ t^k \end{bmatrix},\ \mu_k \geq 0 \right\},$$

where $\sum_{k=1}^{k'} \mu_k = 1$ and the set of vectors $[v_k, {t^k}^T]$ includes all extreme points and directions of epi(ψ).

Proposition 2. *Suppose* $[v_k, {t^k}^T]^T$, $k = 1, \ldots, k'$, *and* $[v_k, {t^k}^T]^T$, $k = k' + 1, \ldots, K$, *is a collection of points and directions in epi*(ψ) $\subset \Re^{m'+1}$. *Then*

$$(3) \qquad \psi(t) \leq \inf \left\{ v : \begin{bmatrix} v \\ t \end{bmatrix} = \sum_{k=1}^{k'} \mu_k \begin{bmatrix} v_k \\ t^k \end{bmatrix} + \sum_{k=k'+1}^{K} \mu_k \begin{bmatrix} v_k \\ t^k \end{bmatrix},\ \mu_k \geq 0 \right\},$$

where $\sum_{k=1}^{k'} \mu_k = 1$.

Proof. This follows directly from Proposition 1 and Jensen's inequality. □

Proposition 2 provides an upper bound on the value of $\psi(t)$ for any given right-hand side t. To implement this approximation, we solve (2) for different right-hand side vectors, t^k, $k = 1, \ldots, k'$, and directions, t^k, $k = k' + 1, \ldots, K$. Then, given a vector t, we write it in terms of t^k by choosing the appropriate values for μ_k. The approximate value of $\psi(t)$ is then $\sum_{k=1}^{K} \mu_k v_k$. Clearly, it is advantageous to select the smallest possible value for v_k for each vector used in the approximation; i.e., $v_k = \psi(t^k)$. We call this approximation polyhedral since it is based on the fact that epi(ψ) is a polyhedron. In the special case of $b = 0$, the function $\psi(t)$ is positively homogeneous and the polyhedral approximation becomes that of Birge and Wets, i.e., a sublinear approximation.

The following proposition states the necessary and sufficient conditions for the polyhedral approximation to be exact.

Proposition 3. *Inequality* (3) *is satisfied as an equality for all* $t \in \Re^{m'}$ *if and only if the set of vectors used in the approximation includes all the extreme points and directions of epi*(ψ).

Proof. If the set of vectors used in the polyhedral approximation includes all extreme points and directions, then

$$\left\{ \sum_{k=1}^{k'} \mu_k \begin{bmatrix} v_k \\ t^k \end{bmatrix} + \sum_{k=k'+1}^{K} \mu_k \begin{bmatrix} v_k \\ t^k \end{bmatrix},\ \sum_{k=1}^{k'} \mu_k = 1,\ \mu_k \geq 0 \right\} = \text{epi}(\psi)$$

by Theorem 18.5 of [10], and the bound is exact. On the other hand, assume that the bound is exact and that an extreme point or direction is not included in the set of approximating vectors. Then, we have a representation of an extreme point or direction in terms of other points and directions in epi(ψ), which leads to a contradiction. □

If the representation, μ, of the given right-hand side, t, is not unique, we can choose the vector μ that provides the lowest upper bound. However, finding the best μ requires solving another linear program of size $m' \times K$. To avoid this difficulty, the set of the representing vectors can be chosen so that the approximation has a unique solution which satisfies the constraint $\sum_{k=1}^{k'} \mu_k = 1$. That is, the vectors t^k form a basis for $\Re^{m'+1}$. The solution vector μ is then found by solving the system of linear equations:

$$\begin{bmatrix} D_p & D_d \\ e^T & 0 \end{bmatrix} \begin{bmatrix} \mu_p \\ \mu_d \end{bmatrix} = \begin{bmatrix} t \\ 1 \end{bmatrix},$$

where $D_p = [t^1, \ldots, t^{k'}]$ and $D_d = [t^{k'+1}, \ldots, t^{m'+1}]$. The objective function value obtained from using the vectors $[D_p, D_d]$ in the approximation is denoted by $\psi_{[D_p,D_d]}$. The problem here is that some components of μ may be negative, in which case the upper bound is assumed to be infinite. One can use a number of bases, D_p^l and D_d^l, $l = 1, \ldots, L$, and choose the solution corresponding to the best bound. Also, using the convex hull of D_p^l plus the nonnegative combination of D_d^l improves the bound obtained.

PROPOSITION 4. *Let D_p^l and D_d^l, $l = 1, \ldots, L$, be a collection of matrices as described above. Then*

$$\psi(t) \leq \psi_{con(D_p^l)+pos(D_d^l)}(t) \leq \inf_l \psi_{[D_p^l,D_d^l]},$$

where

$$\psi_{con(D_p^l)+pos(D_d^l)}(t) = \inf \left\{ v : \begin{bmatrix} v \\ t \end{bmatrix} = \sum_{l=1}^{L} \sum_{i=1}^{k_l'} \mu_i^l \begin{bmatrix} v_i^l \\ t_i^l \end{bmatrix} + \sum_{l=1}^{L} \sum_{i=k_l'+1}^{m'+1} \mu_i^l \begin{bmatrix} v_i^l \\ t_i^l \end{bmatrix} \right\},$$

$\sum_{l=1}^{L} \sum_{i=1}^{k_l'} \mu_i^l = 1$, *and* $\mu_i^l \geq 0$, $i = 1, \ldots, m'+1$, $l = 1, \ldots, L$.

Proof. Since $\text{epi}(\psi)$ is a convex polyhedron, any convex combination of vectors in $\text{epi}(\psi)$, such as the column vectors of D_p^l, $l = 1, \ldots, L$, plus a nonnegative combination of directions in $\text{epi}(\psi)$, such as the column vectors of D_d^l, $l = 1, \ldots, L$, is also in $\text{epi}(\psi)$. That is,

$$\text{epi}(\psi_{\text{con}(D_p^l)+\text{pos}(D_d^l)}) \subseteq \text{epi}(\psi).$$

Also, any feasible solution, μ_i^l, for $\text{epi}(\psi_{[D_p^l,D_d^l]})$ can be used to construct a feasible solution for $\text{epi}(\psi_{\text{con}(D_p^l)+\text{pos}(D_d^l)})$. Therefore,

$$\text{epi}(\psi_{[D_p^l,D_d^l]}) \subseteq \text{epi}(\psi_{\text{con}(D_p^l)+\text{pos}(D_d^l)}).$$

The statement of the proposition follows by taking the infimum of the components corresponding to the objective function value, v, over the three polyhedral sets. □

In order to obtain a good approximation, it is clear that one may need to use a large number of bases. Of course, the goal is to have a small number of bases that cover the range of possible t and that provide reasonably accurate results. As an example, consider the following system of linear equations:

$$\begin{array}{cccccl} x_1 & & +x_3 & -x_4 & & = t_1, \\ & x_2 & +x_3 & & -x_5 & = t_2, \\ x_1, & x_2, & x_3, & x_4, & x_5 & \geq 0 \end{array}$$

with the objective function $x_1 + x_2 + x_3 + 100x_4 + 100x_5$. Let us assume that t can only take its values from the first quadrant. If we choose

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

as an approximating basis, our approximation is exact for all right-hand side vectors such that $t_1 \geq t_2$, but we pay a very large price for vectors of the form $t_1 < t_2$. Also, using the basis

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

results in very large objective function values for vectors that have $t_1 > t_2$. To obtain an exact approximation for all $t \geq 0$, we need to include both of the previous bases in our set of approximating bases. On the other hand, if we use the identity matrix as an approximating basis, we obtain a reasonable approximation for all possible vectors, t, while performing a smaller number of calculations. In other words, one basis that is suboptimal but covers a wide region is better than choosing an optimal basis that may end up very bad if conditions change. To avoid searching for good bases, we suggest using parametric programming in conjunction with our approximation to achieve reasonable bounds while using a single basis. The following section discusses this approach.

3. Using parametric programming. An alternative for the approximation of section 2 is to choose a set of vectors, t^k, and use scalar multiples of these vectors to represent a given right-hand side. In other words, we use the vectors $\lambda_k t^k$, $\lambda_k \in \Re$, to represent any right-hand side, t. Note that the function $\psi(\lambda_k t^k)$ is a piecewise-linear convex function of the parameter λ_k (see, for example, section 8.11 in [9]). The best upper bound is then obtained by solving the following program:

$$\begin{array}{llrcl} & \inf_{\lambda,\mu} & \displaystyle\sum_{k=1}^{K}\mu_k\psi(\lambda_k t^k) & & \\ (4) & \text{subject to} & \displaystyle\sum_{k=1}^{K}\mu_k\lambda_k t^k & = & t, \\ & & \displaystyle\sum_{k=1}^{K}\mu_k & = & 1, \\ & & \mu & \geq & 0. \end{array}$$

Let $\nu_k = \mu_k\lambda_k$. Then, the mathematical program (4) can be written in the form

$$\begin{array}{llrcl} & \inf_{\nu,\mu} & \displaystyle\sum_{k=1}^{K}\mu_k\psi(\frac{\nu_k}{\mu_k} t^k) & & \\ (5) & \text{subject to} & \displaystyle\sum_{k=1}^{K}\nu_k t^k & = & t, \\ & & \displaystyle\sum_{k=1}^{K}\mu_k & = & 1, \\ & & \mu & \geq & 0. \end{array}$$

Note that $\mu_k\psi(\frac{\nu_k}{\mu_k}t^k)$ is convex (see, for example, section 5 in [10]), hence the objective function of (5) is convex. However, the resulting approximation is not linear, and solving it may require more effort than solving the original program (2).

To simplify the calculations, the set of vectors t^k is chosen to form a basis for $\Re^{m'}$. In this case, we can write program (5) as

$$
\begin{array}{lll}
\inf_{\lambda,\mu} & \displaystyle\sum_{k=1}^{m'} \mu_k \psi(\lambda_k t^k) & \\
\text{subject to} & \mu_k \lambda_k = D_{k.}^{-1} t, & \\
 & \displaystyle\sum_{k=1}^{K} \mu_k = 1, & \\
 & \mu \geq 0, &
\end{array}
\tag{6}
$$

where D is the basis formed by taking t^k as its column vectors and $D_{k.}^{-1}$ is the kth row of D^{-1}. Note that in the previous program, the value of each μ_k determines the corresponding λ_k. To look for an approximate solution of (6), one can assume certain values for μ_k, compute the corresponding objective function values, and then take the lowest upper bound as an approximation for the objective function value of (2). The previous operation is not expensive if m' is relatively small.

Since the parametric linear programs, $\psi(\lambda_k t^k)$, $k = 1, \ldots, K$, are solved for all values of λ_k or, at least, for a wide range of positive and negative values, program (6) has an infinite number of feasible solutions for any given vector, t. In this case, a single basis is sufficient to cover the set of all possible right-hand side vectors, $t \in \Re^{m'}$. Here again, we have a generalization of the positive linear bases approach presented by Birge and Wets. They suggest solving the original linear program for each t^k and $-t^k$ to guarantee a representation of any vector t. In our case, we solve parametrically for each t^k and for all possible negative and positive values of the parameter λ_k. Here is an algorithmic description of our approximating procedure.

ALGORITHM. Given the parameters $\lambda_{\min} \leq 0 \leq \lambda_{\max}$, a step size $\Delta > 0$, and a basis, $D = [t^1, \ldots, t^{m'}]$, for A, we use the following procedure to approximate $\psi(t)$.

```
• Initialization. Evaluate $\psi(\lambda t^k)$, $k = 1, \ldots, m'$. The evaluation is performed
  for all values of $\lambda$ in the range $[\lambda_{\min}, \lambda_{\max}]$. Since $\psi$ is a piecewise-linear
  function of the parameter $\lambda$, we only need to evaluate $\psi$ at the break-points.
  Set $v \leftarrow \infty$.
• Compute the inverse $D^{-1}$.
• General step.
  1. Obtain a new right-hand side vector $t$.
  2. Set $\mu_1 = \cdots = \mu_{m'-1} = 0$.
     While $\mu_1 \leq 1$
         While $\mu_2 \leq 1 - \mu_1$
             ...
                 While $\mu_{m'-1} \leq 1 - \sum_{k=1}^{m'-2} \mu_k$
                     Set $\mu_{m'} = 1 - \sum_{k=1}^{m'-1} \mu_k$.
                     For $k = 1$ to $m'$
                         If $\mu_k = 0$, let $\lambda_k = 0$; otherwise, let $\lambda_k = D_{k.}^{-1} t / \mu_k$.
                     If $v > \sum_{k=1}^{m'} \mu_k \psi(\lambda_k t^k)$, then set $v = \sum_{k=1}^{m'} \mu_k \psi(\lambda_k t^k)$.
                 $\mu_{m'-1} \leftarrow \mu_{m'-1} + \Delta$
             ...
         $\mu_2 \leftarrow \mu_2 + \Delta$
     $\mu_1 \leftarrow \mu_1 + \Delta$
  3. The value of $v$ is the best upper bound for $\psi(t)$ using $D$.
```

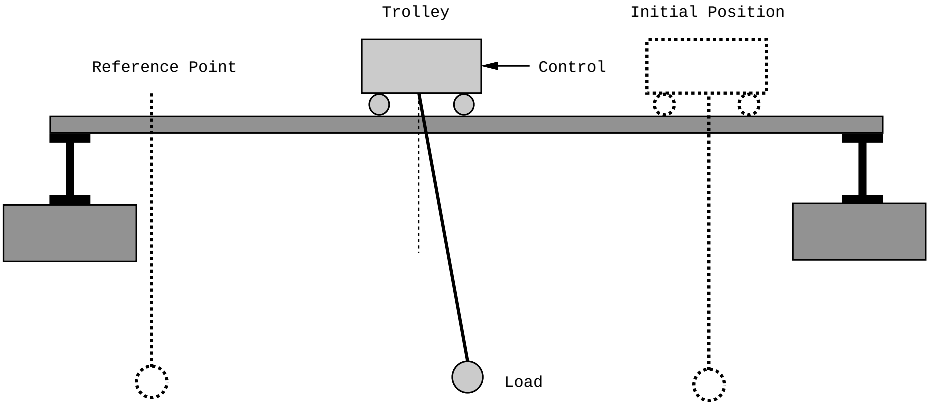


Fig. 1. *A crane control problem.*

4. Go to 1.

In the following section, we test our algorithm by applying it on a crane control problem.

4. Numerical results. To evaluate the procedure of section 3, we consider the feedback control example of [7, 8]. The example, which is depicted in figure 1, is based on a crane with an overhead trolley that runs on frictionless straight rails with its load suspended on an inextensible cable from its center of gravity. The objective is to move the trolley from a given position to a reference point with the load starting and finishing in a stationary position directly beneath the trolley. The state variables, x_1^i, x_2^i, x_3^i, and x_4^i, represent the position of the trolley (from the reference point), its velocity, the angle of the cable, and its angular velocity. The state of the system changes according to the linear equation

$$x^{i+1} = \begin{bmatrix} 1 & 0.3 & 0.036 & 0.0036 \\ & 1 & 0.234 & 0.036 \\ & & 0.92 & 0.292 \\ & & -0.526 & 0.92 \end{bmatrix} x^i + \begin{bmatrix} 0.08 \\ 0.534 \\ -0.08 \\ -0.526 \end{bmatrix} u^i + \alpha\beta w^i, \tag{7}$$

where α and β are scalars and w^i is the random disturbance at stage i. The control variable, u_i, is restricted to the range $[-1, 1]$. Starting at $[-5, 1, 0, 0]^T$ and using a study horizon of $T = 30$, the goal is to be as close as possible to the origin at any point in time. This requirement is approximated by minimizing the expectation over w^i of the function $\sum_{i=1}^{T} \sum_{j=1}^{m'} |x_j^i|$.

Since the size of such a problem may make on-line control practically infeasible, certainty equivalent controllers (CECs) are used to simplify the calculations. CECs apply, at each stage, the control that would be optimal if all uncertain quantities are fixed at their expected values (see, for example, section 4.1 of [2]). That is, given the initial state, x^0, we assume known conditions throughout the study horizon. The state transition equation takes the form $x^{i+1} = \mathcal{A}x^i + \mathcal{B}u^i + \alpha\beta E(w^i),\ i = 0,\ \ldots,\ T-1$, where $E(w^i)$ is the expected value of the noise term at stage i. At each time period, we apply the decision, u^0, obtained from solving the mathematical program

$$\begin{array}{llrl} (8) & \min_{x,u} & \displaystyle\sum_{i=1}^{T}\sum_{j=1}^{m'}|x_j^i| & \\ & \text{subject to} & x^{i+1} = & \mathcal{A}x^i + \mathcal{B}u^i + \alpha\beta E(w^i),\ i = 0,\ \ldots,\ T-1. \end{array}$$

If the resulting problem (8) is still large for an on-line controller, CECs are often combined with limited lookahead policies (see section 4.3 of [2]) to further reduce the size of the problem. The resulting mathematical program is to minimize the cost over the truncated horizon, $\mathcal{T}$,

$$\begin{array}{llrl} (9) & \min_{x,u} & \displaystyle\sum_{i=1}^{\mathcal{T}}\sum_{j=1}^{m'}|x_j^i| & \\ & \text{subject to} & x^{i+1} = & \mathcal{A}x^i + \mathcal{B}u^i + \alpha\beta E(w^i),\ i = 0,\ \ldots,\ \mathcal{T}-1. \end{array}$$

This approach results in significant computational savings at the cost of providing a suboptimal control. In general, the quality of the decision or control tends to improve as the value of $\mathcal{T}$ increases.

Our technique is basically an approximation for (8). The optimization problem (8) is solved parametrically for different right-hand side vectors, t^K, which are used to approximate the optimal value of (8) whenever a new state, x^0, is observed. At each stage, a suboptimal policy is constructed using the optimal policies

$$x^i = \sum_{k=1}^{K} \mu_k X_k^i(\lambda_k), \tag{10}$$

where $X_k^i(\lambda_k)$ is an optimal solution vector for (8) when the initial state of the system is $\lambda_k t^k$. The vectors λ and μ are obtained from solving (6). The error in our approximation is introduced in representing t as a combination of the vectors t^k. Since the effort needed to compute the upper bound of the approximation (6) is independent of the effort needed to approximate $\psi(\lambda_k t^k)$, it is best, at least in the deterministic case, to use the most accurate approximation for $\psi(\lambda_k t^k)$.

Here, we compare our approximation to the limited lookahead policies with horizons, $\mathcal{T}$, between 3 and 30 periods. In order to obtain a quick solution, we use the identity matrix as the approximating basis for (6). That is, the problem in (8) is solved parametrically for the right-hand side vectors $\lambda_k e_i^k$, $i = 1, \ldots, 4$, where λ_k is chosen to be between $\lambda_{\min} = -50$ and $\lambda_{\max} = 50$. The values of μ_i, $i = 1, \ldots, 4$, are varied between 0 and 1 using a step size of $\Delta = 0.2$.

As a third case, we consider a quadratic objective function subject to the constraints of (8) and assume that $T \to \infty$ so that we can derive a stationary linear feedback control. In this case, we assume that the objective function at each stage is to minimize the l_2 norm of the state vector. The resulting quadratic program has a stationary solution in which the optimal control, u^i, at any stage is a linear function of the state of the system

$$u^i = [-(\mathcal{B}^T\mathcal{K}\mathcal{B})^{-1}\mathcal{B}^T\mathcal{K}\mathcal{A}]x^i.$$

Here, $\mathcal{K}$ is the unique solution of the discrete-time Riccati equation

$$\mathcal{K} = \mathcal{A}^T[\mathcal{K} - \mathcal{K}\mathcal{B}(\mathcal{B}^T\mathcal{K}\mathcal{B})^{-1}\mathcal{B}^T\mathcal{K}]\mathcal{A} + I,$$

where I is the identity matrix. For our example, the solution of the Riccati equation is

$$\mathcal{K} = \begin{bmatrix} 7.376 & 5.387 & 0.979 & 4.635 \\ 5.387 & 12.562 & -2.609 & 10.514 \\ 0.979 & -2.609 & 9.347 & -1.531 \\ 4.635 & 10.514 & -1.531 & 10.729 \end{bmatrix},$$

and the resulting control is of the form

$$u^i = [-1.052, -2.329, 1.042, -0.143]x^i.$$

This control law is a rough approximation of an optimal policy for the linear model.

Table 1 compares the policy obtained by applying the parametric programming model of section 3 with both the lookahead policy and the stationary linear feedback policy. The first column in the table, $\mathcal{T}$, contains the number of stages used in the lookahead policy approximation of (9). The second column presents the objective function value of (9) over the full horizon of 50 periods for the deterministic case, $\mu = 0.0$. In general, the accuracy of the lookahead policies improves as the number of stages, $\mathcal{T}$, used in the approximation increases. The row labeled "Quadratic" provides the objective function value for the stationary linear feedback approach, while the row labeled "Polyhedral" provides the objective value if the approximation of section 3 is used. The approximation of section 3 provides a bound of 74.01, compared to a bound of 81.10 for the stationary linear feedback control. The optimal objective value, 69.94, corresponds to the limited lookahead policies when $\mathcal{T} = 30$. Using a lookahead approximation with a horizon smaller than 12 periods yields an inferior solution to the polyhedral approximation. We expect our approximation to improve as more bases are included. The CPU time required to compute the polyhedral approximation is almost the same as that required to solve the linear program (9) with $\mathcal{T} = 3$. The time needed to solve (9) increases as the number of stages considered is increased. The quadratic linear feedback model is 10 times faster than the polyhedral approximation since it only involves matrix multiplication. However, the solution produced by the linear feedback policy is 10.1% worse than that of the polyhedral approximation.

In order to incorporate the effect of random noise into the problem, we assume that the state vector, x^{i+1}, at time $i+1$ is given as a linear function of the state of the system at time i, plus a random noise vector $\alpha\beta w^i$. Each component of w^i is a uniformly distributed random variable over the open interval (–0.5, 0.5). Solving this problem, with an expected value objective as in (7), yields a stochastic program that may be quite difficult to solve in real time. The CEC combined with limited lookaheads is one approach to achieve an approximate solution. In this case, we resolve (9) with the current state, x^i, and expectations for future conditions. Our approximation applies to this problem as well, but here we may benefit by a more robust basis than in limited lookaheads. The linear feedback approximation may also benefit in this case.

We use three values for the coefficient α: 0.01, 0.05, and 0.10. For the limited lookahead strategy of (9), we assume that the value of the coefficient β is a function of the study horizon, $\mathcal{T}$. It grows linearly with the size of the study horizon since the times needed to solve the linear programs involved increase with $\mathcal{T}$. The effect of the number of stages is assumed to be $\beta = \mathcal{T}/3$. The reason for this assumption is that the execution time of the limited lookahead model (9) increases as $\mathcal{T}$ increases, which results in a wider range for the random disturbance. On the other hand, the polyhedral

Table 1
Comparison of different approximations.

$\mathcal{T}$	$\alpha = 0.0$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
3	98.69	100.68	102.50	110.25
4	98.58	97.41	96.46	172.96
5	98.24	104.13	123.53	279.60
6	98.22	103.79	120.90	331.24
7	89.42	94.34	132.64	392.83
8	84.38	88.70	158.41	473.78
9	79.95	85.28	166.99	600.29
10	75.64	86.88	178.43	750.42
11	74.45	90.00	205.48	1017.94
12	74.73	93.46	235.46	1167.03
13	73.64	95.08	273.56	1369.77
14	73.59	98.30	322.20	1568.85
15	73.27	100.51	364.31	1807.80
16	73.06	103.95	439.03	2045.91
17	71.36	104.98	499.91	2273.62
18	71.64	103.57	562.90	2506.10
19	71.60	107.28	633.87	2717.86
20	70.06	108.83	737.30	2898.79
25	69.94	120.41	1209.84	4072.42
30	69.94	131.87	1819.55	5248.97
Quadratic	81.10	81.24	82.01	83.34
Polyhedral	74.01	75.20	85.44	100.26

and quadratic approximation times do not increase with the horizon. Therefore, we can consider β fixed for these approximations. Since the polyhedral approximation requires time comparable to a three stage ($\mathcal{T} = 3$) horizon, we use $\beta = 1$ for that method. The linear feedback control from the quadratic model is 10 times faster than our serial polyhedral implementation, so we choose $\beta = 0.1$ in that case.

The results are shown in columns 3, 4, and 5 of Table 1. Note that the best bound provided by the limited lookahead strategy is always higher than that produced by the polyhedral approximation. This is a result of increasing β with the size of the problem that needs to be solved when limited lookahead is used. Our approximation seems to produce reasonable results even when α is set to 0.10. Note that the linear feedback control produces better results for $\alpha = 0.05$ and $\alpha = 0.10$ since the value of β used is relatively small. However, we expect our results to improve if we use parallel computers to cut down the calculation time.

5. Conclusions. We gave an approximate solution method for linear control problems with varying initial state conditions that require rapid solutions. We showed characteristics of the approximation that guarantee bounds on the solution of the linear optimization problem. As shown in an example, our approximation has an efficiency advantage over an optimal policy for deterministic problems with quite accurate results for longer time horizons (error of 5.8%). The corresponding linear feedback controller had an error of 16.0% for the same problem. In addition, our policy was effective in a stochastic system in reducing the error by more than 14.3% over the limited lookahead policies while maintaining low computational requirements.

Acknowledgments. The authors would like to thank the associate editor and the referees for their constructive comments, especially for pointing out that (5) is convex.

REFERENCES

[1] K. J. Aström, *Introduction to Stochastic Control Theory*, Academic Press, New York, 1970.

[2] D. P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[3] J. R. Birge and R. J.-B. Wets, *On-Line Solution of Linear Programming Using Sublinear Functions*, Tech. report 86-25, University of Michigan, Ann Arbor, MI, 1986.

[4] J. R. Birge and R. J.-B. Wets, *Sublinear upper bounds for stochastic programs with recourse*, Math. Programming, 43 (1989), pp. 131–149.

[5] C. K. Chui and G. Chen, *Linear Systems and Optimal Control*, Springer-Verlag, Berlin, 1989.

[6] H. T. Cuong, *Investigation of Methods for Adaptive Control of Surface Ships*, Ph.D. thesis, University of Michigan, Ann Arbor, MI, 1980.

[7] W. G. Hwang and W. E. Schmitendorf, *Controllability results for systems with a nonconvex target*, IEEE Trans. Automat. Control, AC-31 (1984), pp. 794–802.

[8] S. S. Keerthi and E. G. Gilbert, *Optimal infinite-horizon feedback laws for a general class of constrained discrete-time systems: Stability and moving horizon approximations*, J. Optim. Theory Appl., 57 (1988), pp. 265–294.

[9] K. G. Murty, *Linear Programming*, John Wiley and Sons, New York, 1983.

[10] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

SIAM J. Control Optim.
Vol. 37, No. 1, pp. 177–210

A NEW ALGORITHM FOR STATE-CONSTRAINED SEPARATED CONTINUOUS LINEAR PROGRAMS[*]

XIAODONG LUO[†] AND DIMITRIS BERTSIMAS[‡]

Abstract. During the last few decades, significant progress has been made in solving large-scale finite-dimensional and semi-infinite linear programming problems. In contrast, little progress has been made in solving linear programs in infinite-dimensional spaces despite their importance as models in manufacturing and communication systems. Inspired by the research on separated continuous linear programs, we propose a new class of continuous linear programming problems that has a variety of important applications in communications, manufacturing, and urban traffic control. This class of continuous linear programs contains the separated continuous linear programs as a subclass. Using ideas from quadratic programming, we propose an efficient algorithm for solving large-scale problems in this new class under mild assumptions on the form of the problem data. We prove algorithmically the absence of a duality gap for this class of problems without any boundedness assumptions on the solution set. We show this class of problems admits piecewise constant optimal control when the optimal solution exists. We give conditions for the existence of an optimal solution. We also report computational results which illustrate that the new algorithm is effective in solving large-scale realistic problems (with several hundred continuous variables) arising in manufacturing systems.

Key words. continuous linear programming, strong duality, semi-infinite linear programming, nonlinear programming, jobshop scheduling problems

AMS subject classifications. 49J15, 49K30, 49M35, 49M37, 49M39, 90C34, 90C35

PII. S0363012995292664

1. Introduction.

Bellman [7, 8] introduced the following optimization problem:

$$
\begin{aligned}
(CLP)\ \text{minimize} & \int_0^T c(t)'x(t)\ dt \\
\text{subject to (s.t.)}\ \ & A(t)x(t) + \int_0^t B(s,t)x(s)\ ds \le b(t), \\
& x(t) \ge 0, \qquad t \in [0,\ T],
\end{aligned}
$$

where $A(t)$ and $B(s,t)$ are matrices depending on time (their entries are bounded measurable functions) and $b(t)$ and $c(t)$ are bounded measurable functions. (CLP) is an instance of a *continuous linear program.*

The problem that has attracted the most attention is the *separated continuous linear programming* problem (SCLP), a subclass of the continuous linear programming problem:

$$
(SCLP)\ \text{minimize} \int_0^T c(t)'u(t)\ dt
$$

[*] Received by the editors October 4, 1995; accepted for publication (in revised form) February 9, 1998; published electronically October 7, 1998.
http://www.siam.org/journals/sicon/37-1/29266.html

[†] Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139 (xluo@ athena.mit.edu).

[‡] Sloan School of Management, Massachusetts Institute of Technology, 50 Memorial Dr., Room E53-363, Cambridge, MA 02139 (dbertsim@mit.edu). The research of the author was partially supported by Presidential Young Investigator Award DDM-9158118 with matching funds from Draper Laboratory and NSF grant DMI-9610486.

$$\text{s.t.} \int_0^t Gu(t)\ dt + y(t) = a(t), \tag{1}$$

$$\begin{aligned} &Hu(t) \leq b(t), \\ &y(t),\ u(t) \geq 0, \qquad t \in [0,\ T], \end{aligned}$$

where $y(t)$ and $a(t)$ are absolutely continuous functions. Note that the variables $u(t)$ and $y(t)$ are linked only through (1), in which $u(t)$ appears only under the integration operator and $y(t)$ does not appear under the integration operator. The problem $(SCLP)$ was first introduced by Anderson [4] in order to model job-shop scheduling problems (see also Avram, Bertsimas, and Ricard [6], Weiss [49]).

In this paper, we examine a larger subclass of continuous linear programs which can be used to model a variety of problems that arise in communications, manufacturing, and urban traffic control (see Luo [32]). The problem we consider is the following:

$$(SCSCLP)\ \text{minimize} \int_0^T (c(t)'u(t) + g(t)'y(t))\ dt$$

$$\text{s.t.} \quad \int_0^t Gu(t)\ dt + Ey(t) = a(t), \tag{2}$$

$$Hu(t) \leq b(t), \tag{3}$$

$$Fy(t) \leq h(t), \tag{4}$$

$$u(t) \geq 0, \qquad t \in [0,\ T],$$

where $b(t)$, $c(t)$, $g(t)$, and $h(t)$ are bounded measurable functions and $a(t)$ is an absolutely continuous function. The dimensions of $b(t)$, $a(t)$, $u(t)$, $y(t)$, and $h(t)$ are n_1, n_2, n_3, n_4, and n_5, respectively. We call $(SCSCLP)$ the *state-constrained separated continuous linear programs.* We call $y(t)$ the state variable and $u(t)$ the control variable. We call (2) the state equation (or sometimes we use the term system dynamics) and call (4) the state constraint. We call (3) the control constraint.

Related literature. The computational study of CLP was initiated by Lehman [26] who attempted to develop a simplex-like algorithm for CLP. Drews [10], Hartberger [20], and Segers [44] later followed him. Perold [37, 38] developed the first simplex-like algorithm for CLP (see also Anderson, Nash, and Perold [1] and Anderson and Philpott [3]. Anstreicher [5] continued Perold's work in his Ph.D. thesis, even though both their algorithms were still incomplete. In the meantime, Russian authors such as Ilyutovich [21, 22] treated the problem using Pontryagin's maximum principle. In addition, Ito, Kelley, and Sachs [23] have developed a primal-dual path, following interior point method for CLP. Anderson and Nash in [2] proposed a convex quadratic programming procedure for $(SCLP)$. The series of papers on SCLP by Pullan [41, 42, 43] deals with solution structure, duality theory, and numerical algorithms and to the best of our knowledge represents the state of the art of this area. Philpott and Craddock [39] later specialized Pullan's work to a network version of SCLP and presented encouraging numerical results.

Objective and contributions. In this paper, we will develop a new algorithm for solving SCSCLP problems under Assumption 1 below. The new algorithm uses discretization. Unlike the algorithms mentioned above, it varies the discretization and control simultaneously. Based on the number of constant pieces allowed in the control, we develop a quadratic program with polyhedral constraints. The quadratic program is generally nonconvex. However, we do not need to solve the quadratic program to optimality. We only need to obtain a KKT point. We use the Frank–Wolfe method (see Martos [33] and Murty [36]) or general matrix-splitting algorithms (see Lin and Pang [28], Eckstein [11], Bertsekas and Tsitsiklis [9], Luo and Tseng [31]) to find a KKT point for the quadratic program. By gradually increasing (and occasionally decreasing) the number of pieces allowed in the control, we can improve upon any nonoptimal KKT solution. We call this the successive quadratic programming method. By a KKT solution structural result of Luo and Tseng [31], we show that the iterates of the algorithm move from one polyhedral set to another, with improved cost. By bounding the size of the quadratic programs we encounter, we bound the number of all such polyhedral sets. We show that the new algorithm converges in finite time. The absence of a duality gap and the existence of certain highly structured optimal solutions for ($SCSCLP$) follow as byproducts. Furthermore, we have implemented our algorithm and report computational results which illustrate that the new algorithm is effective in solving large scale realistic problems (with several hundred continuous variables) arising in manufacturing systems.

Structure of the paper. The remainder of this paper is structured as follows. In section 2, we introduce the dual problem for ($SCSCLP$) (called ($SCSCLP^*$)) and state our assumptions. We also prove weak duality results between ($SCSCLP$) and ($SCSCLP^*$) and introduce some standard definitions and notations. In section 3, we develop a quadratic program with polyhedral constraints. In section 3.1, we review some nonlinear programming techniques for calculating a KKT point of a quadratic program with polyhedral constraints. In section 4, we develop a procedure for removing redundant intervals in a feasible solution for ($SCSCLP$). In section 5, we introduce a new discrete approximation for ($SCSCLP$) which is closely related to the dual problem. From this discrete approximation, we derive a criterion to detect whether a feasible solution is optimal for ($SCSCLP$). If the criterion is not satisfied, we derive a descent direction for the feasible solution to ($SCSCLP$). In section 6, we formally state the new algorithm. In section 7, we prove that the new algorithm converges in finite time. In section 8, we use the new algorithm to prove new duality results and new optimal solution structural results for ($SCSCLP$). In section 9, we report computational results that illustrate the effectiveness of the new algorithm in solving large-scale problems. The reader is advised to first read sections 2 and 6 to obtain a general idea of the problem, the assumptions, and the new algorithm.

2. Definitions and notation. First, we reiterate problem ($SCSCLP$) and state our assumptions. We consider the problem

$$
\begin{aligned}
(SCSCLP)\ \text{minimize} & \int_0^T (c(t)'u(t) + g(t)'y(t))\ dt \\
\text{s.t.} \qquad & \int_0^t Gu(t)\ dt + Ey(t) = a(t), \\
& Hu(t) \le b(t), \\
& Fy(t) \le h(t), \\
& u(t) \ge 0, \qquad t \in [0,\ T],
\end{aligned}
$$

and its dual

$$
(SCSCLP^*)\ \text{maximize} - \int_0^T a(t)'\ d\pi(t) - \int_0^T b(t)'\eta(t)\ dt - \int_0^T h(t)'\ d\xi(t)
$$

$$\begin{aligned} \text{s.t.} \qquad & c(t) - G'\pi(t)\ dt + H'\eta(t) \geq 0, \\ & E'\pi(t) + F'\xi(t) = \int_t^T g(t)\ dt, \\ & \pi(t) \text{ bounded measurable with finite variation,} \\ & \xi(t) \text{ monotonic increasing and right continuous} \\ & \qquad \text{on } [0,\ T] \text{ with } \xi(T) = 0, \quad \pi(T) = 0, \\ & \eta(t) \geq 0, \qquad \text{for } t \in [0,\ T], \end{aligned}$$

under the following assumptions (we will give a formal definition of piecewise linear (constant) functions later in this section).

Assumption 1.

a) *$a(t)$ and $h(t)$ are continuous,*

b) *$a(t)$, $c(t)$, and $h(t)$ are piecewise linear,*

c) *$b(t)$ and $g(t)$ are piecewise constant,*

d) *Problem ($SCSCLP$) is feasible and its objective value is bounded from below.*

We require that $u(t)$, $y(t)$, and $\eta(t)$ are bounded and measurable functions on $[0,\ T]$. We remark that the dual problem ($SCSCLP^*$) reduces to the alternative dual problem for ($SCLP$) introduced by Pullan [41] when the primal problem is an SCLP.

We have the following weak duality results for ($SCSCLP$). For completeness, we give its proof.

Proposition 1. *Weak duality holds between ($SCSCLP$) and ($SCSCLP^*$).*

Proof. Consider any two solutions $(u(t),\ y(t))$ and $(\pi(t),\ \eta(t),\ \xi(t))$ which are feasible to ($SCSCLP$) and ($SCSCLP^*$), respectively. Let $z(t) = b(t) - Hu(t)$ and $\bar{z}(t) = h(t) - Fy(t)$. We have

$$\begin{aligned} & \int_0^T (c(t)'u(t) + g(t)'y(t))\ dt - \left(- \int_0^T a(t)'\ d\pi(t) - \int_0^T b(t)'\eta(t)\ dt - \int_0^T h(t)'\ d\xi(t) \right) \\ & = \int_0^T (c(t)'u(t) + g(t)'y(t))\ dt + \int_0^T a(t)'\ d\pi(t) + \int_0^T b(t)'\eta(t)\ dt + \int_0^T h(t)'\ d\xi(t) \\ & = \int_0^T (c(t)'u(t) + g(t)'y(t))\ dt + \int_0^T \left(\int_0^t Gu(s)\ ds + Ey(t) \right)'\ d\pi(t) \\ & \quad + \int_0^T (Hu(t) + z(t))'\eta(t)\ dt + \int_0^T (Fy(t) + \bar{z}(t))'d\xi(t) \\ & = \int_0^T (c(t)'u(t) + g(t)'y(t))\ dt - \int_0^T \pi(t)'Gu(t)\ dt + \int_0^T (Ey(t))'\ d\pi(t) \\ & \quad + \int_0^T (Hu(t) + z(t))'\eta(t)\ dt + \int_0^T (Fy(t) + \bar{z}(t))'d\xi(t) \\ & = \int_0^T (c(t) - G'\pi(t) + H'\eta(t))'\ u(t)\ dt \\ & \quad + \int_0^T y(t)'\ d\left(E'\pi(t) + F'\xi(t) - \int_t^T g(t)\ dt \right) \\ & \quad + \int_0^T z(t)'\eta(t)\ dt + \int_0^T \bar{z}(t)'d\xi(t) \end{aligned}$$

$$= \int_0^T \left(c(t) - G'\pi(t) + H'\eta(t)\right)' u(t)\ dt + \int_0^T z(t)'\eta(t)\ dt + \int_0^T \bar{z}(t)'d\xi(t)$$
$$\geq 0.$$

Note that in general $\int_0^T (Fy(t))'d\xi(t)$ and $\int_0^T \bar{z}(t)'d\xi(t)$ may not exist since neither $y(t)$ nor $\bar{z}(t)$ needs to be continuous. However, since

$$\int_0^T y(t)'\ d\left(E'\pi(t) + F'\xi(t) - \int_t^T g(t)\ dt\right) = 0,$$

we have

$$\begin{aligned}\int_0^T y(t)'\ dF'\xi(t) &= \int_0^T y(t)'\ d\left(E'\pi(t) + F'\xi(t) - \int_t^T g(t)\ dt\right)\\ &\quad - \int_0^T y(t)'\ d\left(E'\pi(t) - \int_t^T g(t)\ dt\right)\\ &= - \int_0^T y(t)'\ d\left(E'\pi(t) - \int_t^T g(t)\ dt\right),\end{aligned}$$

which implies that $\int_0^T y(t)'\ dF'\xi(t)$ exists. The existence of $\int_0^T \bar{z}(t)'d\xi(t)$ now follows from the existence of both $\int_0^T y(t)'\ dF'\xi(t)$ and $\int_0^T (Fy(t) + \bar{z}(t))'d\xi(t)$ since

$$\int_0^T \bar{z}(t)'d\xi(t) = \int_0^T (Fy(t) + \bar{z}(t))'d\xi(t) - \int_0^T y(t)'\ dF'\xi(t). \qquad \square$$

The requirement that $\pi(t)$ is bounded, measurable, and of finite variation in $(SCSCLP^*)$ is important, as it makes the integration by parts valid in the proof of the above proposition (see also Harrison [19]). As a consequence of the proof, we have the following corollary.

COROLLARY 1. *Strong duality holds between* $(SCSCLP)$ *and* $(SCSCLP^*)$ *if and only if there exist* $(u(t),\ y(t))$ *and* $(\pi(t),\ \eta(t),\ \xi(t))$ *which are feasible to* $(SCSCLP)$ *and* $(SCSCLP^*)$*, respectively, and satisfy the following conditions:*

$$\begin{aligned}&\int_0^T \left(c(t) - G'\pi(t) + H'\eta(t)\right)' u(t)\ dt = 0,\\ &\int_0^T (b(t) - Hu(t))'\eta(t)\ dt = 0,\\ &\int_0^T (h(t) - Fy(t))'d\xi(t) = 0.\end{aligned} \tag{5}$$

We call all three equations in (5) the complementary slackness condition for $(SCSCLP)$ and $(SCSCLP^*)$.

The following are standard definitions and notations which we will use throughout the remainder of the paper.

We call a sequence of time epochs $P = \{t_0, \ldots, t_p\}$ a partition of $[0,\ T]$ if

$$0 = t_0 \leq t_1 \leq \cdots \leq t_p = T.$$

We use $|P|$ to denote the cardinality of P. Note that since our development sometimes treats t_i as a variable, we allow $t_i = t_{i-1}$ for some $i \geq 1$ and always treat t_i and t_{i-1} as two different variables.

We say that a function $f(t)$ is piecewise constant (linear) with a partition $P = \{t_0, \ldots, t_p\}$ if $f(t)$ is constant (linear) on $[t_{i-1},\ t_i)$ for $i = 1, \ldots, p$. We say $f(t)$ is piecewise constant (linear) on $[0,\ T]$ if $f(t)$ is piecewise constant (linear) with some partition of $[0,\ T]$.

Let $P = \{t_0, \ldots, t_p\}$ be a partition of $[0,\ T]$. Throughout this paper, we assume Assumption 1 holds. In particular, we assume that $a(t)$, $h(t)$, and $c(t)$ are piecewise linear and $b(t)$ and $g(t)$ are piecewise constant with partition P. Let $\mathcal{B}$ be the set of breakpoints of $a(t)$, $b(t)$, $c(t)$, $g(t)$, and $h(t)$. For each breakpoint in $\mathcal{B}$, we select one element t_i in P such that its value denotes the same time in $[0,\ T]$ as the breakpoint. We always select $t_0 = 0$ and $t_p = T$. We denote D^P to be the set of selected elements of P excluding t_0 and t_p. Let $D_1^P = D^P \bigcup \{t_0, t_p\}$. We sometimes omit the superscript P when the context is clear.

We say that an interval $[t_{i-1},\ t_i]$ is a subinterval of $[t_l,\ t_m]$, where t_l and t_m are two consecutive breakpoints in D_1^P, if $l \leq i-1 < i \leq m$. In this case, we also say that t_{i-1}, t_i, and $[t_{i-1},\ t_i]$ reside on $[t_l,\ t_m]$.

For a function $f(t)$, we will use the notation

$$f(t-) = \lim_{s \to t-} f(s) \qquad \text{and} \qquad f(t+) = \lim_{s \to t+} f(s),$$

when the above limits exist and t is not equal to any breakpoint in D_1^P. If $[t_{i-1},\ t_i]$ is a zero-length subinterval of $[t_l,\ t_m]$, where t_l and t_m are two consecutive breakpoints in D_1^P, we let

$$f(t_i-) = \begin{cases} \lim_{s \to t_i-} f(s) & \text{if } t_i = t_m, \\ \lim_{s \to t_i+} f(s) & \text{if } t_i = t_l, \end{cases}$$

and let $f(t_{i-1}+) = f(t) = f(t_i-)$. We note that the value of $f(t_i)$ is sensitive to both the value of t_i and its index i.

Given $t_i \neq t_{i-1}$ for $i = 1, \ldots, p$ and a set of $2p$ numbers $\hat{f}(t_0+)$, $\hat{f}(t_1-)$, $\hat{f}(t_1+), \ldots,$ $\hat{f}(t_{p-1}+)$, $\hat{f}(t_p-)$, the function $f(t)$ defined by

$$f(t) = \begin{cases} \hat{f}(t_i+) & \text{if } t = t_0, t_1, \ldots,\ t_{p-1}, \\ 0 & \text{if } t = T, \\ \frac{t_i - t}{t_i - t_{i-1}} \hat{f}(t_{i-1}+) + \frac{t - t_{i-1}}{t_i - t_{i-1}} \hat{f}(t_i-) & \text{for } t \in (t_{i-1},\ t_i),\ i = 1, \ldots, p \end{cases}$$

is called the piecewise-linear extension of these $2p$ numbers; for a set of p numbers $\hat{f}(t_0+)$, $\hat{f}(t_1+)$, $\ldots$, $\hat{f}(t_{p-1}+)$, the function $f(t)$ defined by

$$f(t) = \begin{cases} \hat{f}(t_{p-1}+) & t = T, \\ \hat{f}(t_{i-1}+) & \text{for } t \in [t_{i-1},\ t_i),\ i = 1, \ldots, p \end{cases}$$

is called the piecewise constant extension of these p variables.

For two functions $f(t)$ and $g(t)$, we denote $\int_a^b f(t)\ dg(t)$ as the Lebesgue–Stieltjes integral of $f(t)$ with respect to $g(t)$ from a to b, given that the integral exists, including both a and b. For any mathematical program (LP) we let $V(\text{LP})$ be the optimal value of the objective function, which may not be attained. For any feasible solution x of (LP), we let $V(\text{LP}, x)$ be the solution value of x in (LP). For any n-dimensional

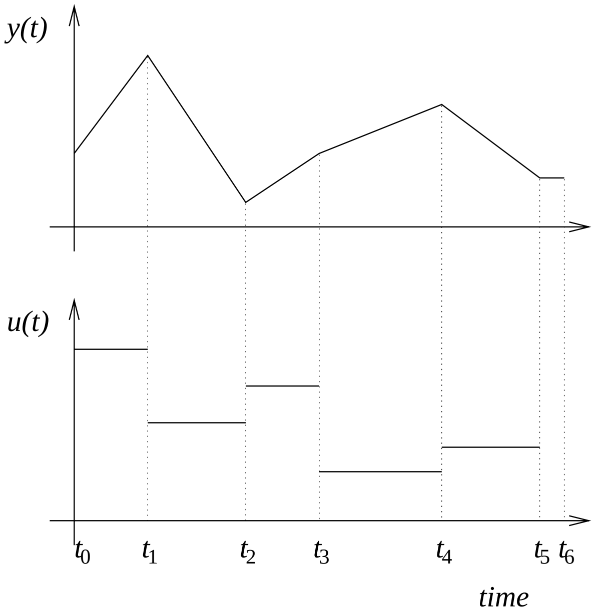


FIG. 1. *A piecewise constant optimal control for* $(SCLP)$.

vector x, we denote by x_i the ith coordinate of x, and, for any nonempty subset $Q \subseteq \{1,\ldots,n\}$, we use x_Q, $[x]_Q$, or $(x)_Q$ to denote the vector with components x_i, $i \in Q$ (with x_i arranged in the same order as in x). For a matrix A, we denote by A_{ij} the jth element of the ith row of matrix A and denote by $A_{i\bullet}$ the ith row of A.

3. A quadratic programming subproblem.

By a result of Pullan [42] (see also Anderson and Nash [2]), there exists an optimal basic feasible solution to $(SCLP)$ whose $u(t)$ is piecewise constant (see Figure 1) when Assumption 1 holds and the solution set to $(SCLP)$ is bounded. We will prove later in the paper that this remains true for $(SCSCLP)$. For any feasible control $u(t)$ that is piecewise constant with respect to a partition P, we have the following standard linear approximation problem (see Pullan [41] and the references therein):

$$
\begin{aligned}
DP(P)\min & \sum_{i=1}^{p}(t_i - t_{i-1})\hat{u}(t_{i-1}+)'c\left(\frac{t_i + t_{i-1}}{2}\right) \\
& + \sum_{i=1}^{p}\frac{t_i - t_{i-1}}{2}(\hat{y}(t_i) + \hat{y}(t_{i-1}))'g(t_{i-1}+) \\
\text{s.t. } & E\hat{y}(t_0) = a(t_0), \\
& (t_i - t_{i-1})G\hat{u}(t_{i-1}+) + E\hat{y}(t_i) - E\hat{y}(t_{i-1}) = a(t_i) - a(t_{i-1}), \\
& \qquad i = 1,\ldots,p, \\
& H\hat{u}(t_{i-1}+) \le b(t_{i-1}+), \qquad i = 1,\ldots,p, \\
& F\hat{y}(t_i) \le h(t_i), \qquad i = 0,\ldots,p, \\
& \hat{u}(t_{i-1}+) \ge 0, \qquad i = 1,\ldots,p,
\end{aligned}
$$

where we have the convention that $c(\frac{t_i+t_{i-1}}{2}) = c(t_i-)$ whenever $t_i = t_{i-1}$. Note that even though it is possible that $t_i = t_{i-1}$ for some $i \geq 1$, we still treat $\hat{u}(t_i+)$ and $\hat{u}(t_{i-1}+)$ as separate variables. If $(\hat{u},\ \hat{y})$ is a feasible solution to $DP(P)$, where partition P satisfies $t_i \neq t_{i-1}$ for all i, the piecewise constant extension of $\hat{u}$, together with the piecewise-linear extension of $\hat{y}$, defines a feasible solution to $(SCSCLP)$ with the same cost, due to Assumption 1. If we fix the partition, $DP(P)$ is a linear programming problem. So, once an optimal partition P is known, an optimal solution can be computed by solving the linear program $DP(P)$.

However, we do not know the optimal partition in advance. The algorithms proposed by Pullan [41] and by Philpott and Craddock [39] alternatively do the following two steps:

1) Improve the control for the current partition.
2) Improve the partition.

In contrast, the algorithm we propose improves both the control and partition at the same time.

By introducing new variables

$$\hat{v}(t_i) = (t_i - t_{i-1})\hat{u}(t_{i-1}+), \tag{6}$$

we can eliminate variable $\hat{u}$ from $DP(P)$ and obtain the following simpler mathematical programming problem in variables $\hat{v}$, $\hat{y}$, and $\hat{t}$, with $\hat{t}$ being the vector of t_i's such that $t_i \notin D_1^P$:

$$\begin{aligned} QP(|P|) \min \sum_{i=1}^{p} \hat{v}(t_i)' c\left(\frac{t_i + t_{i-1}}{2}\right) + \sum_{i=1}^{p} \frac{t_i - t_{i-1}}{2}(\hat{y}(t_i) + \hat{y}(t_{i-1}))' g(t_{i-1}+) \\ \text{s.t. } E\hat{y}(t_0) = a(t_0), \\ G\hat{v}(t_i) + E\hat{y}(t_i) - E\hat{y}(t_{i-1}) = a(t_i) - a(t_{i-1}), \\ i = 1, \dots, p, \\ H\hat{v}(t_i) \leq (t_i - t_{i-1})b(t_{i-1}+), \qquad i = 1, \dots, p, \\ F\hat{y}(t_i) \leq h(t_i), \qquad i = 0, \dots, p, \\ 0 = t_0 \leq t_1 \leq \cdots \leq t_p = T, \\ \hat{v}(t_i) \geq 0, \qquad i = 1, \dots, p, \end{aligned} \tag{7}$$

where $c(\frac{t_i+t_{i-1}}{2}) = c(t_i-)$ whenever $t_i = t_{i-1}$. Note that the breakpoints in D_1^P are fixed and are not variables. We treat both $\hat{v}(t_i)$ and $\hat{y}(t_i)$ as variables. Let t_l and t_m be two consecutive breakpoints in D_1^P. For any $i \in (l,\ m]$, $c(\frac{t_i+t_{i-1}}{2})$, $a(t_i) - a(t_{i-1})$, and $h(t_i)$ are the following linear functions of t_i and t_{i-1} (note that by Assumption 1b, $c(t)$, $a(t)$, and $h(t)$ are piecewise linear and therefore $\dot{c}(t)$, $\dot{a}(t)$, and $\dot{h}(t)$ are piecewise constant):

$$\begin{aligned} c\left(\frac{t_i + t_{i-1}}{2}\right) &= c(t_l) + \frac{t_i + t_{i-1} - 2t_l}{2}\dot{c}(t_l+), \\ a(t_i) - a(t_{i-1}) &= (t_i - t_{i-1})\dot{a}(t_l+), \\ h(t_i) &= h(t_l) + (t_i - t_l)\dot{h}(t_l+), \end{aligned}$$

and $g(t_{i-1}+) = g(t_l)$ and $b(t_{i-1}+) = b(t_l)$ are constant vectors. So, $QP(|P|)$ is a quadratic programming problem with polyhedral constraints.

Given a feasible solution $(\hat{v},\ \hat{y},\ \hat{t})$ to $QP(|P|)$ such that $t_i \neq t_{i-1}$ for all i, we can obtain a feasible solution $(\hat{u}, \hat{y})$ to problem $DP(P)$ with P defined from vector $\hat{t}$ and the breakpoints in D_1^P and $\hat{u}$ defined from

$$\hat{u}(t_{i-1}+) = \frac{\hat{v}(t_i)}{t_i - t_{i-1}}. \tag{8}$$

Equation (6) defines an injective mapping from the solution set to $DP(P)$ to the solution set to $QP(|P|)$. The two related solutions have the same solution value.

However, if $t_i = t_{i-1}$ but $\hat{v}(t_i) \neq 0$ for some i, the right-hand side of (8) is not properly defined, i.e., there may be a solution to $QP(|P|)$ for which the corresponding solution to $DP(P)$ cannot be constructed. We overcome this difficulty by constantly removing redundant zero-length intervals in a feasible solution and by using only the solution $(\hat{v},\ \hat{y},\ \hat{t})$ to $QP(|P|)$ that satisfies

$$t_i \neq t_{i-1} \quad \text{for all} \quad i \geq 1 \tag{9}$$

to construct a feasible solution for $DP(P)$ (and so for $(SCSCLP)$). When some zero-length intervals cannot be removed, we show that there is a series of feasible solutions to $QP(|P|)$ that satisfies (9) whose solution value becomes arbitrarily close to that of the feasible solution to $QP(|P|)$. This is key to understanding the absence of a duality gap result between $(SCSCLP)$ and $(SCSCLP^*)$, as we will see later on.

LEMMA 1. *Suppose $u(t)$ in all feasible solutions to $(SCSCLP)$ is bounded. Let $(\hat{v},\ \hat{y},\ \hat{t})$ be a feasible solution to $QP(|P|)$. Then*

$$\hat{v}(t_i) = 0 \quad \textit{whenever} \quad t_i = t_{i-1}. \tag{10}$$

Proof. Suppose $t_i = t_{i-1}$ for some i, but $\hat{v}(t_i) \neq 0$. Let $[t_l,\ t_m]$ be the interval t_i resides on, where t_l and t_m are two consecutive breakpoints in D_1^P. Without loss of generality, we may assume that there exists a positive-length subinterval of $[t_l,\ t_m]$ that is adjacent to $[t_{i-1},\ t_i]$ (since we can switch the values of $\hat{v}(t_j)$ and $\hat{y}(t_j)$ with adjacent zero-length subintervals on $[t_l,\ t_m]$ and maintain the feasibility of the solution). We assume that the adjacent positive-length subinterval on $[t_l,\ t_m]$ is $[t_{i-2},\ t_{i-1}]$. When the adjacent positive-length subinterval of $[t_l,\ t_m]$ is $[t_i,\ t_{i+1}]$, a similar analysis applies.

For any $\tau \in (0,\ 1)$, it is easy to verify that the following solution is feasible for $QP(|P|)$:

$$\tilde{t}_j^\tau = \begin{cases} t_j & \text{if } j \neq i-1 \text{ and } j \neq i, \\ \tau t_{i-2} + (1-\tau)t_{i-1} & \text{if } j = i-1, \\ t_i & \text{if } j = i, \end{cases}$$

$$\tilde{v}^\tau(\tilde{t}_j^\tau) = \begin{cases} \hat{v}(t_j) & \text{if } j \neq i-1 \text{ and } j \neq i, \\ (1-\tau)\hat{v}(t_{i-1}) & \text{if } j = i-1, \\ \tau\hat{v}(t_{i-1}) + \hat{v}(t_i) & \text{if } j = i, \end{cases}$$

$$\tilde{y}^\tau(\tilde{t}_j^\tau) = \begin{cases} \hat{y}(t_j) & \text{if } j \neq i-1 \text{ and } j \neq i, \\ \tau\hat{y}(t_{i-2}) + (1-\tau)\hat{y}(t_{i-1}) & \text{if } j = i-1, \\ \hat{y}(t_i) & \text{if } j = i. \end{cases}$$

The basic idea is to split interval $[t_{i-2},\ t_{i-1}]$ into two intervals $[t_{i-2},\ \tau t_{i-2}+(1-\tau)t_{i-1}]$ and $[\tau t_{i-2} + (1-\tau)t_{i-1},\ t_{i-1}]$ and combine the second interval with $[t_{i-1},\ t_i]$. It is easy to check that $(\tilde{v}^\tau,\ \tilde{y}^\tau,\ \tilde{t}^\tau)$ is feasible for $QP(|P|)$ and has one less zero-length interval than $(\hat{v},\ \hat{y},\ \hat{t})$. Applying the same process repeatedly, we can eliminate all the zero-length intervals in the solution $(\hat{v},\ \hat{y},\ \hat{t})$.

Let $(\bar{v}^{\tau},\ \bar{y}^{\tau},\ \bar{t}^{\tau})$ be the resulting solution and let Q be the resulting partition. Hence, $(\bar{v}^{\tau},\ \bar{y}^{\tau},\ \bar{t}^{\tau})$ is feasible for $QP(|Q|)$. From this solution, we can construct a feasible solution for $DP(Q)$ (and thus for $(SCSCLP)$) by using (8). However, as τ tends to zero, the corresponding feasible solution to $DP(P)$ is unbounded from above (since the denominator in (8) goes to zero, but the numerator is bounded away from zero). Thus $u(t)$ in $(SCSCLP)$ is unbounded and this creates a contradiction. □

We remark that Lemma 1 implies that if $u(t)$ is bounded and E is an identity matrix (e.g., a bounded and feasible $(SCLP)$), then $\hat{y}(t_{i-1}) = \hat{y}(t_i)$ whenever $t_{i-1} = t_i$. In general, when (10) holds, it is possible that $\hat{y}(t_{i-1}) \neq \hat{y}(t_i)$ even if $t_{i-1} = t_i$. If in addition to (10), $\hat{y}(t_{i-1}) = \hat{y}(t_i)$ for some i such that $t_{i-1} = t_i$, then we can eliminate the zero-length interval $[t_{i-1},\ t_i]$ from $(\hat{v},\ \hat{y},\ \hat{t})$ while maintaining the feasibility and improving the solution value of the solution. This fact will be used later in section 4 to remove redundant intervals.

In general, $u(t)$ may not be bounded in a feasible solution to $(SCSCLP)$. It is possible that there is no feasible solution to $(SCSCLP)$ that is optimal for $(SCSCLP)$. This perhaps is the key difficulty in establishing the absence of a duality gap between $(SCSCLP)$ and $(SCSCLP^*)$ by conventional methods. Hence, we have the following relationship between $(SCSCLP)$ and $QP(|P|)$.

LEMMA 2. *Given any feasible solution* $(\hat{v},\ \hat{y},\ \hat{t})$ *to* $QP(|P|)$, *there exists a series of feasible solutions* $(\hat{v}^k,\ \hat{y}^k,\ \hat{t}^k)$ *to* $QP(|P|)$ *that satisfies* (9) *and whose solution value becomes arbitrarily close to that of* $(\hat{v},\ \hat{y},\ \hat{t})$ *as* k *tends to infinity.*

Proof. By using the same procedure used to prove Lemma 1, we can construct a solution $(\bar{v}^{\tau},\ \bar{y}^{\tau},\ \bar{t}^{\tau})$ which is feasible to $QP(|Q|)$ for some partition Q and satisfies (9). It is easily verified that the solution value of $(\bar{v}^{\tau},\ \bar{y}^{\tau},\ \bar{t}^{\tau})$ to $QP(|Q|)$ becomes arbitrarily close to that of $(\hat{v},\ \hat{y},\ \hat{t})$ as τ goes to zero. □

In fact, we can have $t_i \neq t_{i+1}$ and $t_{i-1} \neq t_{i-2}$ whenever $t_i = t_{i-1}$ in a local optimum for $QP(|P|)$. The existence of $\hat{v}(t_i) \neq 0$ but $t_i = t_{i-1}$ indicates the presence of the Dirac δ function in $u(t)$ at time t_i.

A direct consequence of Lemma 2 is $V((SCSCLP)) \leq V(QP(|P|))$ for all P. This fact enables us to solve $(SCSCLP)$ through solving $QP(|P|)$ for a series of partitions. We note that $V(QP(|P|)) = V(SCSCLP)$ does not imply that there is a feasible solution for $(SCSCLP)$, whose solution value is equal to $V(QP|P|)$, due to the possible presence of zero-length intervals in P.

3.1. Finding a KKT point for $QP(|P|)$.

We do not need to solve the nonconvex quadratic program $QP(|P|)$ to optimality, as we will see in section 6. We only need to compute a series of KKT points (or equivalently, stationary points) of a set of quadratic programs. We use the Frank–Wolfe method (see Martos [33] and Murty [36]) or general matrix-splitting algorithms (see Lin and Pang [28], Eckstein [11], Bertsekas and Tsitsiklis [9], Luo and Tseng [31]) to find a KKT point for the quadratic program. There are other methods for obtaining a KKT point, such as those proposed by Ye [51] and Kojima, Noma, and Yoshise [24].

4. Removing redundant intervals.

After finding a KKT point of $QP(|P|)$, it is possible that some zero-length intervals can be removed, as we noted following Lemma 1. It is also possible that some adjacent intervals can be merged while improving the solution value. The reduction of unnecessary control pieces in the solution is a key feature of the new algorithm. This enables us to prove the convergence of the new algorithm without requiring the norm of the maximal length interval in the discretization to tend to zero (cf. Pullan [41] and Philpott and Craddock [39]).

To do this, let $(\hat{v},\ \hat{y},\ \hat{t})$ be a feasible solution to $QP(|P|)$ and let $[t_{i-1},\ t_i]$ and $[t_i,\ t_{i+1}]$ be two adjacent intervals that reside on $[t_l,\ t_m]$, where t_l and t_m are two consecutive breakpoints in D_1^P. We eliminate t_i from P (or equivalently, combine $[t_{i-1},\ t_i]$ and $[t_i,\ t_{i+1}]$) and define a new feasible solution $(\tilde{v},\ \tilde{y},\ \tilde{t})$ for $QP(|P \setminus \{t_i\}|)$ as follows. Let $\tilde{v}$ be the vector formed by removing $\hat{v}(t_{i+1})$ from $\hat{v}$ and then replacing $\hat{v}(t_i)$ with $\hat{v}(t_i) + \hat{v}(t_{i+1})$, let $\tilde{y}$ be the vector formed by removing $\hat{y}(t_i)$ from $\hat{y}$, and let $\tilde{t}$ be the vector formed by removing t_i from $\hat{t}$.

Lemma 3. *Let $[t_{i-1},\ t_i]$ and $[t_i,\ t_{i+1}]$ be two adjacent intervals that reside on $[t_l,\ t_m]$, where t_l and t_m are two consecutive breakpoints in D_1^P. If*

$$(t_{i+1} - t_i)\hat{v}(t_i)'\dot{c}(t_{i-1}+) + (t_{i+1} - t_i)\hat{y}(t_{i-1})'g(t_{i-1}+) + (t_i - t_{i-1})\hat{y}(t_{i+1})'g(t_{i-1}+)$$

$$\leq (t_i - t_{i-1})\hat{v}(t_{i+1})'\dot{c}(t_{i-1}+) + (t_{i+1} - t_{i-1})\hat{y}(t_i)'g(t_{i-1}+), \tag{11}$$

then we can combine $[t_{i-1},\ t_i]$ and $[t_i,\ t_{i+1}]$ while maintaining the feasibility and improving the solution value of a feasible solution to $QP(|P|)$.

Proof. The difference between the solution value of $(\hat{v},\ \hat{y},\ \hat{t})$ and that of the solution $(\tilde{v},\ \tilde{y},\ \tilde{t})$ is the following:

$$\begin{aligned}
&\hat{v}(t_i)'c\left(\frac{t_i + t_{i-1}}{2}\right) + \hat{v}(t_{i+1})'c\left(\frac{t_{i+1} + t_i}{2}\right) + \frac{t_i - t_{i-1}}{2}(\hat{y}(t_i) + \hat{y}(t_{i-1}))'g(t_{i-1}+)\\
&+ \frac{t_{i+1} - t_i}{2}(\hat{y}(t_{i+1}) + \hat{y}(t_i))'g(t_{i-1}+) - (\hat{v}(t_i) + \hat{v}(t_{i+1}))'c\left(\frac{t_{i+1} + t_{i-1}}{2}\right)\\
&- \frac{t_{i+1} - t_{i-1}}{2}(\hat{y}(t_{i+1}) + \hat{y}(t_{i-1}))'g(t_{i-1}+)\\
= &-\frac{1}{2}\left((t_{i+1} - t_i)\,\hat{v}(t_i)'\dot{c}(t_{i-1}+) + (t_{i+1} - t_i)\hat{y}(t_{i-1})'g(t_{i-1}+)\right.\\
&\left.+ (t_i - t_{i-1})\hat{y}(t_{i+1})'g(t_{i-1}+)\right)\\
&+ \frac{1}{2}\left((t_i - t_{i-1})\hat{v}(t_{i+1})'\dot{c}(t_{i-1}+) + (t_{i+1} - t_{i-1})\hat{y}(t_i)'g(t_{i-1}+)\right).
\end{aligned}$$

We see that the new solution has a smaller solution value if and only if (11) holds. □

A direct corollary to Lemma 3 is the following.

Corollary 2. *Let t_l and t_m be two consecutive breakpoints in D_1^P. We can combine adjacent zero-length intervals in $[t_l,\ t_m]$ while maintaining the feasibility and improving the solution value of a feasible solution to $QP(|P|)$.*

Proof. Let $[t_{i-1},\ t_i]$ and $[t_i,\ t_{i+1}]$ be two adjacent zero-length intervals that reside on $[t_l,\ t_m]$. Since $t_{i-1} = t_i = t_{i+1}$, (11) is trivially satisfied. By Lemma 3, we can combine $[t_{i-1},\ t_i]$ and $[t_i,\ t_{i+1}]$ and maintain the feasibility and improve the solution value of the feasible solution to $QP(|P|)$. □

By Corollary 2, we can combine adjacent zero-length intervals. The following lemma implies that all the zero-length intervals except those at the breakpoints in D_1^P can be eliminated.

Lemma 4. *Let $[t_{i-1},\ t_i]$ be a zero-length interval that resides on $[t_l,\ t_m]$, where t_l and t_m are two consecutive breakpoints in D_1^P. Suppose $[t_{i-2},\ t_{i-1}]$ and $[t_i,\ t_{i+1}]$ are two positive-length intervals that also reside on $[t_l,\ t_m]$. We can either*

(a) *combine $[t_{i-2},\ t_{i-1}]$ and $[t_{i-1},\ t_i]$, or*

(b) *combine $[t_{i-1},\ t_i]$ and $[t_i,\ t_{i+1}]$,* *while maintaining the feasibility and improving the solution value of the feasible solution to* $QP(|P|)$.

Proof. Since $t_{i-1} = t_i$, by Lemma 3, we can combine $[t_{i-2},\ t_{i-1}]$ and $[t_{i-1},\ t_i]$ if the following relation holds:

$$(12)\quad (t_{i-1} - t_{i-2})\hat{y}(t_i)'g(t_{i-2}+) \leq (t_{i-1} - t_{i-2})(\hat{v}(t_i)'\dot{c}(t_{i-2}+) + \hat{y}(t_{i-1})'g(t_{i-2}+)).$$

By Lemma 3 again, we can combine $[t_{i-1},\ t_i]$ and $[t_i,\ t_{i+1}]$ if the following relation holds:

$$(13)\qquad (t_{i+1} - t_i)\hat{y}(t_i)'g(t_i+) \geq (t_{i+1} - t_i)(\hat{v}(t_i)'\dot{c}(t_i+) + \hat{y}(t_{i-1})'g(t_i+)).$$

By assumption, we have $t_{i+1} - t_i > 0$ and $t_{i-1} - t_{i-2} > 0$. Since $c(t)$ is linear and $g(t)$ is constant on $[t_l,\ t_m]$, we have

$$g(t_{i-2}+) = g(t_i+) \quad \text{and} \quad \dot{c}(t_{i-2}+) = \dot{c}(t_i+).$$

So either (12) or (13) is true. This proves the lemma. □

We next propose the following procedure for removing redundant intervals on $[t_l,\ t_m]$, where t_l and t_m are two consecutive breakpoints in D_1^P.

Procedure PURIFY. Repeatedly combine two adjacent intervals $[t_{i-1}, t_i]$ and $[t_i, t_{i+1}]$ in $[t_l,\ t_m]$ if (11) is satisfied.

When more than one pair of adjacent intervals satisfies (11), we can combine them in an arbitrary order, one pair at a time. Let $\tilde{P}$ be the resulting partition of $[0,\ T]$ after we apply the above procedure to P for all consecutive breakpoints in D_1^P. We call $\tilde{P}$ a *purified partition* of $[0,\ T]$. Note that the remaining zero-length intervals are located at the breakpoints in D_1^P and there are at most $2|D_1^P|$ zero-length intervals in P.

5. Improving a nonoptimal solution. One major step of the new algorithm is to calculate a KKT point of the system $QP(|P|)$ for some partition P of $[0,\ T]$. However, the problem $QP(|P|)$ is nonconvex. To obtain a global optimal solution for $(SCSCLP)$, we must be able to improve a solution that is not globally optimal for $(SCSCLP)$. In this section, we give descent directions for solutions that are not globally optimal for $(SCSCLP)$. To do so, we first introduce a new discrete approximation for $(SCSCLP)$ which is closely related to the dual problem $(SCSCLP^*)$. From this new approximation, we derive a criterion that detects whether a solution is globally optimal for $(SCSCLP)$. If this criterion is not satisfied, we give a descent direction for the current solution and thus improve the solution value. We show that instead of using the direction constructed in section 5.3, an algorithm for $(SCSCLP)$ can also use the Frank–Wolfe method or the matrix-splitting algorithm to find a descent direction. We also show that the first iterate of the Frank–Wolfe method provides an upper bound on the current duality gap.

5.1. A new discrete approximation. For partition $P = \{t_0, \ldots, t_p\}$, we let $P' = \left\{t_0, \frac{t_0+t_1}{2}, t_1, \ldots, \frac{t_{p-1}+t_p}{2}, t_p\right\}$ be a refined partition of P. Consider the following new discrete approximation to $(SCSCLP)$, a close variation of the second discretization in Pullan [41]:

$$AP1(P) \quad \min \sum_{i=1}^{p} \frac{t_i - t_{i-1}}{2} \Bigg(c(t_{i-1}+)'\hat{u}(t_{i-1}+) + c(t_i-)'\hat{u}(t_i-)$$

$$+2\hat{y}\left(\frac{t_{i-1}+t_i}{2}\right)' g(t_{i-1}+)\Bigg)$$

$$\text{s.t. } E\hat{y}(t_0) = a(t_0),$$

$$\begin{aligned}
&\left(\frac{t_i - t_{i-1}}{2}\right) G\hat{u}(t_i-) + E\hat{y}(t_i) - E\hat{y}\left(\frac{t_i + t_{i-1}}{2}\right)\\
&= a(t_i) - a\left(\frac{t_i + t_{i-1}}{2}\right), \\
&\qquad\qquad\qquad\qquad\qquad\qquad i = 1, \ldots, p,\\
&\left(\frac{t_i - t_{i-1}}{2}\right) G\hat{u}(t_{i-1}+) + E\hat{y}\left(\frac{t_i + t_{i-1}}{2}\right) - E\hat{y}(t_{i-1})\\
&= a\left(\frac{t_i + t_{i-1}}{2}\right) - a(t_{i-1}),\\
&\qquad\qquad\qquad\qquad\qquad\qquad i = 1, \ldots, p,\\
&H\hat{u}(t_{i-1}+) \le b(t_{i-1}+), \qquad i = 1, \ldots, p,\\
&H\hat{u}(t_i-) \le b(t_i-), \qquad i = 1, \ldots, p,\\
&F\hat{y}(t_i) \le h(t_i), \qquad i = 0, \ldots, p,\\
&F\hat{y}\left(\frac{t_i + t_{i-1}}{2}\right) \le h\left(\frac{t_i + t_{i-1}}{2}\right), \qquad i = 1, \ldots, p,\\
&\hat{u}(t_i-), \hat{u}(t_{i-1}+) \ge 0, \qquad i = 1, \ldots, p.
\end{aligned}$$

Problem $AP1(P)$ is closely related to the dual problem. The linear programming dual of $AP1(P)$ gives rise to feasible solutions for the dual problem $(SCSCLP^*)$. Thus an optimal solution to $AP1(P)$ contains the dual information. We will construct a descent solution for $(SCSCLP)$ based on a solution for $AP(P)$, a closely related linear program, to be defined shortly.

It is clear that the set of feasible solutions to $AP1(P)$ is the same as the set of feasible solutions to $DP(P')$ if we identify $\hat{u}(t_i-)$ in $AP1(P)$ with $\hat{u}((\frac{t_{i-1}+t_i}{2})+)$ in $DP(P')$. There are two differences between $DP(P')$ and $AP1(P)$, both of which reside in the objective function. First, instead of averaging the cost coefficients of $u(t)$ over each subinterval, the instantaneous values of the cost coefficients at the original breakpoints of P are used. Second, instead of using the average values of the state variable $y(t)$ in each subinterval, the values of $y(t)$ at the midpoint of each subinterval of P are used. It can be checked that any feasible solution for $DP(P)$ defines a feasible solution for $DP(P')$ and thus for $AP1(P)$, and these two solutions have the same solution value.

Similar to $QP(|P|)$, we introduce $\hat{v}$ to eliminate $\hat{u}$, where

$$\hat{v}(t_{i-1}+) = \frac{t_i - t_{i-1}}{2}\hat{u}(t_{i-1}+) \quad \text{and} \quad \hat{v}(t_i-) = \frac{t_i - t_{i-1}}{2}\hat{u}(t_i-). \tag{14}$$

Now $AP1(P)$ is transformed into the following linear program in $\hat{v}$ and $\hat{y}$:

$$AP(P) \quad \min \sum_{i=1}^{p} \Bigg(c(t_{i-1}+)'\hat{v}(t_{i-1}+) + c(t_i-)'\hat{v}(t_i-)$$

$$
\begin{aligned}
&\qquad\qquad +(t_i - t_{i-1})\hat{y}\left(\frac{t_{i-1}+t_i}{2}\right)' g(t_{i-1}+)\Bigg) \\
&\qquad \text{s.t. } E\hat{y}(t_0) = a(t_0), \\
&G\hat{v}(t_i-) + E\hat{y}(t_i) - E\hat{y}\left(\frac{t_i+t_{i-1}}{2}\right) = a(t_i) - a\left(\frac{t_i+t_{i-1}}{2}\right), \\
&\qquad\qquad i = 1,\dots,p, \\
&G\hat{v}(t_{i-1}+) + E\hat{y}\left(\frac{t_i+t_{i-1}}{2}\right) - E\hat{y}(t_{i-1}) = a\left(\frac{t_i+t_{i-1}}{2}\right) - a(t_{i-1}), \\
&\qquad\qquad i = 1,\dots,p, \\
&H\hat{v}(t_{i-1}+) \le \left(\frac{t_i-t_{i-1}}{2}\right) b(t_{i-1}+), \qquad i = 1,\dots,p, \\
&H\hat{v}(t_i-) \le \left(\frac{t_i-t_{i-1}}{2}\right) b(t_i-), \qquad i = 1,\dots,p, \\
&F\hat{y}(t_i) \le h(t_i), \qquad i = 0,\dots,p, \\
&F\hat{y}\left(\frac{t_i+t_{i-1}}{2}\right) \le h\left(\frac{t_i+t_{i-1}}{2}\right), \qquad i = 1,\dots,p, \\
&\hat{v}(t_i-),\ \hat{v}(t_{i-1}+) \ge 0, \quad i = 1,\dots,p.
\end{aligned}
$$

Similar to $AP1(P)$ and $DP(P')$, $AP(P)$ and $QP(|P'|)$ have the same feasible solution set if the partition in $QP(|P'|)$ is fixed to P'. We note that the actual value of $\hat{y}(t_0)$ does not affect the objective value of $AP(P)$ as long as $E\hat{y}(t_0) = a(t_0)$ and $F\hat{y}(t_0) \le h(t_0)$ (which is indeed feasible by assumption). The dual problem for $AP(P)$ (after eliminating $\hat{y}(t_0)$) can be written as

$$
\begin{aligned}
AP^*(P) \max\ & \hat{\pi}(t_0+)' a(t_0) \\
&+ \sum_{i=1}^{p} (\hat{\pi}(t_{i-1}+) + \hat{\pi}(t_i-))' \left(a(t_i) - a\left(\frac{t_i+t_{i-1}}{2}\right)\right) \\
&- \sum_{i=1}^{p} \left(\frac{t_i-t_{i-1}}{2}\right) (\hat{\eta}(t_{i-1}+) + \hat{\eta}(t_i-))' b(t_i-) \\
&+ \sum_{i=1}^{p} \left(\hat{\xi}(t_i)' h(t_i) + \hat{\xi}\left(\frac{t_{i-1}+t_i}{2}\right)' h\left(\frac{t_{i-1}+t_i}{2}\right)\right) \\
\text{s.t. } & c(t_i-) - G'\hat{\pi}(t_i-) + H'\hat{\eta}(t_i-) \ge 0, \quad i = 1,\dots,p, \\
& c(t_{i-1}+) - G'\hat{\pi}(t_{i-1}+) + H'\hat{\eta}(t_{i-1}+) \ge 0, \quad i = 1,\dots,p, \\
& E'(-\hat{\pi}(t_i-) + \hat{\pi}(t_{i-1}+)) + F'\hat{\xi}\left(\frac{t_{i-1}+t_i}{2}\right) = (t_i - t_{i-1}) g(t_{i-1}+), \\
& \qquad\qquad i = 1,\dots,p, \\
& E'(-\hat{\pi}(t_i+) + \hat{\pi}(t_i-)) + F'\hat{\xi}(t_i) = 0, \quad i = 1,\dots,p-1, \\
& E'(\hat{\pi}(t_p-)) + F'\hat{\xi}(t_p) = 0, \\
& \hat{\eta}(t_i-), \hat{\eta}(t_{i-1}+) \ge 0, \quad i = 1,\dots,p, \\
& \hat{\xi}(t_i), \hat{\xi}\left(\frac{t_{i-1}+t_i}{2}\right) \le 0, \quad i = 1,\dots,p.
\end{aligned}
$$

Similar to the second discretization in Pullan [41], the importance of $AP(P)$ lies in the fact that feasible solutions for its dual problem $AP^*(P)$ can be used either to define a feasible solution for $(SCSCLP^*)$ with the same solution value or to define a sequence of feasible solutions for $(SCSCLP^*)$ whose solution value converges to that of the original solution to $AP^*(P)$, as shown in the following theorem.

THEOREM 1. *Suppose that P is a purified partition of $[0,\ T]$ (as defined at the end of section 4). Given any feasible solution $(\hat{\pi},\ \hat{\eta},\ \hat{\xi})$ to $AP^*(P)$, if (9) holds for P, then there exists a feasible solution $(\pi(t),\ \eta(t),\ \xi(t))$ to $(SCSCLP^*)$ whose solution value equals that of $(\hat{\pi},\ \hat{\eta},\ \hat{\xi})$. Otherwise, there exists a series of feasible solutions $(\pi^k(t),\ \eta^k(t),\ \xi^k(t))$ to $(SCSCLP^*)$ that are piecewise linear with partition P^k, whose solution value converges to that of $(\hat{\pi},\ \hat{\eta},\ \hat{\xi})$ with P^k satisfying (9).*

Proof. When there are no zero-length intervals in P (i.e., (9) holds), we let

$$\xi(t) = \begin{cases} \sum_{j=i+1}^{p} \left(\hat{\xi}\left(\frac{t_j+t_{j-1}}{2}\right) + \hat{\xi}(t_j)\right) & \text{if } t = t_i,\ i = 0, 1, \dots,\ p-1, \\ 0 & \text{if } t = T. \end{cases}$$

For $t \in (t_{i-1},\ t_i)$, we let

$$\xi(t) = \frac{t_i - t}{t_i - t_{i-1}}\xi(t_{i-1}) + \frac{t - t_{i-1}}{t_i - t_{i-1}}\left(\xi(t_i) + \hat{\xi}(t_i)\right).$$

We note that $\xi(t)$ is monotonically increasing and right continuous (albeit discontinuous). Let $\pi(t)$ and $\eta(t)$ be the piecewise-linear extensions of $\hat{\pi}$ and $\hat{\eta}$, respectively. It can be shown that $(\pi(t),\ \eta(t),\ \xi(t))$ is a feasible solution for $(SCSCLP^*)$ by virtue of the piecewise linearity of the problem data. Now, let us check the relationship between the solution value of the newly constructed solution of $(SCSCLP^*)$ and the original solution of $AP^*(P)$. Through integration by parts, we have

$$\begin{aligned} & -\int_0^T a(t)'\ d\pi(t) \\ &= -a(t)'\pi(t)\ |_0^T + \int_0^T \pi(t)'\ da(t) \\ &= \hat{\pi}(t_0+)'a(t_0) + \sum_{i=1}^{p}\left(\frac{a(t_i) - a(t_{i-1})}{t_i - t_{i-1}}\right)' \int_{t_{i-1}}^{t_i} \pi(t)\ dt \\ &= \hat{\pi}(t_0+)'a(t_0) + \sum_{i=1}^{p}(\hat{\pi}(t_{i-1}+) + \hat{\pi}(t_i-))'\left(a(t_i) - a\left(\frac{t_i + t_{i-1}}{2}\right)\right). \end{aligned} \tag{15}$$

Since $\eta(t)$ is piecewise linear and $b(t)$ is piecewise constant with partition P, we have

$$\int_{t_{i-1}}^{t_i} b(t)'\eta(t)\ dt = \left(\frac{t_i - t_{i-1}}{2}\right)(\hat{\eta}(t_{i-1}+) + \hat{\eta}(t_i-))'b(t_i-), \quad i = 1, \dots, p.$$

So

$$-\int_0^T b(t)'\eta(t)\ dt = -\sum_{i=1}^{p}\left(\frac{t_i - t_{i-1}}{2}\right)(\hat{\eta}(t_{i-1}+) + \hat{\eta}(t_i-))'b(t_i-). \tag{16}$$

Direct calculation gives

$$-\int_0^T h(t)'\ d\xi(t)$$

$$
\begin{aligned}
&= -h(t)'\xi(t)\big|_0^T + \int_0^T \xi(t)'\, dh(t) \\
&= h(t_0)'\xi(t_0) + \sum_{i=1}^{p} \left(\frac{h(t_i) - h(t_{i-1})}{t_i - t_{i-1}} \right)' \int_{t_{i-1}}^{t_i} \xi(t)\, dt \\
&= h(t_0)' \sum_{j=1}^{p} \left(\hat{\xi}\left(\frac{t_j + t_{j-1}}{2} \right) + \hat{\xi}(t_j) \right) \\
&\quad + \sum_{i=1}^{p} \left(\frac{h(t_i) - h(t_{i-1})}{2} \right)' \\
&\quad \times \left(2 \sum_{j=i+1}^{p} \left(\hat{\xi}\left(\frac{t_j + t_{j-1}}{2} \right) + \hat{\xi}(t_j) \right) + \hat{\xi}\left(\frac{t_i + t_{i-1}}{2} \right) + 2\hat{\xi}(t_i) \right)
\end{aligned}
$$

$$
= \sum_{i=1}^{p} \left(\hat{\xi}(t_i)'h(t_i) + \hat{\xi}\left(\frac{t_{i-1} + t_i}{2} \right)' h\left(\frac{t_{i-1} + t_i}{2} \right) \right). \tag{17}
$$

Combining (15), (16), and (17), we see that $(\pi(t),\ \eta(t),\ \xi(t))$ has the same solution value as $(\hat{\pi},\ \hat{\eta},\ \hat{\xi})$. This proves the first part of the theorem.

Now, suppose (9) does not hold for P. Since P is a purified partition, by Corollary 2 and Lemma 4, the zero-length intervals in P can be located only at the breakpoints in D_1^P. So for any zero-length interval $[t_{i-1},\ t_i]$ that resides on $[t_l,\ t_m]$, where t_l and t_m are two consecutive breakpoints in D_1^P, either $t_{i-1} = t_l$ or $t_i = t_m$. Let $\tau \in (0,\ 1)$. We define a new solution $(\tilde{\pi}^\tau,\ \tilde{\eta}^\tau,\ \tilde{\xi}^\tau)$ in the following way.

If $t_{i-1} = t_l$, we let

$$
\begin{aligned}
\tilde{t}_{i-1}^\tau &= t_{i-1}, \\
\tilde{t}_i^\tau &= (1-\tau)t_i + \tau t_{i+1}, \\
\tilde{t}_{i+1}^\tau &= t_{i+1}, \\
\tilde{\pi}^\tau(\tilde{t}_i^\tau -) &= (1-\tau)\hat{\pi}(t_i -) + \tau\hat{\pi}(t_i +), \\
\tilde{\pi}^\tau(\tilde{t}_i^\tau +) &= (1-\tau)\hat{\pi}(t_i +) + \tau\hat{\pi}(t_{i+1} -), \\
\tilde{\eta}^\tau(\tilde{t}_i^\tau -) &= (1-\tau)\hat{\eta}(t_i -) + \tau\hat{\eta}(t_i +), \\
\tilde{\eta}^\tau(\tilde{t}_i^\tau +) &= (1-\tau)\hat{\eta}(t_i +) + \tau\hat{\eta}(t_{i+1} -), \\
\tilde{\xi}^\tau\left(\frac{\tilde{t}_i^\tau + \tilde{t}_{i-1}^\tau}{2} \right) &= (1-\tau)\hat{\xi}\left(\frac{t_i + t_{i-1}}{2} \right) + \tau\hat{\xi}(t_i), \\
\tilde{\xi}^\tau(\tilde{t}_i^\tau) &= (1-\tau)\hat{\xi}(t_i) + \tau\hat{\xi}\left(\frac{t_i + t_{i+1}}{2} \right), \\
\tilde{\xi}^\tau\left(\frac{\tilde{t}_i^\tau + \tilde{t}_{i+1}^\tau}{2} \right) &= (1-\tau)\hat{\xi}\left(\frac{t_i + t_{i+1}}{2} \right).
\end{aligned}
$$

If $t_i = t_m$, we let

$$
\begin{aligned}
\tilde{t}_{i-2}^\tau &= t_{i-2}, \\
\tilde{t}_{i-1}^\tau &= (1-\tau)t_{i-1} + \tau t_{i-2}, \\
\tilde{t}_i^\tau &= t_i, \\
\tilde{\pi}^\tau(\tilde{t}_{i-1}^\tau -) &= (1-\tau)\hat{\pi}(t_{i-1} -) + \tau\hat{\pi}(t_{i-2} +),
\end{aligned}
$$

$$\begin{aligned}
\tilde{\pi}^\tau(\tilde{t}^\tau_{i-1}+) &= (1-\tau)\hat{\pi}(t_{i-1}+) + \tau\hat{\pi}(t_{i-1}-),\\
\tilde{\eta}^\tau(\tilde{t}^\tau_{i-1}-) &= (1-\tau)\hat{\eta}(t_{i-1}-) + \tau\hat{\eta}(t_{i-2}+),\\
\tilde{\eta}^\tau(\tilde{t}^\tau_{i-1}+) &= (1-\tau)\hat{\eta}(t_{i-1}+) + \tau\hat{\eta}(t_{i-1}-),\\
\tilde{\xi}^\tau\left(\frac{\tilde{t}^\tau_i + \tilde{t}^\tau_{i-1}}{2}\right) &= (1-\tau)\hat{\xi}\left(\frac{t_i + t_{i-1}}{2}\right) + \tau\hat{\xi}(t_{i-1}),\\
\tilde{\xi}^\tau(\tilde{t}^\tau_{i-1}) &= (1-\tau)\hat{\xi}(t_{i-1}) + \tau\hat{\xi}\left(\frac{t_{i-1} + t_{i-2}}{2}\right),\\
\tilde{\xi}^\tau\left(\frac{\tilde{t}^\tau_{i-1} + \tilde{t}^\tau_{i-2}}{2}\right) &= (1-\tau)\hat{\xi}\left(\frac{t_{i-1} + t_{i-2}}{2}\right).
\end{aligned}$$

For all the other quantities not defined in the above cases, we let $\tilde{t}^\tau_j = t_j$, $\tilde{\pi}^\tau(\tilde{t}^\tau_j-) = \hat{\pi}(t_j-)$, $\tilde{\pi}^\tau(\tilde{t}^\tau_j+) = \hat{\pi}(t_j+)$, $\tilde{\eta}^\tau(\tilde{t}^\tau_j-) = \hat{\eta}(t_j-)$, $\tilde{\eta}^\tau(\tilde{t}^\tau_j+) = \hat{\eta}(t_j+)$, $\tilde{\xi}^\tau(\tilde{t}^\tau_{j-1}) = \hat{\xi}(t_{j-1})$, and $\tilde{\xi}^\tau\left(\frac{\tilde{t}^\tau_j+\tilde{t}^\tau_{j-1}}{2}\right) = \hat{\xi}\left(\frac{t_j+t_{j-1}}{2}\right)$.

Let P^τ be the partition defined from $\tilde{t}^\tau$. It is easy to check the feasibility of $(\tilde{\pi}^\tau,\ \tilde{\eta}^\tau,\ \tilde{\xi}^\tau)$ to $AP^*(P^\tau)$. Since $(\tilde{\pi}^\tau,\ \tilde{\eta}^\tau,\ \tilde{\xi}^\tau)$ converges to $(\hat{\pi},\ \hat{\eta},\ \hat{\xi})$ and $\tilde{t}^\tau$ converges to $\hat{t}$ as τ tends to zero, we see that the solution value of $(\tilde{\pi}^\tau,\ \tilde{\eta}^\tau,\ \tilde{\xi}^\tau)$ in $AP^*(P^\tau)$ converges to the solution value of $(\hat{\pi},\ \hat{\eta},\ \hat{\xi})$ in $AP^*(P)$. Furthermore, (9) holds for P^τ. Applying the first part of the theorem to P^τ, we conclude that the theorem is true for P. □

We may now summarize the relationship among the values of various discrete approximations in the following theorem (see also Theorem 3.5 in Pullan [41]).

THEOREM 2. *For any partitions P and Q,*

$$V(AP(P)) = V(AP^*(P)) \le V((SCSCLP^*)) \le V((SCSCLP)) \le V(DP(Q)).$$

Proof. By the strong duality result for finite-dimensional linear programming, the value of the optimal solution to $AP(P)$ is the value of the optimal solution to its dual $AP^*(P)$. By Theorem 1, the solution value of this solution can be closely approximated by a sequence of feasible solutions to $(SCSCLP^*)$. It then follows that this value is a lower bound on $V((SCSCLP^*))$, and thus a lower bound on $V((SCSCLP))$ by Proposition 1. The final inequality follows from the definition of $DP(Q)$. □

COROLLARY 3. *For any partitions P and Q, if*

$$V(AP(P)) \ge V(QP(|Q|)),$$

then the optimal solution value of $QP(|Q|)$ gives the optimal solution value to $(SCSCLP)$. In particular, if a solution $(\hat{v},\ \hat{y},\ \hat{t})$ is feasible for $QP(|Q|)$ and has the same cost as the optimal value of $AP(P)$, then $(\hat{v},\ \hat{y},\ \hat{t})$ gives the optimal solution value for $(SCSCLP)$ which can be closely approximated by a sequence of feasible solutions to $(SCSCLP)$.

Proof. By Lemma 2, the solution value of any feasible solution to $QP(|Q|)$ is an upper bound on $V((SCSCLP))$, and the result follows directly from Theorem 2. □

5.2. The doubling of breakpoints. Based on a new discrete approximation of $(SCLP)$ similar to $AP1(P)$, Pullan [41] found a descent solution for $(SCLP)$ (consequently, a descent direction can be constructed) by patching together the current solution and a solution that has a better solution value in $AP1(P)$ than the current solution. The new solution has a strictly improved solution value in $(SCLP)$ but usually has three times as many constant control pieces as the original solution. In the following, we give a construction for a feasible solution to $(SCSCLP)$ that produces, at most, approximately twice as many breakpoints as the original feasible solution.

Let P be a partition of $[0,\ T]$, and define a new partition as follows:

$$\bar{P} = \{t_0, t_0, t_1, t_1, \ldots, t_i, t_i, t_i, \ldots, t_p, t_p\},$$

where each breakpoint in D^P has two duplicates and all the other breakpoints have only one duplicate. Intuitively, we have placed a zero-length interval at the beginning of every breakpoint of P and put a zero-length interval at the end of each breakpoint in D^P. Under this configuration, the set of intervals in $\bar{P}$ is the union of the intervals in P and a set of zero-length intervals. We let $\bar{t}_i$ denote the $(i+1)$th element of $\bar{P}$. $D_1^{\bar{P}}$ is the set of breakpoints in $\bar{P}$ that correspond to the breakpoints in D_1^P. For the ith interval (i.e., $[t_{i-1},\ t_i]$) in P, we have a corresponding interval $[\bar{t}_{j-1},\ \bar{t}_j]$ in $\bar{P}$, where $\bar{t}_{j-1} = t_{i-1}$ and $\bar{t}_j = t_i$. We call this interval in $\bar{P}$ an old interval. All the other intervals in $\bar{P}$ are called new intervals. Note that all the new intervals have zero length but not vice versa.

Given a solution $(\hat{v},\ \hat{y},\ \hat{t})$ to $QP(|P|)$, we first construct a feasible solution $(\bar{v},\ \bar{y},\ \bar{t})$ to $QP(|\bar{P}|)$ and then show a descent direction for this solution in $QP(|\bar{P}|)$. The descent direction will be used in the proof of convergence. We need not use the same direction in the new algorithm, as we will see in the last remark in section 6. This solution has the same solution value in $QP(|\bar{P}|)$ as the current solution in $QP(|P|)$ and has approximately twice as many intervals, fewer than the one constructed by Pullan [41].

Let $(\hat{v},\ \hat{y},\ \hat{t})$ be a feasible solution for $QP(|P|)$. For the ith interval in $\bar{P}$, if it is an old interval, we let interval j be the corresponding interval in P and set

$$\bar{v}(\bar{t}_i) = \hat{v}(t_j), \qquad \bar{y}(\bar{t}_i) = \hat{y}(t_j). \tag{18}$$

We let $\bar{v}(\bar{t}_i) = 0$ if interval i in $\bar{P}$ is a new interval and let $\bar{y}(\bar{t}_i) = \hat{y}(t_j)$, where j is the interval in P that corresponds to the closest old interval in $\bar{P}$ to the left of $[t_{i-1},\ t_i]$ (with the convention that $\bar{y}(\bar{t}_1) = y(t_0)$ and $\bar{y}(\bar{t}_0) = y(t_0)$).

It is easy to verify that $(\bar{v},\ \bar{y},\ \bar{t})$ is feasible for $QP(|\bar{P}|)$ and has the same solution value in $QP(|\bar{P}|)$ as $(\hat{v},\ \hat{y},\ \hat{t})$ in $QP(|P|)$.

5.3. A descent direction.

According to Corollary 3, a feasible solution $(\hat{v},\ \hat{y},\ \hat{t})$ to $QP(|P|)$ gives the optimal solution value of $(SCSCLP)$ if the optimal solution to $AP(P)$ has an equal or larger solution value. If so, we can stop the algorithm. Otherwise, there exists $(\tilde{\bar{v}},\ \tilde{\bar{y}},\ \tilde{\bar{t}})$ feasible for $AP(P)$ and with a strictly smaller solution value in $AP(P)$, i.e., we have

$$\delta \overset{\text{def}}{=} V(AP(P), (\tilde{\bar{v}}, \tilde{\bar{y}}, \tilde{\bar{t}})) - V(QP(|P|), (\hat{v},\ \hat{y},\ \hat{t})) < 0. \tag{19}$$

Note that $|\delta|$ is an upper bound on the duality gap between $(SCSCLP)$ and $(SCSCLP^*)$.

Let $\epsilon \in [0,\ 1]$. For every interval $[t_{i-1},\ t_i]$, we define

$$\epsilon_i = \frac{(t_i - t_{i-1})\epsilon}{2}.$$

We define a new partition P^ϵ of $[0,\ T]$ as follows:

$$P^\epsilon \overset{\text{def}}{=} \{t_0, t_0 + \epsilon_1, t_1 - \epsilon_1, t_1 + \epsilon_2, \ldots, t_i - \epsilon_i, t_i, t_i + \epsilon_{i+1}, \ldots, t_p - \epsilon_p, t_p\},$$

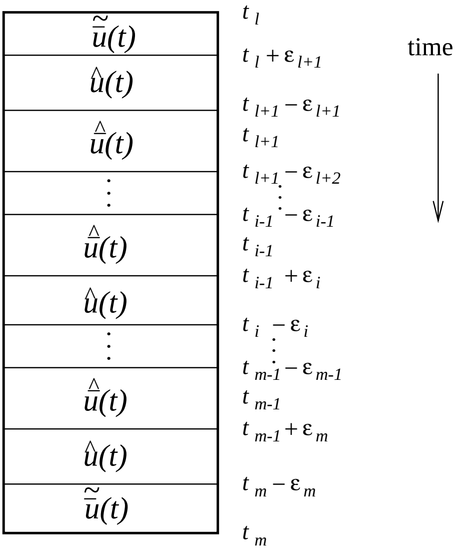

$\tilde{\bar{u}}(t)$
$\hat{u}(t)$
$\overset{\triangle}{u}(t)$
$\overset{\triangle}{u}(t)$
$\hat{u}(t)$
$\overset{\triangle}{u}(t)$
$\hat{u}(t)$
$\tilde{\bar{u}}(t)$
t_l
$t_l + \varepsilon_{l+1}$
$t_{l+1} - \varepsilon_{l+1}$
t_{l+1}
$t_{l+1} - \varepsilon_{l+2}$
$t_{i-1} - \varepsilon_{i-1}$
t_{i-1}
$t_{i-1} + \varepsilon_i$
$t_i - \varepsilon_i$
$t_{m-1} - \varepsilon_{m-1}$
t_{m-1}
$t_{m-1} + \varepsilon_m$
$t_m - \varepsilon_m$
t_m
time

FIG. 2. *The construction of a descent solution where t_l and t_m are two consecutive breakpoints in D_1.*

where we replace the breakpoint t_i in $P \setminus D_1^P$ with two elements $t_i - \epsilon_i$ and $t_i + \epsilon_{i+1}$ and add two elements $t_i - \epsilon_i$ and $t_i + \epsilon_{i+1}$. For breakpoint t_i in D^P, we add $t_0 + \epsilon_1$ and $t_p - \epsilon_p$ for t_0 and t_p, respectively. We define the vector t^ϵ from P^ϵ by mapping t_i^ϵ to the $(i+1)$th element in P^ϵ. We construct a descent solution $(v^\epsilon,\ y^\epsilon,\ t^\epsilon)$ with partition P^ϵ as follows.

When P does not have any zero-length intervals, let $\tilde{\bar{u}}(t)$, $\hat{u}(t)$, and $\hat{\bar{u}}(t)$ be the piecewise constant extensions of $\tilde{\bar{u}}$, $\hat{u}$, and $\hat{\bar{u}}$, respectively, where $\tilde{\bar{u}}$ is defined from $\tilde{\bar{v}}$ by

$$\tilde{\bar{u}}(t_{i-1}+) = 2\frac{\tilde{\bar{v}}(t_{i-1}+)}{t_i - t_{i-1}}, \quad \tilde{\bar{u}}(t_i-) = 2\frac{\tilde{\bar{v}}(t_i-)}{t_i - t_{i-1}},$$

$\hat{u}$ is defined from $\hat{v}$ by (8), and $\hat{\bar{u}}$ is defined as

$$\hat{\bar{u}}(t_{i-1}+) = \frac{\tilde{\bar{v}}(t_{i-1}+) + \tilde{\bar{v}}(t_i-)}{t_i - t_{i-1}}.$$

We construct the new control by patching together $\tilde{\bar{u}}(t)$, $\hat{u}(t)$, and $\hat{\bar{u}}(t)$ as follows:

$$(20) \qquad u^\epsilon(t) = \begin{cases} \tilde{\bar{u}}(t), & t \in [t_{i-1},\ t_{i-1} + \epsilon_i) \bigcup [t_i - \epsilon_i,\ t_i),\ t_i \in D^P, \\ \tilde{\bar{u}}(t), & t \in [t_{p-1},\ t_{p-1} + \epsilon_p) \bigcup [t_p - \epsilon_p,\ t_p], \\ \hat{u}(t), & t \in [t_{i-1} + \epsilon_i,\ t_i - \epsilon_i), \\ \hat{\bar{u}}(t), & \text{otherwise.} \end{cases}$$

Having constructed the control, the construction of the state variables for $(SCSCLP)$ is straightforward. Our construction of a descent solution $(v^\epsilon,\ y^\epsilon,\ t^\epsilon)$ for $(\hat{v},\ \hat{y},\ \hat{t})$ is illustrated in Figure 2.

However, if $t_{i-1} = t_i$ for some i, the u variables in the previous paragraph are not properly defined. Fortunately, we can bypass this difficulty by working on the v variables. We define v^ϵ as follows. Let t_l and t_m be two consecutive breakpoints in

D_1^P. Let $[t_i+\epsilon_{i+1},\ t_{i+1}-\epsilon_{i+1}]$ and $[t_{i+1}-\epsilon_{i+1},\ t_{i+1}+\epsilon_{i+2}]$ be two intervals that reside on $[t_l,\ t_m]$. If t_j^ϵ is the breakpoint in P^ϵ that is mapped to $t_{i+1}-\epsilon_{i+1}$, we let

$$\begin{aligned} v^\epsilon(t_j^\epsilon) &= (1-\epsilon)\hat{v}(t_{i+1}),\\ v^\epsilon(t_{j+1}^\epsilon) &= \epsilon(\tilde{\bar{v}}(t_{i+1}-)+\tilde{\bar{v}}(t_{i+1}+)). \end{aligned}\tag{21}$$

If t_j^ϵ is the breakpoint in P^ϵ that is mapped to t_l, we let

$$v^\epsilon(t_{j+1}^\epsilon) = \epsilon\tilde{\bar{v}}(t_l+).\tag{22}$$

If t_j^ϵ is the breakpoint in P^ϵ that is mapped to t_m, we let

$$v^\epsilon(t_j^\epsilon) = \epsilon\tilde{\bar{v}}(t_m-).\tag{23}$$

We define y^ϵ in three different cases as follows. For the breakpoint t_j^ϵ in P^ϵ that is mapped to $t_{i-1}+\epsilon_i$, we let

$$y^\epsilon(t_j^\epsilon) = (1-\epsilon)\hat{y}(t_{i-1}) + \epsilon\tilde{\bar{y}}\left(\frac{t_i+t_{i-1}}{2}\right).\tag{24}$$

For the breakpoint t_j^ϵ in P^ϵ that is mapped to $t_i-\epsilon_i$, we let

$$y^\epsilon(t_j^\epsilon) = (1-\epsilon)\hat{y}(t_i) + \epsilon\tilde{\bar{y}}\left(\frac{t_i+t_{i-1}}{2}\right).\tag{25}$$

For the breakpoint t_j^ϵ in P^ϵ that is mapped to t_i, we let

$$y^\epsilon(t_j^\epsilon) = (1-\epsilon)\hat{y}(t_i) + \epsilon\tilde{\bar{y}}(t_i).\tag{26}$$

When ϵ is small, $(v^\epsilon,\ y^\epsilon,\ t^\epsilon)$ is a descent solution as shown in the following theorem.

THEOREM 3. *If* (19) *holds, then* $(v^\epsilon,\ y^\epsilon,\ t^\epsilon)$ *is a feasible solution to* $QP(|\bar{P}|)$ *and*

$$V(QP(|\bar{P}|),(v^\epsilon,\ y^\epsilon,\ t^\epsilon)) - V(QP(|\bar{P}|),(\bar{v},\ \bar{y},\ \bar{t})) = \epsilon\delta + o(\epsilon),\tag{27}$$

where δ *is defined in* (19). *For* ϵ *small enough,* $(v^\epsilon,\ y^\epsilon,\ t^\epsilon)$ *has a strictly smaller solution value than* $(\bar{v},\ \bar{y},\ \bar{t})$.

Proof. The feasibility of $(v^\epsilon,\ y^\epsilon,\ t^\epsilon)$ follows easily. By definition, we have

$$\begin{aligned} V(QP(|\bar{P}|),(\bar{v},\ \bar{y},\ \bar{t})) &= V(QP(|P|),(\hat{v},\ \hat{y},\ \hat{t}))\\ &= \sum_{i=1}^{p}\hat{v}(t_i)'c\left(\frac{t_i+t_{i-1}}{2}\right) + \sum_{i=1}^{p}\frac{t_i-t_{i-1}}{2}(\hat{y}(t_i)\\ &\quad + \hat{y}(t_{i-1}))'g(t_{i-1}+),\\ V(QP(|\bar{P}|),(v^\epsilon,\ y^\epsilon,\ t^\epsilon)) &= \sum_{i=1}^{|P^\epsilon|-1} c\left(\frac{t_i^\epsilon+t_{i-1}^\epsilon}{2}\right)' v^\epsilon(t_i^\epsilon)\\ &\quad + \sum_{i=1}^{|P^\epsilon|-1}\frac{t_i^\epsilon-t_{i-1}^\epsilon}{2}(y^\epsilon(t_i^\epsilon)+y^\epsilon(t_{i-1}^\epsilon))'g(t_{i-1}^\epsilon+). \end{aligned}$$

Let t_l and t_m be two consecutive breakpoints in D_1^P and let $t_{\bar{l}}^\epsilon$ and $t_{\bar{m}}^\epsilon$ be the corresponding breakpoints in $D_1^{P^\epsilon}$. We have

$$\sum_{i=l+1}^{m}\hat{v}(t_i)'c\left(\frac{t_i+t_{i-1}}{2}\right) - \sum_{i=\bar{l}+1}^{\bar{m}} c\left(\frac{t_i^\epsilon+t_{i-1}^\epsilon}{2}\right)' v^\epsilon(t_i^\epsilon)$$

$$
\begin{aligned}
&= \sum_{i=l+1}^{m} \hat{v}(t_i)'c\left(\frac{t_i+t_{i-1}}{2}\right) - \sum_{i=l+1}^{m}(1-\epsilon)\hat{v}(t_i)'c\left(\frac{t_i+t_{i-1}}{2}\right)\\
&\quad - \sum_{i=l+1}^{m-1}\epsilon(\tilde{\bar{v}}(t_i+)+\tilde{\bar{v}}(t_i-))'c\left(t_i+\frac{\epsilon_{i+1}-\epsilon_i}{2}\right)\\
&\quad - \epsilon\tilde{\bar{v}}(t_l+)'c\left(t_l+\frac{\epsilon_{i+1}}{2}\right) - \epsilon\tilde{\bar{v}}(t_m-)'c\left(t_m-\frac{\epsilon_m}{2}\right)\\
&= \sum_{i=l+1}^{m}\epsilon\left(\hat{v}(t_i)'c\left(\frac{t_i+t_{i-1}}{2}\right) - (\tilde{\bar{v}}(t_{i-1}+)'c(t_{i-1}+) + \ \tilde{\bar{v}}(t_i-)'c(t_i-))\right) + o(\epsilon),
\end{aligned}
$$

(28)

and

$$
\begin{aligned}
&\sum_{i=l+1}^{m}\frac{t_i-t_{i-1}}{2}(\hat{y}(t_i)+\hat{y}(t_{i-1}))'g(t_{i-1}+) - \sum_{i=\bar{l}+1}^{\bar{m}}\frac{t_i^{\epsilon}-t_{i-1}^{\epsilon}}{2}(y^{\epsilon}(t_i^{\epsilon})+y^{\epsilon}(t_{i-1}^{\epsilon}))'g(t_{i-1}^{\epsilon}+)\\
&= \sum_{i=l+1}^{m}\frac{t_i-t_{i-1}}{2}(\hat{y}(t_i)+\hat{y}(t_{i-1}))'g(t_{i-1}+)\\
&\quad - \sum_{i=l+1}^{m}(1-\epsilon)\frac{t_i-t_{i-1}}{2}\left((1-\epsilon)(\hat{y}(t_i)+\hat{y}(t_{i-1}))+2\epsilon\tilde{\bar{y}}\left(\frac{t_i+t_{i-1}}{2}\right)\right)'g(t_{i-1}+)\\
&\quad - \sum_{i=l+1}^{m}\frac{\epsilon_i+\epsilon_{i+1}}{2}\left(2(1-\epsilon)\hat{y}(t_i)+\epsilon\left(\tilde{\bar{y}}\left(\frac{t_i+t_{i-1}}{2}\right)+\tilde{\bar{y}}\left(\frac{t_{i+1}+t_i}{2}\right)\right)\right)'g(t_{i-1}+)\\
&\quad -\frac{\epsilon_{l+1}}{2}\left(2(1-\epsilon)\hat{y}(t_l)+\epsilon\left(\tilde{\bar{y}}(t_l)+\tilde{\bar{y}}\left(\frac{t_{l+1}+t_l}{2}\right)\right)\right)'g(t_l+)\\
&\quad -\frac{\epsilon_m}{2}\left(2(1-\epsilon)\hat{y}(t_m)+\epsilon\left(\tilde{\bar{y}}(t_m)+\tilde{\bar{y}}\left(\frac{t_{m-1}+t_m}{2}\right)\right)\right)'g(t_m-)\\
&= \sum_{i=l+1}^{m}\epsilon\frac{t_i-t_{i-1}}{2}(\hat{y}(t_i)+\hat{y}(t_{i-1}))'g(t_{i-1}+)\\
&\quad - \sum_{i=l+1}^{m}(t_i-t_{i-1})\epsilon\tilde{\bar{y}}\left(\frac{t_i+t_{i-1}}{2}\right)'g(t_{i-1}+)+o(\epsilon).
\end{aligned}
$$

(29)

Summing up (28) and (29) over all pairs of consecutive breakpoints in D_1, we have

$$
\begin{aligned}
&V(QP(|\bar{P}|),(v^{\epsilon},\ y^{\epsilon},\ t^{\epsilon})) - V(QP(|\bar{P}|),(\bar{v},\ \bar{y},\ \bar{t}))\\
&= V(QP(|\bar{P}|),(v^{\epsilon},\ y^{\epsilon},\ t^{\epsilon})) - V(QP(|P|),(\hat{v},\ \hat{y},\ \hat{t}))\\
&= \sum_{i=1}^{p}\epsilon\left(\tilde{\bar{v}}(t_{i-1}+)'c(t_{i-1}+)+\tilde{\bar{v}}(t_i-)'c(t_i-)-\hat{v}(t_i)'c\left(\frac{t_i+t_{i-1}}{2}\right)\right.\\
&\qquad\qquad \left. -\frac{t_i-t_{i-1}}{2}(\hat{y}(t_i)+\hat{y}(t_{i-1}))'g(t_{i-1}+)\right)+o(\epsilon)\\
&= \epsilon(V(AP(P),(\tilde{\bar{v}},\tilde{\bar{y}},\tilde{\bar{t}})) - V(QP(|P|),(\hat{v},\ \hat{y},\ \hat{t}))) + o(\epsilon)\\
&= \epsilon\delta + o(\epsilon).
\end{aligned}
$$

Since $\delta < 0$, when ϵ is small enough, $(v^\epsilon,\ y^\epsilon,\ t^\epsilon)$ is a strictly improved feasible solution to $QP(|\bar{P}|)$. □

Interestingly, the new solution $(v^\epsilon,\ y^\epsilon,\ t^\epsilon)$ gives a descent direction for $(\bar{v},\ \bar{y},\ \bar{t})$ in $QP(|\bar{P}|)$. This solution can also be used to show that the first Frank–Wolfe iterate for $(\bar{v},\ \bar{y},\ \bar{t})$ provides an upper bound on the current duality gap, as we next illustrate.

Let $[t_l,\ t_m]$ be two consecutive breakpoints in D_1^P. We define a new partition $\bar{\bar{P}}$ as follows. The set of breakpoints of $\bar{\bar{P}}$ that resides on $[t_l,\ t_m]$ is $\{t_l,\ \frac{t_l+t_{l+1}}{2},\ \frac{t_l+t_{l+1}}{2}, \dots,\ t_m\}$, i.e., the union of $\{t_l,\ t_m\}$ with the set of midpoints of the intervals in P, and each midpoint appears exactly twice. We construct $(\bar{\bar{v}},\ \bar{\bar{y}},\ \bar{\bar{t}})$ as follows. The set of breakpoints of $(\bar{\bar{v}},\ \bar{\bar{y}},\ \bar{\bar{t}})$ is $\bar{\bar{P}}$. Let

$$\bar{\bar{v}}_j = \begin{cases} \tilde{\bar{v}}(t_{i+1}-) + \tilde{\bar{v}}(t_{i+1}+) & \text{if the } j\text{th interval of } \bar{\bar{P}} \text{ is } [\frac{t_i+t_{i+1}}{2},\ \frac{t_{i+1}+t_{i+2}}{2}], \\ \tilde{\bar{v}}(t_l+) & \text{if the } j\text{th interval of } \bar{\bar{P}} \text{ is } [t_l,\ \frac{t_l+t_{l+1}}{2}], \\ \tilde{\bar{v}}(t_m-) & \text{if the } j\text{th interval of } \bar{\bar{P}} \text{ is } [\frac{t_{m-1}+t_m}{2},\ t_m], \\ 0 & \text{otherwise,} \end{cases}$$

$\bar{\bar{y}}(\bar{\bar{t}}_0) = y(t_0)$, and

$$\bar{\bar{y}}(\bar{\bar{t}}_j) = \begin{cases} \tilde{\bar{y}}\left(\frac{t_{i+1}+t_{i+2}}{2}\right) & \text{if the } j\text{th interval of } \bar{\bar{P}} \text{ is } [\frac{t_i+t_{i+1}}{2},\ \frac{t_{i+1}+t_{i+2}}{2}], \\ \tilde{\bar{y}}\left(\frac{t_l+t_{l+1}}{2}\right) & \text{if the } j\text{th interval of } \bar{\bar{P}} \text{ is } [t_l,\ \frac{t_l+t_{l+1}}{2}], \\ \tilde{\bar{y}}(t_m) & \text{if the } j\text{th interval of } \bar{\bar{P}} \text{ is } [\frac{t_{m-1}+t_m}{2},\ t_m], \\ \tilde{\bar{y}}\left(\frac{t_{i+1}+t_i}{2}\right) & \text{if the } j\text{-th interval of } \bar{\bar{P}} \text{ is } [\frac{t_i+t_{i+1}}{2},\ \frac{t_i+t_{i+1}}{2}]. \end{cases}$$

THEOREM 4. *For $\epsilon \in [0,\ 1]$, let t^ϵ be defined by P^ϵ. Let $(v^\epsilon,\ y^\epsilon,\ t^\epsilon)$ be the solution to $QP(|\bar{P}|)$ defined by* (21)–(26). *We have*

$$\begin{aligned} v^\epsilon &= \epsilon\bar{\bar{v}} + (1-\epsilon)\,\bar{v}, \\ y^\epsilon &= \epsilon\bar{\bar{y}} + (1-\epsilon)\,\bar{y}, \\ t^\epsilon &= \epsilon\bar{\bar{t}} + (1-\epsilon)\,\bar{t}, \end{aligned}$$

and $(\bar{\bar{v}},\ \bar{\bar{y}},\ \bar{\bar{t}})$ is feasible for $QP(|\bar{P}|)$.

Proof. Theorem 4 obtains the direct consequence of the definition of $(v^\epsilon,\ y^\epsilon,\ t^\epsilon)$ and $(\bar{\bar{v}},\ \bar{\bar{y}},\ \bar{\bar{t}})$. □

If we pick $(\tilde{\bar{v}}, \tilde{\bar{y}},\ \tilde{\bar{t}})$ introduced in (19) as an optimal solution for $AP(P)$, then by Theorem 2, $|\delta|$ is an upper bound on the current duality gap. By (27) and Theorem 4, the negative objective value of the first Frank–Wolfe iterate for $(\bar{v},\ \bar{y},\ \bar{t})$ gives an upper bound on the current duality gap.

6. A new algorithm for $(SCSCLP)$.

In this section, we give a generic successive quadratic programming algorithm for $(SCSCLP)$.

```
Algorithm $\mathcal{A}$ $(E, F, G, H, a(t), b(t), c(t), g(t), h(t), T, \beta)$.
Let $k = 0$. Let $d$ be the current duality gap initially set to infinity.
Let $(v^k,\ y^k,\ t^k)$ be a feasible solution to $QP(|P^0|)$. Let $P^0$ be a partition on $[0,\ T]$,
such that $a(t)$, $c(t)$, and $h(t)$ are piecewise linear with $P^0$ and $b(t)$ and $g(t)$ are
piecewise constant with $P^0$.
while $d > \beta$ do
    1. Calculate a KKT point of $QP(|P^k|)$ which has an equal or better solution
       value than $(v^k,\ y^k,\ t^k)$.
```

2. Recursively remove redundant intervals in P^k as follows.
   Apply Procedure $PURIFY$ to all pairs of consecutive breakpoints in $D_1^{P^k}$. Let $(\tilde{v}^k,\ \tilde{y}^k,\ \tilde{t}^k)$ be the resulting solution and let Q be the resulting partition. If $(\tilde{v}^k,\ \tilde{y}^k,\ \tilde{t}^k)$ is not a KKT point of $QP(|Q|)$, let $(v^k,\ y^k,\ t^k) = (\tilde{v}^k,\ \tilde{y}^k,\ \tilde{t}^k)$ and $P^k = Q$ and go to Step 1. Otherwise, we denote the resulting purified partition as $\tilde{P}^k = \{t_0, t_1, \dots, t_p\}$.
3. Double the number of intervals. Define P^{k+1} as
   $$P^{k+1} = \{t_0, t_0, t_1, t_1, \dots, t_i, t_i, t_i, \dots, t_p, t_p\},$$
   where each breakpoint in D has two duplicates and all the other breakpoints have only one duplicate. Construct a feasible solution $(\bar{v}^{k+1},\ \bar{y}^{k+1},\ \bar{t}^{k+1})$ for $QP(|P^{k+1}|)$ as in (18).
4. Calculate the current duality gap d. If the solution value of $(\tilde{v}^k,\ \tilde{y}^k,\ \tilde{t}^k)$ is the same as the optimal value of $AP(\tilde{P}^k)$, stop the algorithm. Otherwise go to Step 5.
5. Get a strictly improved solution $(v^{k+1},\ y^{k+1},\ t^{k+1})$ from $(\bar{v}^{k+1},\ \bar{y}^{k+1},\ \bar{t}^{k+1})$ for $QP(|P^{k+1}|)$.
6. Let $k = k + 1$.

end while

Remarks.

1. In Step 1 of Algorithm $\mathcal{A}$, we can use the Frank–Wolfe method or general matrix-splitting algorithms to compute a KKT point of $QP(|P^k|)$.
2. Algorithm $\mathcal{A}$ will not loop between Step 1 and Step 2 forever, because every time Algorithm $\mathcal{A}$ goes from Step 2 to Step 1, the cardinality of P^k is reduced at least by 1.
3. In Step 4 of Algorithm $\mathcal{A}$, we can let $d = V(QP(|\tilde{P}^k|)) - V(AP(\tilde{P}^k))$. We can also let d be the negative objective value of the first Frank–Wolfe iterate for $(\bar{v},\ \bar{y},\ \bar{t})$ and so, instead of checking whether the solution value of $(\tilde{v}^k,\ \tilde{y}^k,\ \tilde{t}^k)$ is the same as the optimal value of $AP(\tilde{P}^k)$, we can check whether the objective value of the first Frank–Wolfe iterate for $(\bar{v},\ \bar{y},\ \bar{t})$ is zero.
4. In Step 5 of Algorithm $\mathcal{A}$, we can use the direction constructed in section 5.3 (cf. $(v^\epsilon,\ y^\epsilon,\ t^\epsilon)$). We can also use the Frank–Wolfe method or general matrix-splitting algorithms to find a descent direction for $(\bar{v}^{k+1},\ \bar{y}^{k+1},\ \bar{t}^{k+1})$. By Theorem 3, we are guaranteed to find a descent direction.

7. Convergence of the new algorithm. In this section we prove that the new algorithm converges. We first describe the argument we will use to show the convergence informally. We use the Frank–Wolfe method or general matrix-splitting algorithms to compute a series of KKT points to a series of generally nonconvex quadratic programs. These KKT points have nondecreasing solution values. By Corollary 3, we can detect whether a KKT point gives an optimal solution to $(SCSCLP)$. If it does, we terminate the algorithm. If not, by Theorem 3, we can find a new solution with approximately twice as many constant control pieces as the current solution but with a strictly improved cost. Since there is only a finite number of different solution values for the KKT points of every quadratic program constructed, and there is an upper bound on the size of the quadratic programs we encounter (see Corollary 4 below), a finite convergence result follows readily. Based on the primal solution, we can compute an optimal dual solution for $(SCSCLP^*)$.

Contrary to the convergence analysis of a variety of algorithms for $(SCLP)$, we do not need to let the norm of the maximal length interval in the discretization tend to zero (as in Pullan [41]). Moreover, neither do we need the explicit knowledge of all the extreme points of a certain set of finite-dimensional linear programs (as in Anderson and Nash [2]). Most importantly, we prove the absence of a duality gap result as a byproduct of the new algorithm, even when there is no optimal solution for $(SCSCLP)$.

In the following, we give upper bounds on the cardinality of $\tilde{P}^k$, the purified partition in Step 2 of Algorithm $\mathcal{A}$. Since by Lemma 4 and Corollary 2 we know that the total number of zero-length intervals in $\tilde{P}^k$ is at most $2|D_1^{\tilde{P}^k}|$, we need only to bound the number of positive-length intervals in $\tilde{P}^k$. We map each positive-length interval of $\tilde{P}^k$ to an extreme point of a certain set of linear programs and then show that the mapping is injective. Before doing so, we give some more notation and several useful lemmas.

Let t_l and t_m be two consecutive breakpoints in $D_1^{\tilde{P}^k}$. By definition, $a(t)$, $c(t)$, and $h(t)$ are linear and $b(t)$ and $g(t)$ are constant on $[t_l,\ t_m)$. Let $[\tilde{t}_{i-1},\ \tilde{t}_i]$ and $[\tilde{t}_i,\ \tilde{t}_{i+1}]$ be two adjacent positive-length intervals in partition $\tilde{P}^k$ such that $[\tilde{t}_{i-1},\ \tilde{t}_{i+1}] \subseteq [t_l,\ t_m]$. Let $\Delta t_i = \tilde{t}_i - \tilde{t}_{i-1}$ and $\Delta t_{i+1} = \tilde{t}_{i+1} - \tilde{t}_i$. We have $\Delta t_i > 0$ and $\Delta t_{i+1} > 0$ by assumption. Let $(\tilde{v}^k, \tilde{y}^k, \tilde{t}^k)$ be the resulting solution in Step 2 of Algorithm $\mathcal{A}$. Let J_i be the set of indices of the constraints in $F\tilde{y}^k(\tilde{t}_i) \le h(\tilde{t}_i)$ that are binding. Let

$$\tilde{u}^k(\tilde{t}_{i-1}+) = \frac{\tilde{v}^k(\tilde{t}_i)}{\Delta t_i}$$

and

$$\tilde{u}^k(\tilde{t}_i+) = \frac{\tilde{v}^k(\tilde{t}_{i+1})}{\Delta t_{i+1}}.$$

It is obvious that $(\tilde{u}^k(\tilde{t}_i+),\ \frac{\tilde{y}^k(\tilde{t}_{i+1}) - \tilde{y}^k(\tilde{t}_i)}{\Delta t_{i+1}})$ is a feasible solution to the following linear system:

$$\begin{aligned}(SYS_{J_i})\ & G\tilde{u}^k(\tilde{t}_i+) + E\frac{\tilde{y}^k(\tilde{t}_{i+1}) - \tilde{y}^k(\tilde{t}_i)}{\Delta t_{i+1}} = \dot{a}(\tilde{t}_i),\\ & H\tilde{u}^k(\tilde{t}_i+) \le b(\tilde{t}_i),\\ & \left(F\frac{\tilde{y}^k(\tilde{t}_{i+1}) - \tilde{y}^k(\tilde{t}_i)}{\Delta t_{i+1}}\right)_{J_i} \le \dot{h}(\tilde{t}_i),\\ & \tilde{u}^k(\tilde{t}_i+) \ge 0.\end{aligned}$$

By introducing new variables, we can eliminate $\frac{\tilde{y}^k(\tilde{t}_{i+1}) - \tilde{y}^k(\tilde{t}_i)}{\Delta t_{i+1}}$ in (SYS_{J_i}) and transform (SYS_{J_i}) into the following linear system:

$$\begin{aligned}(SYS1_{J_i})\ & G\tilde{u}^k(\tilde{t}_i+) + E(w_{i+1} - w_i) = \dot{a}(\tilde{t}_i),\\ & H\tilde{u}^k(\tilde{t}_i+) + \tilde{z}^k(\tilde{t}_i+) = b(\tilde{t}_i),\\ & (F(w_{i+1} - w_i))_{J_i} + x = \dot{h}(\tilde{t}_i),\\ & x \ge 0, w_{i+1} \ge 0, w_i \ge 0, \tilde{u}^k(\tilde{t}_i+) \ge 0, \tilde{z}^k(\tilde{t}_i+) \ge 0.\end{aligned}$$

Every extreme point of the linear program defined by maximizing some linear function over $(SYS1_{J_i})$ defines a unique feasible solution to (SYS_{J_i}), which is called a generalized extreme point for (SYS_{J_i}). Every extreme ray of this linear program defines a unique ray to (SYS_{J_i}), which is called a generalized extreme ray for (SYS_{J_i}). Since this is a feasible finite-dimensional linear program in standard form, the resolution theorem applies. After translating the result into variables in (SYS_{J_i}), we have the following analogue of the resolution theorem for (SYS_{J_i}).

LEMMA 5. *Every feasible solution of (SYS_{J_i}) can be written as the sum of a convex combination of the generalized extreme points of (SYS_{J_i}) and a linear combination (with nonnegative coefficients) of generalized extreme rays to (SYS_{J_i}).*

By Lemma 5, we have

$$\begin{aligned} \tilde{u}^k(\tilde{t}_i+) &= \sum_{j=1}^{k^{(i)}} \lambda_j^{(i)} s_j^{(i)} + \sum_{j=1}^{q^{(i)}} \mu_j^{(i)} r_j^{(i)}, \\ \frac{\tilde{y}^k(\tilde{t}_{i+1}) - \tilde{y}^k(\tilde{t}_i)}{\Delta t_{i+1}} &= \sum_{j=1}^{k^{(i)}} \lambda_j^{(i)} \bar{s}_j^{(i)} + \sum_{j=1}^{q^{(i)}} \mu_j^{(i)} \bar{r}_j^{(i)} \end{aligned} \tag{30}$$

for some positive $k^{(i)} \geq 1$ and nonnegative $q^{(i)} \geq 0$, where $\lambda_j^{(i)} > 0$, $\sum_{j=1}^{k^{(i)}} \lambda_j^{(i)} = 1$, and $\mu_j^{(i)} > 0$, the $(s_j^{(i)},\ \bar{s}_j^{(i)})$ are generalized extreme points to system (SYS_{J_i}), and the $(r_j^{(i)},\ \bar{r}_j^{(i)})$ are generalized extreme rays to system (SYS_{J_i}). Without loss of generality, assume that we have sorted $(s_j^{(i)},\ \bar{s}_j^{(i)})$ in the following order:

$$\dot{c}(\tilde{t}_i)' s_j^{(i)} - g(t_l+)' \bar{s}_j^{(i)} \geq \dot{c}(\tilde{t}_i)' s_{j+1}^{(i)} - g(t_l+)' \bar{s}_{j+1}^{(i)} \qquad \text{for all } j. \tag{31}$$

We have the following result on $(\tilde{v}^k,\ \tilde{y}^k,\ \tilde{t}^k)$.

LEMMA 6.

$$V(QP(|\tilde{P}^k|), (\tilde{v}^k,\ \tilde{y}^k,\ \tilde{t}^k)) \leq V(QP(|P^k|), (v^k,\ y^k,\ t^k)).$$

Proof. Since Procedure *PURIFY* does not increase the solution value of the current solution, the result immediately follows. □

LEMMA 7. *Suppose* (30) *and* (31) *hold for $\tilde{u}^k(\tilde{t}_{i-1}+)$ and $\tilde{u}^k(\tilde{t}_i+)$. Furthermore, suppose $[\tilde{t}_{i-1},\ \tilde{t}_i]$ is not the first positive-length interval that resides on $[t_l,\ t_m]$. Then we have*

$$\dot{c}(\tilde{t}_i)' s_1^{(i-1)} - g(t_l+)' \bar{s}_1^{(i-1)} > \dot{c}(\tilde{t}_i)' s_1^{(i)} - g(t_l+)' \bar{s}_1^{(i)}$$

for the two adjacent positive-length intervals $[\tilde{t}_{i-1},\ \tilde{t}_i]$ and $[\tilde{t}_i,\ \tilde{t}_{i+1}]$ that reside on $[t_l,\ t_m]$.

Proof. We first show that

$$\dot{c}(\tilde{t}_i)' r_j^{(i)} - g(t_l+)' \bar{r}_j^{(i)} \leq 0 \tag{32}$$

for every $j \leq q^{(i)}$ without assuming that $[\tilde{t}_{i-1},\ \tilde{t}_i]$ is not the first positive-length interval that resides on $[t_l,\ t_m]$.

Let $\tau \in (0,\ 1)$. Suppose

$$\tilde{u}^k(\tilde{t}_i+) = \tau u_1 + (1-\tau) u_2$$

and

$$\frac{\tilde{y}^k(\tilde{t}_{i+1}) - \tilde{y}^k(\tilde{t}_i)}{\Delta t_{i+1}} = \tau y_1 + (1-\tau) y_2,$$

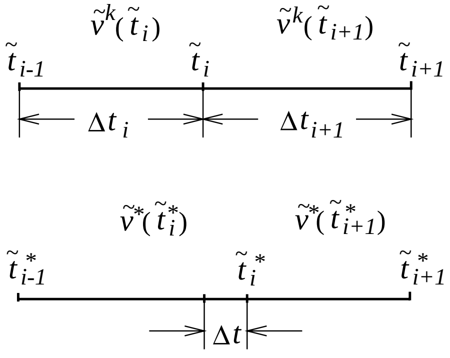


FIG. 3. *Perturbation of the solution.*

where $(u_1,\ y_1)$ and $(u_2,\ y_2)$ are two feasible solutions for (SYS_{J_i}). Let γ be the largest scalar in $(0,\ \tau\Delta t_{i+1}]$ such that $F(\tilde y^k(\tilde t_i) + \gamma y_1) \le h(\tilde t_i)$. Such a γ exists by virtue of the feasibility of $(u_1,\ y_1)$ and $(u_2,\ y_2)$ to system (SYS_{J_i}). For any $\Delta t \in (0,\ \gamma)$, we consider the following perturbation of $(\tilde v^k, \tilde y^k, \tilde t^k)$, as shown in Figure 3:

$$\tilde t_j^* = \begin{cases} \tilde t_i + \Delta t & \text{if } j = i,\\ \tilde t_j & \text{otherwise,} \end{cases}$$

$$\tilde v^*(\tilde t_j^*) = \begin{cases} \tilde v^k(\tilde t_i) + u_1\Delta t & \text{if } j = i,\\ \tilde v^k(\tilde t_j) - u_1\Delta t & \text{if } j = i+1,\\ \tilde v^k(\tilde t_j) & \text{otherwise,} \end{cases}$$

$$\tilde y^*(\tilde t_j^*) = \begin{cases} \tilde y^k(\tilde t_i) + y_1\Delta t & \text{if } j = i,\\ \tilde y^k(\tilde t_j) & \text{otherwise.} \end{cases}$$

We can easily check the feasibility of $(\tilde v^*,\ \tilde y^*,\ \tilde t^*)$ to $QP(|\tilde P|)$. So

$$\begin{aligned}
&\sum_{j=1}^{p} \tilde v^*(\tilde t_j^*)' c\left(\frac{\tilde t_{i-1}^* + \tilde t_i^*}{2}\right) - \sum_{j=1}^{p} \tilde v^k(\tilde t_j)' c\left(\frac{\tilde t_{i-1} + \tilde t_i}{2}\right)\\
&= \left(\frac{c(\tilde t_{i-1}) + c(\tilde t_i^*)}{2}\right)' \tilde v^*(\tilde t_i^*) + \left(\frac{c(\tilde t_i^*) + c(\tilde t_{i+1})}{2}\right)' \tilde v^*(\tilde t_{i+1}^*)\\
&\quad - \left(\frac{c(\tilde t_{i-1}) + c(\tilde t_i)}{2}\right)' \tilde v^k(\tilde t_i) - \left(\frac{c(\tilde t_i) + c(\tilde t_{i+1})}{2}\right)' \tilde v^k(\tilde t_{i+1})\\
&= \left(\frac{c(\tilde t_i^*) - c(\tilde t_i)}{2}\right)' \tilde v^k(\tilde t_i) + \left(\frac{c(\tilde t_i^*) - c(\tilde t_i)}{2}\right)' \tilde v^k(\tilde t_{i+1})\\
&\quad - \frac{\Delta t}{2}(\Delta t_i + \Delta t_{i+1})\dot c(\tilde t_i)' u_1\\
&= \frac{\Delta t \Delta t_i}{2}(\dot c(\tilde t_i)' \tilde u^k(\tilde t_i -) - \dot c(\tilde t_i)' u_1)\\
&\quad + \frac{\Delta t \Delta t_{i+1}}{2}(\dot c(\tilde t_i)' \tilde u^k(\tilde t_i +) - \dot c(\tilde t_i)' u_1).
\end{aligned} \tag{33}$$

Also,

$$\sum_{j=1}^{p} \frac{\tilde t_i^* - \tilde t_{i-1}^*}{2}(\tilde y^*(\tilde t_i^*) + \tilde y^*(\tilde t_{i-1}^*))' g(\tilde t_{i-1}^* +)$$

$$
\begin{aligned}
&-\sum_{j=1}^{p}\frac{\tilde{t}_i-\tilde{t}_{i-1}}{2}(\tilde{y}^k(\tilde{t}_i)+\tilde{y}^k(\tilde{t}_{i-1}))'g(\tilde{t}_{i-1}+)\\
&=\frac{\tilde{t}^*_i-\tilde{t}^*_{i-1}}{2}(\tilde{y}^*(\tilde{t}^*_i)+\tilde{y}^*(\tilde{t}^*_{i-1}))'g(\tilde{t}^*_{i-1}+)\\
&\quad+\frac{\tilde{t}^*_{i+1}-\tilde{t}^*_i}{2}(\tilde{y}^*(\tilde{t}^*_{i+1})+\tilde{y}^*(\tilde{t}^*_i))'g(\tilde{t}^*_i+)\\
&\quad-\frac{\tilde{t}_i-\tilde{t}_{i-1}}{2}(\tilde{y}^k(\tilde{t}_i)+\tilde{y}^k(\tilde{t}_{i-1}))'g(\tilde{t}_{i-1}+)\\
&\quad-\frac{\tilde{t}_{i+1}-\tilde{t}_i}{2}(\tilde{y}^k(\tilde{t}_{i+1})+\tilde{y}^k(\tilde{t}_i))'g(\tilde{t}_i+)\\
&=\frac{\Delta t_i+\Delta t}{2}(\tilde{y}^k(\tilde{t}_i)+\tilde{y}^k(\tilde{t}_{i-1})+\Delta t y_1)'g(\tilde{t}_{i-1}+)\\
&\quad+\frac{\Delta t_{i+1}-\Delta t}{2}(\tilde{y}^k(\tilde{t}_{i+1})+\tilde{y}^k(\tilde{t}_i)+\Delta t y_1)'g(\tilde{t}_i+)\\
&\quad-\frac{\Delta t_i}{2}(\tilde{y}^k(\tilde{t}_i)+\tilde{y}^k(\tilde{t}_{i-1}))'g(\tilde{t}_{i-1}+)-\frac{\Delta t_{i+1}}{2}(\tilde{y}^k(\tilde{t}_{i+1})+\tilde{y}^k(\tilde{t}_i))'g(\tilde{t}_i+)\\
&=\frac{\Delta t}{2}(\tilde{y}^k(\tilde{t}_{i-1})-\tilde{y}^k(\tilde{t}_{i+1})+(\Delta t_i+\Delta t_{i+1})y_1)'g(t_l+). \qquad (34)
\end{aligned}
$$

Combining (33) and (34), we derive

$$
\begin{aligned}
&V(QP(|\tilde{P}^k|),(\tilde{v}^*,\ \tilde{y}^*,\ \tilde{t}^*))-V(QP(|\tilde{P}^k|),(\tilde{v}^k,\ \tilde{y}^k,\ \tilde{t}^k))\\
&=\frac{\Delta t\Delta t_i}{2}\left(\dot{c}(\tilde{t}_i)'\tilde{u}^k(\tilde{t}_i-)-g(\tilde{t}_i+)'\frac{\tilde{y}^k(\tilde{t}_i)-\tilde{y}^k(\tilde{t}_{i-1})}{\Delta t_i}-(\dot{c}(\tilde{t}_i)'u_1-y_1'g(t_l+))\right)\\
&\quad+\frac{\Delta t\Delta t_{i+1}}{2}\left(\dot{c}(\tilde{t}_i)'\tilde{u}^k(\tilde{t}_i+)-g(\tilde{t}_i+)'\frac{\tilde{y}^k(\tilde{t}_{i+1})-\tilde{y}^k(\tilde{t}_i)}{\Delta t_{i+1}}-(\dot{c}(\tilde{t}_i)'u_1-y_1'g(t_l+))\right). \qquad (35)
\end{aligned}
$$

By the definition of a KKT point and the discussion following it in section 3.1, a feasible solution to $QP(|P|)$ is a KKT point if and only if there is no feasible descent direction for this solution. Hence

$$
V(QP(|\tilde{P}^k|),(\tilde{v}^*,\ \tilde{y}^*,\ \tilde{t}^*))-V(QP(|\tilde{P}^k|),(\tilde{v}^k,\ \tilde{y}^k,\ \tilde{t}^k))\geq 0. \qquad (36)
$$

Thus (35) implies that $\dot{c}(\tilde{t}_i)'u_1-y_1'g(t_l+)$ is uniformly bounded from above for any possible choice of (u_1, y_1).

For any $\bar{j}\leq q^{(i)}$ and any $\epsilon\in(0\ ,1)$, we have

$$
\begin{aligned}
\tilde{u}^k(\tilde{t}_i+)&=\epsilon\lambda_1^{(i)}(s_1^{(i)}+\frac{\mu_{\bar{j}}^{(i)}}{\epsilon\lambda_1^{(i)}}r_{\bar{j}}^{(i)})\\
&\quad+(1-\epsilon\lambda_1^{(i)})\left(\sum_{j=2}^{k^{(i)}}\frac{\lambda_j^{(i)}}{1-\epsilon\lambda_1^{(i)}}s_j^{(i)}+\frac{\lambda_1^{(i)}(1-\epsilon)}{1-\epsilon\lambda_1^{(i)}}s_1^{(1)}+\sum_{j=1,j\neq\bar{j}}^{q^{(i)}}\frac{\mu_j^{(i)}}{1-\epsilon\lambda_1^{(i)}}r_j^{(i)}\right)
\end{aligned}
$$

and

$$
\frac{\tilde{y}^k(\tilde{t}_{i+1})-\tilde{y}^k(\tilde{t}_i)}{\Delta t_{i+1}}
$$

$$
\begin{aligned}
&= \epsilon\lambda_1^{(i)}\left(\bar{s}_1^{(i)} + \frac{\mu_{\bar{j}}^{(i)}}{\epsilon\lambda_1^{(i)}}\bar{r}_{\bar{j}}^{(i)}\right)\\
&\quad + (1-\epsilon\lambda_1^{(i)})\left(\sum_{j=2}^{k^{(i)}}\frac{\lambda_j^{(i)}}{1-\epsilon\lambda_1^{(i)}}\bar{s}_j^{(i)} + \frac{\lambda_1^{(i)}(1-\epsilon)}{1-\epsilon\lambda_1^{(i)}}\bar{s}_1^{(1)} + \sum_{j=1,j\neq\bar{j}}^{q^{(i)}}\frac{\mu_j^{(i)}}{1-\epsilon\lambda_1^{(i)}}\bar{r}_j^{(i)}\right).
\end{aligned}
$$

By letting

$$
u_1 = s_1^{(i)} + \frac{\mu_{\bar{j}}^{(i)}}{\epsilon\lambda_1^{(i)}} r_{\bar{j}}^{(i)}, \qquad y_1 = \bar{s}_1^{(i)} + \frac{\mu_{\bar{j}}^{(i)}}{\epsilon\lambda_1^{(i)}}\bar{r}_{\bar{j}}^{(i)},
$$

and letting ϵ tend to zero, the above boundedness result on $\dot{c}(\tilde{t}_i)'u_1^{(i)} - y_1' g(t_l+)$ implies (32). Since $[t_{i-1},\ t_i]$ is not the first positive-length interval that resides on $[t_l,\ t_m]$, we can similarly have

$$
\dot{c}(\tilde{t}_i)' r_j^{(i-1)} - g(t_l+)'\bar{r}_j^{(i-1)} \leq 0 \quad \text{for all } j.
$$

These together with (30) and (31) give

$$
\dot{c}(\tilde{t}_i)' s_1^{(i-1)} - g(t_l+)'\bar{s}_1^{(i-1)} \geq \dot{c}(\tilde{t}_i)'\tilde{u}^k(\tilde{t}_{i-1}+) - g(\tilde{t}_i+)'\frac{\tilde{y}^k(\tilde{t}_i) - \tilde{y}^k(\tilde{t}_{i-1})}{\Delta t_i}. \tag{37}
$$

Similarly,

$$
\dot{c}(\tilde{t}_i)' s_1^{(i)} - g(t_l+)'\bar{s}_1^{(i)} \geq \dot{c}(\tilde{t}_i)'\tilde{u}^k(\tilde{t}_i+) - g(\tilde{t}_i+)'\frac{\tilde{y}^k(\tilde{t}_{i+1}) - \tilde{y}^k(\tilde{t}_i)}{\Delta t_{i+1}}. \tag{38}
$$

Since $\tilde{P}$ is a purified partition, by Procedure *PURIFY*, the opposite of (11) holds, which is equivalent to

$$
\dot{c}(\tilde{t}_i)'\tilde{u}^k(\tilde{t}_{i-1}+) - g(\tilde{t}_i+)'\frac{\tilde{y}^k(\tilde{t}_i) - \tilde{y}^k(\tilde{t}_{i-1})}{\Delta t_i} > \dot{c}(\tilde{t}_i)'\tilde{u}^k(\tilde{t}_i+) - g(\tilde{t}_i+)'\frac{\tilde{y}^k(\tilde{t}_{i+1}) - \tilde{y}^k(\tilde{t}_i)}{\Delta t_{i+1}}. \tag{39}
$$

Now, suppose

$$
\dot{c}(\tilde{t}_i)' s_1^{(i-1)} - g(t_l+)'\bar{s}_1^{(i-1)} \leq \dot{c}(\tilde{t}_i)' s_1^{(i)} - g(t_l+)'\bar{s}_1^{(i)}.
$$

By (37) and (39), we have

$$
\dot{c}(\tilde{t}_i)'\tilde{u}^k(\tilde{t}_i+) - g(\tilde{t}_i+)'\frac{\tilde{y}^k(\tilde{t}_{i+1}) - \tilde{y}^k(\tilde{t}_i)}{\Delta t_{i+1}} < \dot{c}(\tilde{t}_i)' s_1^{(i)} - g(t_l+)'\bar{s}_1^{(i)}
$$

and

$$
\dot{c}(\tilde{t}_i)'\tilde{u}^k(\tilde{t}_{i-1}+) - g(\tilde{t}_i+)'\frac{\tilde{y}^k(\tilde{t}_i) - \tilde{y}^k(\tilde{t}_{i-1})}{\Delta t_i} \leq \dot{c}(\tilde{t}_i)' s_1^{(i)} - g(t_l+)'\bar{s}_1^{(i)}.
$$

Let $u_1 = s_1^{(i)}$ and $y_1 = \bar{s}_1^{(i)}$. Then the above relationship together with (35) gives

$$
V(QP(|\tilde{P}^k|), (\tilde{v}^*,\ \tilde{y}^*,\ \tilde{t}^*)) - V(QP(|\tilde{P}^k|), (\tilde{v}^k,\ \tilde{y}^k,\ \tilde{t}^k)) < 0,
$$

which contradicts that $(\tilde{v}^k,\ \tilde{y}^k,\ \tilde{t}^k)$ is a KKT point for $QP(|\tilde{P}^k|)$ (cf. (36)). □

Since $\dot{c}(\tilde{t}_i)$ is a constant vector over $[t_l,\ t_m]$, as a consequence of Lemma 7, every nonzero-length interval that resides on $[t_l,\ t_m]$ (except the first nonzero-length interval) corresponds to a different generalized extreme point of some system (SYS_{J_i}). Since only a finite number of different systems (SYS_{J_i}) exists, and for each (SYS_{J_i}) there is a finite number of generalized extreme points, we see there is only a finite number of nonzero-length intervals that reside on $[t_l,\ t_m]$. Since the number of zero-length intervals that reside on $[t_l,\ t_m]$ is at most two (one on each end of $[t_l,\ t_m]$), there is also a finite number of breakpoints in $[t_l,\ t_m]$. Thus we have the following corollary.

COROLLARY 4. *There is a finite number of breakpoints in* $\tilde{P}^k$.

There is only a finite number of different solution values for all the KKT points of $QP(|P|)$, as shown in the following lemma.

LEMMA 8. *The KKT points for* $QP(|P|)$ *are the union of a finite number of connected sets. Over each connected component of KKT points of* $QP(|P|)$*, the objective value is a constant. Furthermore, the number of connected sets is bounded from above by a number that depends on* $|P|$ *only.*

Proof. It is easily seen that a solution to $QP(|P|)$ is a KKT point of $QP(|P|)$ if and only if it is a solution to a feasible symmetric affine variational inequality problem whose dimension depends only on $|P|$ (cf. section 3.1). The lemma now follows directly from Lemma 3.1 of Luo and Tseng [31]. □

We now present the main convergence result of the paper.

THEOREM 5. *Algorithm* $\mathcal{A}$ *will terminate after a finite number of iterations.*

Proof. Suppose Algorithm $\mathcal{A}$ does not terminate after a finite number of iterations. It is guaranteed by Theorem 3 that Step 4 of Algorithm $\mathcal{A}$ would produce a strictly improved solution, and thus every iteration of Algorithm $\mathcal{A}$ would give a KKT point of a certain $QP(|P|)$ that has a strictly better solution value. By Lemma 8, the KKT points generated by $QP(|P|)$ should lie on a different connected KKT points component of $QP(|P|)$ for every $|P|$. This means that the cardinality of P is unbounded and contradicts Corollary 4. □

8. New structural and duality results. As a result of Algorithm $\mathcal{A}$ and Theorem 5, we have the following new structural result for $(SCSCLP)$.

THEOREM 6. *Under Assumption* 1*, Algorithm* $\mathcal{A}$ *terminates with a solution to* $QP(|P|)$ *for some* P *that gives the optimal objective value of* $(SCSCLP)$ *and can be closely approximated by a series of piecewise constant controls for* $(SCSCLP)$*. When the solution set for* $(SCSCLP)$ *is bounded and* E *is an identity matrix, Algorithm* $\mathcal{A}$ *terminates with a piecewise constant optimal control with partition* P *such that* $t_i \neq t_{i-1}$ *for all* i*. Furthermore, over each interval* $[t_{i-1},\ t_i)$*,* $(u(t_i+), \frac{y(t_{i+1})-y(t_i)}{t_{i+1}-t_i})$ *is a convex combination of the generalized extreme points of linear system* (SYS_{J_i})*, where* J_i *is a subset of* $\{1, \ldots, n_2\}$.

Proof. The first part of the theorem is a direct consequence of Theorem 5. The second part of the theorem follows from Lemma 5 and the remark following the proof of Lemma 1. □

We remark that when the solution set for $(SCSCLP)$ is unbounded, it is possible that the optimal solution value is not attained. We next derive the following new duality result for $(SCSCLP)$.

THEOREM 7. *Under Assumption* 1*, there is no duality gap between* $(SCSCLP)$ *and* $(SCSCLP^*)$*. There always exists an optimal solution for* $(SCSCLP^*)$ *that is piecewise linear. Furthermore, there exists a bounded measurable optimal solution for* $(SCSCLP)$ *if and only if Algorithm* $\mathcal{A}$ *terminates with such a solution.*

Proof. The first part of the theorem is a direct consequence of Theorem 5.

Denote $\tilde{P}^k$ as the final purified partition when Algorithm $\mathcal{A}$ terminates. To prove the second part of the theorem, we first show that the zero-length intervals in $\tilde{P}^k$ can be eliminated in the dual problem $AP^*(P)$. Let $[t_{i-1},\ t_i]$ be a zero-length interval that resides on $[t_l,\ t_m]$, where t_l and t_m are two consecutive breakpoints in $D_1^{\tilde{P}^k}$. By Lemma 4, the zero-length intervals can be located only at the breakpoints in $D_1^{\tilde{P}^k}$. We assume $t_i = t_m$ (the case $t_{i-1} = t_l$ can be treated similarly).

Let $(\hat{\pi},\ \hat{\eta},\ \hat{\xi})$ be an optimal solution for $AP^*(\tilde{P}^k)$. Then we can construct a new solution $(\tilde{\pi},\ \tilde{\eta},\ \tilde{\xi})$ for $AP^*(\tilde{P}^k)$ in the following way. Let $(\tilde{\pi},\ \tilde{\eta},\ \tilde{\xi})$ equal $(\hat{\pi},\ \hat{\eta},\ \hat{\xi})$ except

$$\begin{aligned}
\tilde{\pi}(t_{i-1}+) &= \hat{\pi}(t_{i-1}-), \quad \tilde{\pi}(t_i-) = \hat{\pi}(t_{i-1}-),\\
\tilde{\eta}(t_{i-1}+) &= \hat{\eta}(t_{i-1}-), \quad \tilde{\eta}(t_i-) = \hat{\eta}(t_{i-1}-),\\
\tilde{\xi}\left(\frac{t_{i-1}+t_i}{2}\right) &= 0, \qquad \tilde{\xi}(t_{i-1}) = 0,\\
\tilde{\xi}(t_i) &= \hat{\xi}\left(\frac{t_{i-1}+t_i}{2}\right) + \hat{\xi}(t_{i-1}) + \hat{\xi}(t_i).
\end{aligned}$$

It is easy to check the feasibility of $(\tilde{\pi},\ \tilde{\eta},\ \tilde{\xi})$. It is a fact that $(\tilde{\pi},\ \tilde{\eta},\ \tilde{\xi})$ and $(\hat{\pi},\ \hat{\eta},\ \hat{\xi})$ have the same solution value in $AP^*(\tilde{P}^k)$. Let $\bar{P}$ be $\tilde{P}^k \setminus \{t_{i-1}\}$. By eliminating the elements $\tilde{\pi}(t_{i-1}-)$, $\tilde{\pi}(t_{i-1}+)$, $\tilde{\eta}(t_{i-1}-)$, $\tilde{\eta}(t_{i-1}+)$, $\tilde{\xi}(t_{i-1})$, and $\tilde{\xi}\left(\frac{t_{i-1}+t_i}{2}\right)$ from $(\tilde{\pi},\ \tilde{\eta},\ \tilde{\xi})$, we can get a feasible solution $(\bar{\pi},\ \bar{\eta},\ \bar{\xi})$ for $AP(\bar{P})$. Also, $(\bar{\pi},\ \bar{\eta},\ \bar{\xi})$ has the same solution value as $(\tilde{\pi},\ \tilde{\eta},\ \tilde{\xi})$.

By repeating this process, we can eliminate all the zero-length intervals in $\tilde{P}^k$ and define a feasible solution for $AP^*(P)$ from the resulting partition P. From this feasible solution, we can construct an optimal solution for $(SCSCLP^*)$ that is piecewise linear. This proves the second part of the theorem.

One direction of the third part of the theorem is quite obvious. The other direction (i.e., if there exists a bounded measurable optimal solution for $(SCSCLP)$, then Algorithm $\mathcal{A}$ will find such a solution) can be shown as follows. Let the bounded measurable solution $(u(t),\ y(t))$ be optimal for $(SCSCLP)$. By the second part of the theorem, there always exists an optimal solution $(\pi(t),\ \eta(t),\ \xi(t))$ for $(SCSCLP^*)$ that is piecewise linear with partition P (defined by removing all the zero-length intervals from $\tilde{P}^K$). By Corollary 1, the complementary slackness condition (5) is satisfied. Let $\bar{u}(t)$ be the piecewise constant extensions of $u(t_0+),\ u(t_1+), \dots, u(t_{p-1}+)$. Let $\bar{y}(t)$ be the piecewise-linear extension of $y(t_0+), y(t_1-)\ y(t_1+), \dots, y(t_{p-1}+), y(t_p-)$. The solution $(\bar{u}(t),\ \bar{y}(t))$ is a feasible solution for $(SCSCLP)$ which together with $(\pi(t),\ \eta(t),\ \xi(t))$ satisfies (5). Therefore, Corollary 1 again, $(\bar{u}(t),\ \bar{y}(t))$ is optimal for $(SCSCLP)$. □

9. Computational results. Algorithm $\mathcal{A}$ has been implemented and tested on a Sparc 10/41. The program is written in *C*. We used the academic version of *LOQO* Version 1.08 by Vanderbei [48]. We call its subroutines to solve intermediate linear programming and quadratic programming subproblems.

The implementation of Algorithm $\mathcal{A}$ consists of four modules: the input data processing module, the output module, the successive quadratic programming module, and the lower bound module. The successive quadratic programming module uses the Frank–Wolfe method to iteratively solve a sequence of quadratic programs, as outlined in Algorithm $\mathcal{A}$. The lower bound module uses the partition generated by

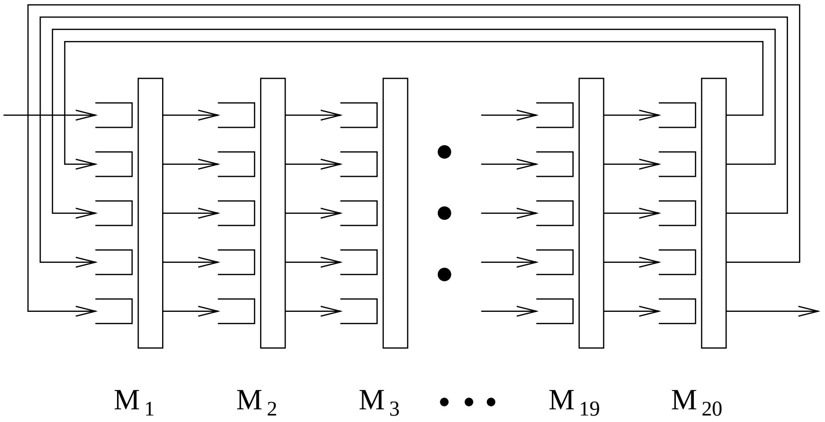


FIG. 4. *The reentrant line for the example.*

the successive quadratic programming module to calculate a dual feasible solution for the problem.

We next give a numerical example that arises in manufacturing systems. The example is a reentrant line, as in Kumar [25]. A reentrant line is a multiclass queuing network with fixed routing.

The reentrant line we consider is shown in Figure 4. We have from left to right n stations (in Figure 4, we have 20 stations), and each station services 5 different classes of customers. There are $5n$ classes of customers in total. Class i customers will be served at machine $\lfloor (i-1)/n + 1 \rfloor$. After class i customer finishes service, it will become class $i+1$ customer if $i < 5n$ and exit the system otherwise. For this system, we assume the exogenous arrival rate for class 1 customer is 1 and is zero for all other classes. We generate randomly the mean service time, the cost per unit time, and the initial number of customers for each class of customers. Our objective is to find an optimal control policy (involving both routing and sequencing decisions) that minimizes the cumulated cost of queuing over a fixed time horizon $[0,\ T]$.

We can formulate the problem as an $(SCSCLP)$. Let $y_i(t)$ be the queue length of class i customers at time t. If class i customers are served at machine j, we let $u_i(t)$ be the proportion of machine capacity of machine j that is devoted to class i customers at time t. The G matrix of the $(SCSCLP)$ is the node-arc incidence matrix for the following line digraph: Node i of the graph corresponds to class i and the edges are $(i,\ i+1)$ for $i = 1, \ldots, 5n-1$. The matrix H is a block diagonal matrix, with each block a row vector of mean service times of the customers served at the same machine. F is a negative identity matrix. $c(t)$ is a zero vector. $g(t)$ is a randomly generated vector. $a(t) = y(0) + e_1 t$ with e_1 the unit vector whose first component is one and all the other components are zero. $b(t)$ is a vector of all ones and $h(t)$ is a zero vector.

The computational sequences are shown in Table 9.1.

When we fix the precision requirement and vary the number of stations in the example, we find that the computational time grows almost quadratically with the problem dimension, as shown in Figure 5. This is due to the fact that the number of control pieces grows almost linearly with the problem dimension and the total number of nonzero elements in the intermediate problems grows almost quadratically with the number of stations. Notice that for the largest example in Figure 5 (25 stations) there are 250 continuous variables.

Table 9.1
Test results for the example.

# Iter.	Obj. Value	# Pieces	Dual obj.	Time in sec.
0	20987.1355	7		
1	5986.7656	7	5956.7923	134.05
2	5965.1006	15	5962.7291	1738.2
3	5963.6674	29	5963.2700	2436.61

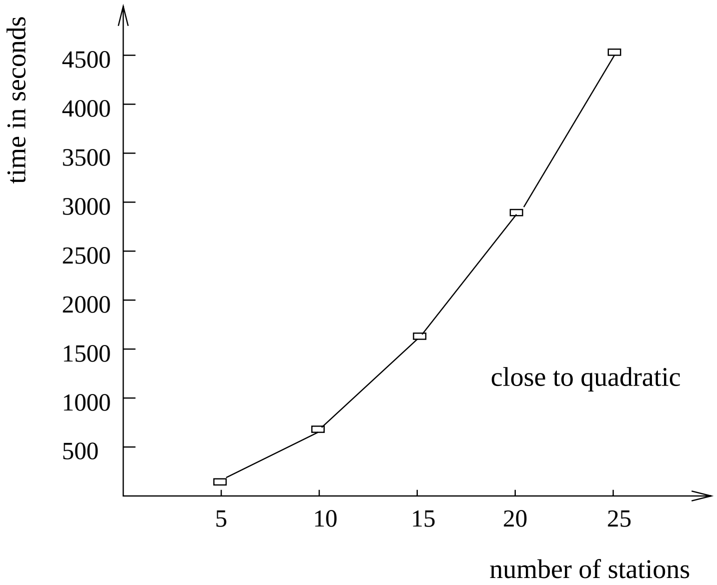


Fig. 5. *Computation time versus the number of stations* (*with precision fixed at* 0.0001).

This problem demonstrates that our algorithm can solve rather large problems. It is our experience that $(SCSCLP)$ is easier to approximate than to solve exactly. The computational time grows almost exponentially with the accuracy requirement. A key feature of Algorithm $\mathcal{A}$ is that it keeps the number of breakpoints as small as possible, which in turn makes the size of intermediate quadratic programming subproblems small. It is this feature that makes the algorithm efficient. We believe that Algorithm $\mathcal{A}$ can be made even more efficient if the special structure of the intermediate quadratic programs is exploited.

REFERENCES

[1] E. J. Anderson, P. Nash, and A. F. Perold, *Some properties of a class of continuous linear programs*, SIAM J. Control Optim., 21 (1983), pp. 758–765.

[2] E. J. Anderson and P. Nash, *Linear Programming in Infinite Dimensional Spaces*, Wiley-Interscience, Chichester, 1987.

[3] E. J. Anderson and A. B. Philpott, *A continuous–time network simplex algorithm*, Networks, 19 (1989), pp. 395–425.

[4] E. J. Anderson, *A Continuous Model for Job-Shop Scheduling*, Ph.D. thesis, University of Cambridge, Cambridge, UK, 1978.

[5] K. M. Anstreicher, *Generation of Feasible Descent Directions in Continuous Time Linear*

Programming, Tech. report SOL 83-18, Department of Operations Research, Stanford University, Stanford, CA, 1983.

[6] F. Avram, D. Bertsimas, and M. Ricard, *Optimization of multiclass queueing networks: A linear control approach*, in Stochastic Networks, Proc. IMA, F. Kelly and R. Williams, eds., Springer-Verlag, New York, 1995, pp. 199–234.

[7] R. Bellman, *Bottleneck problem and dynamic programming*, Proc. Nat. Acad. Sci., 39 (1953), pp. 947–951.

[8] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

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

[10] W. P. Drews, *A simplex-like algorithm for continuous-time linear optimal control problems*, in Optimization Methods for Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak and Co., New York, 1974, pp. 309–322.

[11] J. Eckstein, *Splitting Methods for Monotone Operators with Applications to Parallel Optimization*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1989.

[12] J. Filipiak, *Modeling and Control of Dynamic Flows in Communication Networks*, Springer-Verlag, Berlin, 1988.

[13] R. C. Grinold, *Continuous Programming*, Ph.D. thesis, Operations Research Center, University of California, Berkeley, CA, 1968.

[14] R. C. Grinold, *Continuous programming part one: Linear objectives*, J. Math. Anal. Appl., 28 (1969), pp. 32–51.

[15] R. C. Grinold, *Symmetric duality for continuous linear programs*, SIAM J. Appl. Math., 18 (1970), pp. 84–97.

[16] W. W. Hager and S. K. Mitter, *Lagrange duality theory for convex control problems*, SIAM J. Control Optim., 14 (1976), pp. 843–856.

[17] W. W. Hager, *Lipschitz continuity for constrained processes*, SIAM J. Control Optim., 17 (1979), pp. 321–338.

[18] B. Hajek and R. G. Ogier, *Optimal dynamic routing in communication networks with continuous traffic*, Networks, 14 (1984), pp. 457–487.

[19] J. M. Harrison, *Brownian Motion and Stochastic Flow Systems*, Graduate School of Business, Stanford University, Stanford, CA, Robert E. Krieger Publishing, Malabar, FL, 1990.

[20] J. Hartberger, *Representation extended to continuous time*, in Optimization Methods for Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak and Co., New York, 1974, pp. 309–322.

[21] A. E. Ilyutovich, *Iterative optimization method for linear programming problems in functional spaces*, in Studies in Linear and Nonlinear Programming, Stanford University Press, Stanford, CA, 1976.

[22] A. E. Ilyutovich, *Piecewise-continuous solutions of linear dynamic problems in economic planning*, Automat. Remote Control, 41 (1976), pp. 501–508.

[23] S. Ito, C. T. Kelley, and E. W. Sachs, *Inexact primal-dual interior point iteration for linear program in function spaces*, Comput. Optim. Appl., 4 (1995), pp. 189–202.

[24] M. Kojima, T. Noma, and A. Yoshise, *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*, Research report RJ 7493, IBM Almaden Research Center, San Jose, CA, 1990.

[25] P. R. Kumar, *Re-Entrant Lines*, Queueing Systems Theory Appl., 17 (1993), pp. 87–110.

[26] R. S. Lehman, *On the Continuous Simplex Method*, Tech. report RM–1386, Rand Corporation, Santa Monica, CA, 1954.

[27] N. Levinson, *A class of continuous linear programming problems*, J. Math. Anal. Appl., 16 (1966), pp. 73–83.

[28] Y. Y. Lin and J.-S. Pang, *Iterative methods for large convex quadratic programs: A survey*, SIAM J. Control Optim., 25 (1987), pp. 383–411.

[29] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[30] D. G. Luenberger, *Introduction to Linear and Nonlinear Programming*, Addison–Wesley, Reading, MA, 1973.

[31] Z.-Q. Luo and P. Tseng, *Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem*, SIAM J. Optim., 2 (1992), pp. 43–54.

[32] X. D. Luo, *Continuous Linear Programming: Theory, Algorithms and Applications*, Ph.D. thesis, MIT Operations Research Center, Cambridge MA, 1995.

[33] B. Martos, *Nonlinear Programming, Theory, and Methods*, North–Holland, Amsterdam, 1975.

[34] F. H. Moss and A. Segall, *An optimal control approach to dynamic routing in networks*,

IEEE Trans. Automat. Control, AC–27 (1982), pp. 329–339.

[35] F. H. Moss, *The Application of Optimal Control Theory to Dynamic Routing in Data Communication Networks*, Ph.D. thesis, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA, 1977.

[36] K. G. Murty, *Linear Complementarity, Linear and Nonlinear Programming*, Helderman, Verlag, Berlin, 1988.

[37] A. F. Perold, *Fundamentals of a Continuous Time Simplex Method*, Tech. report Sol 78–26, Department of Operations Research, Stanford University, Stanford, CA, 1978.

[38] A. F. Perold, *Extreme points and basic feasible solutions in continuous time linear programming,* SIAM J. Control Optim., 19 (1981), pp. 52–63.

[39] A. B. Philpott and M. Craddock, *An Adaptive Discretization Algorithm for a Class of Continuous Network Programs*, Department of Engineering Science, University of Auckland, Auckland, New Zealand, submitted.

[40] M. C. Pullan, A. Mason, M. Craddock, and A.B. Philpott, *Computer Programs for Solving the Capacitated Transshipment Problem*, School of Engineering Report 521, University of Auckland, Auckland, New Zealand, 1992.

[41] M. C. Pullan, *An algorithm for a class of continuous linear programming programs*, SIAM J. Control Optim., 31 (1993), pp. 1558–1577.

[42] M. C. Pullan, *Forms of optimal solutions for separated continuous linear programs*, SIAM J. Control Optim., 33 (1995), pp. 1952–1977.

[43] M. C. Pullan, *A duality theory for separated continuous linear programs,* SIAM J. Control Optim., 34 (1996), pp. 931–965.

[44] R. G. Segers, *A generalized function setting for dynamic optimal control problems*, in Optimization Methods for Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak and Co., New York, 1974, pp. 279–296.

[45] S. P. Sethi, W. P. Drews, and R. G. Segers, *A unified framework for linear control problems with state-variable inequality constraints*, J. Optim. Theory Appl., 36 (1982), pp. 93–109.

[46] W. F. Tyndall, *A duality theorem for a class of continuous linear programming problems*, SIAM J. Appl. Math., 13 (1965), pp. 644–666.

[47] W. F. Tyndall, *An extended duality theorem for continuous linear programming problems*, SIAM J. Appl. Math., 15 (1967), pp. 1294–1298.

[48] R. J. Vanderbei, *LOQO User's Manual*, Program in Statistics and Operations Research, Princeton University, Princeton, NJ.

[49] G. Weiss, *On the optimal draining of re-entrant fluid lines*, in Stochastic Networks, Proc. IMA, F. Kelly and R. Williams, eds., Springer-Verlag, New York, 1995.

[50] M. Yamasaki, *Duality theorems in mathematical programmings and their applications*, J. Sci. Hiroshima Univ. Ser. A-1, 32 (1969), pp. 331–356.

[51] Y. Ye, *A fully polynomial-time approximation algorithm for computing a stationary point of the general linear complementarity problem*, Math. Oper. Res., 18 (1993), pp. 334–345.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 1, pp. 211–229

A CONVEX OPTIMIZATION APPROACH TO THE RATIONAL COVARIANCE EXTENSION PROBLEM*

CHRISTOPHER I. BYRNES†, SERGEI V. GUSEV‡, AND ANDERS LINDQUIST§

Abstract. In this paper we present a convex optimization problem for solving the rational covariance extension problem. Given a partial covariance sequence and the desired zeros of the modeling filter, the poles are uniquely determined from the unique minimum of the corresponding optimization problem. In this way we obtain an algorithm for solving the covariance extension problem, as well as a constructive proof of Georgiou's seminal existence result and his conjecture, a stronger version of which we have resolved in [Byrnes et al., *IEEE Trans. Automat. Control*, AC-40 (1995), pp. 1841–1857].

Key words. rational covariance extension, partial stochastic realization, trigonometric moment problem, spectral estimation, speech processing, stochastic modeling

AMS subject classifications. 30E05, 60G35, 62M15, 93A30, 93E12

PII. S0363012997321553

1. Introduction.

In [7] a solution to the problem of parameterizing all rational extensions of a given window of covariance data has been given. This problem has a long history, with antecedents going back to potential theory in the work of Carathéodory, Toeplitz, and Schur [9, 10, 31, 30], and continuing in the work of Kalman, Georgiou, Kimura, and others [18, 14, 21]. It has been of more recent interest due to its significant interface with problems in signal processing and speech processing [11, 8, 25, 20] and in stochastic realization theory and system identification [2, 32, 22]. Indeed, the recent solution to this problem, which extended a result by Georgiou and confirmed one of his conjectures [13, 14], has shed some light on the stochastic (partial) realization problem through the development of an associated Riccati-type equation, whose unique positive semidefinite solution has as its rank the minimum dimension of a stochastic linear realization of the given rational covariance extension [6]. In both its form as a complete parameterization of rational extensions to a given covariance sequence and as an indefinite Riccati-type equation, one of the principal problems which remains open is that of developing effective computational methods for the approximate solution of this problem. In this paper, motivated by the effectiveness of interior point methods for solving nonlinear convex optimization problems, we recast the fundamental problem as such an optimization problem.

In section 2 we describe the principal results for the rational covariance extension problem and set notation we shall need throughout. The only solution to this problem for which there have been simple computational procedures is the so-called *maximum entropy* solution, which is the particular solution that maximizes the entropy gain.

*Received by the editors May 15, 1997; accepted for publication December 3, 1997; published electronically October 7, 1998. This research was supported in part by grants from AFOSR, NSF, TFR, the Göran Gustafsson Foundation, the Royal Swedish Academy of Sciences, and Southwestern Bell.

http://www.siam.org/journals/sicon/37-1/32155.html

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

‡Department of Mathematics and Mechanics, St. Petersburg University, St. Petersburg 198904, Russia (sergei@gusev.niimm.spb.su).

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

In section 3 we demonstrate that the infinite-dimensional optimization problem for determining this solution has a simple finite-dimensional dual. This motivates the introduction in section 4 of a nonlinear, strictly convex functional defined on a closed convex set naturally related to the covariance extension problem. We first show that any solution of the rational covariance extension problem lies in the interior of this convex set and that, conversely, an interior minimum of this convex functional will correspond to the unique solution of the covariance extension problem. Our interest in this convex optimization problem is, therefore, twofold: as a starting point for the computation of an explicit solution and as a means of providing an alternative proof of the rational covariance extension theorem.

Concerning the existence of a minimum, we show that this functional is proper and bounded below, i.e., that the sublevel sets of this functional are compact. From this, it follows that there exists a minimum. Since uniqueness follows from strict convexity of the functional, the central issue which needs to be addressed in order to solve the rational covariance extension problem is whether, in fact, this minimum is an interior point. Indeed, our formulation of the convex functional, which contains a barrier-like term, was inspired by interior point methods. However, in contrast to interior point methods, the barrier function we have introduced does not become infinite on the boundary of our closed convex set. Nonetheless, we are able to show that the gradient, rather than the value, of the convex functional becomes infinite on the boundary. The existence of an interior point which minimizes the functional then follows from this observation.

In section 5, we apply these convex minimization techniques to the rational covariance extension problem, noting that, as hinted above, we obtain a new proof of Georgiou's conjecture. Moreover, this proof, unlike our previous proof [7] and the existence proof of Georgiou [14], is constructive. Consequently, we have also obtained an algorithmic procedure for solving the rational covariance extension problem. In section 6 we report some computational results and present some simulations.

2. The rational covariance extension problem.

It is well known that the spectral density $\Phi(z)$ of a purely nondeterministic stationary random process $\{y(t)\}$ is given by the Fourier expansion

$$\Phi(e^{i\theta}) = \sum_{-\infty}^{\infty} c_k e^{ik\theta} \tag{2.1}$$

on the unit circle, where the covariance lags

$$c_k = \mathrm{E}\{y_{t+k}y_t\}, \quad k = 0, 1, 2, \dots \tag{2.2}$$

play the role of the Fourier coefficients

$$c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\Phi(e^{i\theta})d\theta. \tag{2.3}$$

In spectral estimation [8], identification [2, 22, 32], speech processing [11, 25, 24, 29], and several other applications in signal processing and systems and control, we are faced with the inverse problem of finding a spectral density which is *coercive*, i.e., positive on the unit circle, given only

$$c = (c_0, c_1, \dots, c_n), \tag{2.4}$$

which is a *partial covariance sequence* positive in the sense that

$$T_n = \begin{bmatrix} c_0 & c_1 & \cdots & c_n \\ c_1 & c_0 & \cdots & c_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ c_n & c_{n-1} & \cdots & c_0 \end{bmatrix} > 0, \tag{2.5}$$

i.e., the Toeplitz matrix T_n is positive definite.

In fact, the covariance lags (2.2) are usually estimated from an approximation

$$\frac{1}{N-k+1} \sum_{t=0}^{N-k} y_{t+k} y_t$$

of the ergodic limit

$$c_k = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} y_{t+k} y_t,$$

since only a finite string

$$y_0, y_1, y_2, y_3, \dots, y_N$$

of observations of the process $\{y(t)\}$ is available, and therefore we can only estimate a finite partial covariance (2.4), where $n << N$.

The corresponding inverse problem is a version of the *trigonometric moment problem* [1, 16]: Given a sequence (2.4) of real numbers satisfying the positivity condition (2.5), find a coercive spectral density $\Phi(z)$ such that (2.3) is satisfied for $k = 0, 1, 2, \dots, n$. Of course there are infinitely many such solutions, and we shall shortly specify some additional properties which we would like the solution to have.

The trigonometric moment problem, as stated above, is equivalent to the *Carathéodory extension problem* to determine an extension

$$c_{n+1}, c_{n+2}, c_{n+3}, \dots, \tag{2.6}$$

with the property that the function

$$v(z) = \frac{1}{2} c_0 + c_1 z^{-1} + c_2 z^{-2} + \cdots \tag{2.7}$$

is *strictly positive real*, i.e., is analytic on and outside the unit circle (so that the Laurent expansion (2.7) holds for all $|z| \geq 1$) and satisfies

$$v(z) + v(z^{-1}) > 0 \quad \text{on the unit circle.} \tag{2.8}$$

In fact, given such a $v(z)$,

$$\Phi(z) = v(z) + v(z^{-1}) \tag{2.9}$$

is a solution to the trigonometric moment problem. Conversely, any coercive spectral density $\Phi(z)$ uniquely defines a strictly positive real function $v(z)$ via (2.9).

These problems are classical and go back to Carathéodory [9, 10], Toeplitz [31], and Schur [30]. In fact, Schur parameterized all solutions in terms of what are now known as the *Schur parameters*, or, more commonly in the circuits and systems literature, as *reflection coefficients*, and which are easily determined from the covariance lags via the Levinson algorithm [27]. More precisely, modulo the choice of c_0, there is a one-to-one correspondence between infinite covariance sequences $c_0, c_1, c_2, \dots$ and Schur parameters $\gamma_0, \gamma_1, \dots$ such that

$$|\gamma_t| < 1 \quad \text{for } t = 0, 1, 2, \dots, \tag{2.10}$$

under which partial sequences (2.4) correspond to partial sequences $\gamma_0, \gamma_1, \dots, \gamma_{n-1}$ of Schur parameters. Therefore, covariance extension (2.6) amounts precisely to finding a continuation

$$\gamma_n, \gamma_{n+1}, \gamma_{n+2}, \dots \tag{2.11}$$

of Schur parameters satisfying (2.10). Each such solution is only guaranteed to yield a $v(z)$ which is meromorphic.

In circuits and systems theory, however, we are generally only interested in solutions which yield a rational $v(z)$ of at most degree n, or, equivalently, a rational spectral density $\Phi(z)$ of at most degree $2n$. Then the unique rational, stable, minimum-phase function $w(z)$ having the same degree as $v(z)$ and satisfying

$$w(z)w(z^{-1}) = \Phi(z) \tag{2.12}$$

is the transfer function of a *modeling filter*, which shapes white noise into a random process with the first $n+1$ covariance lags given by (2.4); see, e.g., [7, 6] for more details.

Setting all free Schur parameters (2.11) equal to zero, which clearly satisfies the condition (2.10), yields a rational solution

$$\Phi(z) = \frac{1}{a(z)a(z^{-1})}, \tag{2.13}$$

where $a(z)$ is a polynomial given by

$$a(z) = a_0 z^n + a_1 z^{n-1} + \cdots + a_n \quad (a_0 > 0), \tag{2.14}$$

which is easily computed via the Levinson algorithm [27]. This so-called *maximum entropy solution* is an all-pole or AR solution, and the corresponding modeling filter

$$w(z) = \frac{z^n}{a(z)} \tag{2.15}$$

has all its zeros at the origin.

However, in many applications a wider variety in the choice of zeros is required in the spectral density $\Phi(z)$. To illustrate this point, consider in Figure 2.1 a spectral density in the form of a periodogram determined from a speech signal sampled over 20 milliseconds (in which time interval it represents a stationary process) together with a maximum entropy solution corresponding to $n = 6$. As can be seen, the latter yields a rather flat spectrum which is unable to approximate the valleys or the "notches" in the speech spectrum, and therefore in speech synthesis, the maximum entropy solution results in artificial speech which sounds quite flat. This is a manifestation of the fact that all the zeros of the maximum entropy filter (2.15) are located at the origin and thus do not give rise to a frequency where the power spectrum vanishes. However,

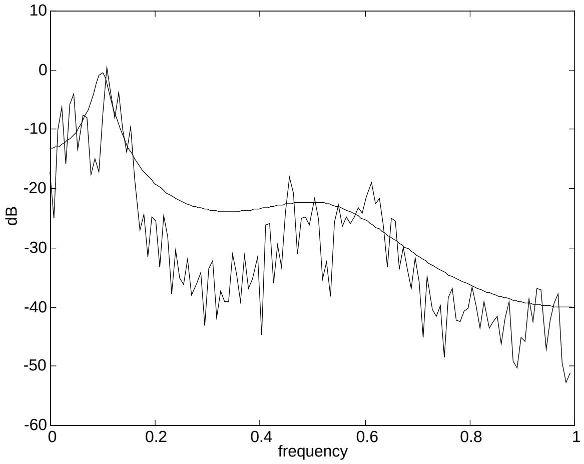


FIG. 2.1. *Spectral envelope of a maximum entropy solution.*

were we able to place some zeros of the modeling filter reasonably close to the unit circle, these would produce notches in the spectrum at approximately the frequency of the arguments of those zeros.

For this reason, it is widely appreciated in the signal and speech processing community that regeneration of human speech requires the design of filters having nontrivial zeros [3, p. 1726], [24, pp. 271–272], [29, pp. 76–78]. Indeed, while all-pole filters can reproduce many human speech sounds, the acoustic theory teaches that nasals and fricatives require both zeros and poles [24, pp. 271–272], [29, p. 105].

Therefore, we are interested in modeling filters

$$w(z) = \frac{\sigma(z)}{a(z)}, \tag{2.16}$$

for which (2.14) and

$$\sigma(z) = z^n + \sigma_1 z^{n-1} + \cdots + \sigma_n \tag{2.17}$$

are *Schur polynomials*, i.e., polynomials with all roots in the open unit disc. In this context, the maximum entropy solution corresponds to the choice $\sigma(z) = z^n$.

An important mathematical question, therefore, is to what extent it is possible to assign desired zeros and still satisfy the interpolation condition that the partial covariance sequence (2.4) is as prescribed. In [13] (see also [14]), Georgiou proved that for any prescribed zero polynomial $\sigma(z)$ there exists a modeling filter $w(z)$ and conjectured that this correspondence would yield a complete parameterization of all rational solutions of at most degree n, i.e., that the correspondence between v and a choice of positive sequence (2.4) and a choice of Schur polynomial (2.14) would be a bijection. This is a nontrivial and highly nonlinear problem, since generally there is no method to see which choices of free Schur parameters will yield rational solutions. In [7] we resolved this long-standing conjecture by proving the following theorem as a corollary of a more general theorem on complementary foliations of the space of all rational positive real functions of degree at most n.

THEOREM 2.1 (see [7]). *Given any partial covariance sequence* (2.4) *and Schur polynomial* (2.17)*, there exists a unique Schur polynomial* (2.14) *such that* (2.16) *is* *a minimum-phase spectral factor of a spectral density* $\Phi(z)$ *satisfying*

$$\Phi(z) = c_0 + \sum_{k=1}^{\infty} \hat{c}_k (z^k + z^{-k}),$$

where

$$\hat{c}_k = c_k \quad \textit{for} \quad i = 1, 2, \dots, n.$$

In particular, the solutions of the rational positive extension problem are in one-to-one correspondence with self-conjugate sets of n points (counted with multiplicity) lying in the open unit disc, i.e., with all possible zero structures of modeling filters. Moreover, this correspondence is bianalytic.

Consequently, we not only proved Georgiou's conjecture that the family of all rational covariance extensions of (2.4) of degree at most n is completely parameterized in terms of the zeros of the corresponding modeling filters $w(z)$, but also that the modeling filter $w(z)$ depends analytically on the covariance data and the choice of zeros, a strong form of well-posedness increasing the likelihood of finding a numerical algorithm.

In fact, both Georgiou's existence proof and our proof of Theorem 2.1 are nonconstructive. However, in this paper we present for the first time an algorithm which, given the partial covariance sequence (2.4) and the desired zero polynomial (2.17), computes the unique pole polynomial (2.14). This is done via the convex optimization problem to minimize the value of the function $\varphi : \mathbb{R}^{n+1} \to \mathbb{R}$, defined by

$$\begin{aligned} \varphi(q_0, q_1, \dots, q_n) &= c_0 q_0 + c_1 q_1 + \cdots + c_n q_n \\ &\quad - \frac{1}{2\pi} \int_{-\pi}^{\pi} \log Q(e^{i\theta}) |\sigma(e^{i\theta})|^2 d\theta \end{aligned} \tag{2.18}$$

over all $q_0, q_1, \dots, q_n$ such that

$$Q(e^{i\theta}) = q_0 + q_1 \cos\theta + q_2 \cos 2\theta + \cdots + q_n \cos n\theta > 0 \quad \text{for all } \theta. \tag{2.19}$$

In sections 4 and 5 we show this problem has a unique minimum. In this way we shall also provide a new and constructive proof of the weaker form of Theorem 2.1 conjectured by Georgiou.

Using this convex optimization problem, a sixth-degree modeling filter with zeros at the appropriate frequencies can be constructed for the speech segment represented by the periodogram of Figure 2.1. In fact, Figure 2.2 illustrates the same periodogram together with the spectral density of such a filter. As can be seen, this filter yields a much better description of the notches than does the maximum entropy filter.

Before turning to the main topic of this paper, the convex optimization problem for solving the rational covariance extension problem for arbitrarily assigned zeros, we shall provide a motivation for this approach in terms of the maximum entropy solution.

3. The maximum entropy solution. As a preliminary we shall first consider the maximum entropy solution discussed in section 2. The reason for this is that, as indicated by its name, this particular solution corresponds to an optimization problem. Hence, this section will be devoted to clarifying the relation between this particular optimization problem and the class of problems solving the general problem. Thus

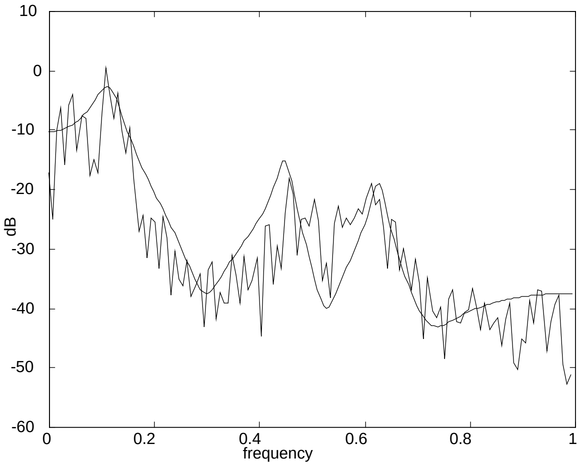


FIG. 2.2. *Spectral envelope obtained with appropriate choice of zeros.*

our interest is not in the maximum entropy solution per se, but in showing that it can be determined from a constrained convex minimization problem in $\mathbb{R}^{n+1}$, which naturally is generalized to a problem with arbitrary prescribed zeros.

Let us briefly recall the problem at hand. Given the partial covariance sequence

$$c_0, c_1, \dots, c_n,$$

determine a coercive, rational spectral density

$$\Phi(z) = \hat{c}_0 + \sum_{k=1}^{\infty} \hat{c}_k (z^k + z^{-k}) \tag{3.1}$$

of degree at most $2n$ such that

$$\hat{c}_k = c_k \quad \text{for} \quad i = 1, 2, \dots, n. \tag{3.2}$$

Of course there are many solutions to this problem, and it is well known that the maximum entropy solution is the one which maximizes the entropy gain

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \log \Phi(e^{i\theta}) d\theta \tag{3.3}$$

(see, e.g., [19]), and we shall now consider this constrained optimization problem.

We begin by setting up the appropriate spaces. Recall from classical realization theory that a rational function

$$v(z) = \frac{1}{2}\hat{c}_0 + \hat{c}_1 z^{-1} + \hat{c}_2 z^{-2} + \cdots$$

of degree n has a representation

$$\hat{c}_k = h' F^{k-1} g \quad k = 1, 2, 3, \dots$$

for some choice of $(F, g, h) \in \mathbb{R}^{n \times n} \times \mathbb{R}^n \times \mathbb{R}^n$. Therefore, if in addition $v(z)$ is strictly positive real, implying that all eigenvalues of F are less than one in modulus, $\hat{c}_k$ tends exponentially to zero as $k \to \infty$. Hence, in particular,

$$\hat{c} := (\hat{c}_0, \hat{c}_1, \hat{c}_2, \dots)$$

must belong to ℓ_1. Moreover, the requirement that (3.1) be a coercive spectral density adds another constraint, namely that $\hat{c}$ belongs to the set

$$\mathcal{F} := \left\{ \hat{c} \in \ell_1 \mid \hat{c}_0 + \sum_{k=1}^{\infty} \hat{c}_k (e^{ik\theta} + e^{-ik\theta}) > 0 \right\}. \tag{3.4}$$

Now, let

$$\psi(\hat{c}) = -\frac{1}{2\pi} \int_{-\pi}^{\pi} \log \left[\hat{c}_0 + \sum_{k=1}^{\infty} \hat{c}_k (e^{ik\theta} + e^{-ik\theta}) \right] d\theta \tag{3.5}$$

be a functional $\mathcal{F} \to \mathbb{R}$, and consider the infinite-dimensional convex constrained optimization problem to minimize $\psi(\hat{c})$ over $\mathcal{F}$ given the finite number of constraints (3.2). Thus we have relaxed the optimization problem to allow also for nonrational spectral densities.

Since the optimization problem is convex, the Lagrange function

$$L(\hat{c}, \lambda) = \psi(\hat{c}) + \sum_{k=0}^{n} \lambda_k (\hat{c}_k - c_k) \tag{3.6}$$

has a saddle point [26, p. 458] provided the stationary point lies in the interior of $\mathcal{F}$, and, in this case, the optimal Lagrange vector $\lambda = (\lambda_0, \lambda_1, \dots, \lambda_n) \in \mathbb{R}^{n+1}$ can be determined by solving the *dual problem* to maximize

$$\rho(\lambda) = \min_{\hat{c} \in \mathcal{F}} L(\hat{c}, \lambda). \tag{3.7}$$

To this end, first note that

$$\frac{\partial L}{\partial \hat{c}_k} = -\frac{1}{2\pi} \int_{-\pi}^{\pi} (e^{ik\theta} + e^{-ik\theta}) \Phi^{-1}(e^{i\theta}) d\theta + \lambda_k \quad \text{for } k = 0, 1, 2, \dots, n, \tag{3.8}$$

and that

$$\frac{\partial L}{\partial \hat{c}_k} = -\frac{1}{2\pi} \int_{-\pi}^{\pi} (e^{ik\theta} + e^{-ik\theta}) \Phi^{-1}(e^{i\theta}) d\theta \quad \text{for } k = n+1, n+2, \dots. \tag{3.9}$$

Then, setting the gradient equal to zero, we obtain from (3.9) that

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} (e^{ik\theta} + e^{-ik\theta}) \Phi^{-1}(e^{i\theta}) d\theta = 0 \quad \text{for } |k| > n,$$

from which it follows that Φ^{-1} must be a pseudopolynomial

$$Q(z) = q_0 + \frac{1}{2} q_1 (z + z^{-1}) + \cdots + \frac{1}{2} q_n (z^n + z^{-n}) \tag{3.10}$$

of degree at most n, i.e.,

$$\Phi^{-1}(z) = Q(z), \tag{3.11}$$

yielding a spectral density Φ which is rational and of at most degree $2n$, and thus belongs to the original (nonrelaxed) class of spectral densities. Likewise we obtain from (3.8) that

$$\lambda_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} (e^{ik\theta} + e^{-ik\theta}) \Phi^{-1}(e^{i\theta}) d\theta \tag{3.12}$$

for $k = 0, 1, 2, \dots, n$, which together with (3.11) yields

$$\lambda_k = q_k \quad \text{for } k = 0, 1, 2, \dots, n. \tag{3.13}$$

However, the minimizing $\hat{c}$ is given by

$$\hat{c}_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{1}{2}(e^{ik\theta} + e^{-ik\theta})Q(e^{i\theta})^{-1}d\theta \tag{3.14}$$

and consequently

$$\sum_{k=0}^{n} q_k\hat{c}_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} Q(e^{i\theta})Q(e^{i\theta})^{-1}d\theta = 1. \tag{3.15}$$

To determine the optimal (saddle point) Lagrange multipliers we turn to the dual problem. In view of (3.11), (3.13), and (3.15), the dual function is

$$\rho(q) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log Q(e^{i\theta})d\theta + 1 - c'q,$$

where $c \in \mathbb{R}^{n+1}$ is the vector with components $c_0, c_1, \dots, c_n$. Consequently, the dual problem is equivalent to minimizing

$$\varphi(q) = c'q - \frac{1}{2\pi}\int_{-\pi}^{\pi} \log Q(e^{i\theta})d\theta \tag{3.16}$$

over all $q \in \mathbb{R}^{n+1}$ such that the pseudopolynomial (3.10) is nonnegative on the unit circle, i.e.,

$$Q(e^{i\theta}) > 0 \quad \text{for all } \theta, \tag{3.17}$$

and, if the dual problem has an optimal solution satisfying (3.17), the optimal Q solves the primal problem when inserted into (3.11).

The dual problem to minimize (3.16) given (3.17) is a finite-dimensional convex optimization problem, which is simpler than the original (primal) problem. Clearly it is a special case of the optimization problem (2.18)–(2.19), obtained by setting $|\sigma(e^{i\theta})|^2 = 1$ as required for the maximum entropy solution. Figure 3.1 depicts a typical cost function φ in the case $n = 1$. As can be seen, it is convex and attains its optimum in an interior point so that the spectral density Φ has all its poles in the open unit disc as required. That this is the case in general will be proven in section 5.

We stress again that the purpose of this section is not primarily to derive an algorithm for the maximum entropy solution, for which we already have the simple Levinson algorithm, but to motivate an algorithm for the case with prescribed zeros in the spectral density. This is the topic of the next two sections.

4. The general convex optimization problem. Given a partial covariance sequence $c = (c_0, c_1, \dots, c_n)'$ and a Schur polynomial $\sigma(z)$, we know from section 2 that there exists a Schur polynomial

$$a(z) = a_0z^n + a_1z^{n-1} + \cdots + a_n \quad (a_0 > 0)$$

such that

$$\Phi(z) = \frac{\sigma(z)\sigma(z^{-1})}{a(z)a(z^{-1})} = c_0 + \sum_{k=1}^{\infty} \hat{c}_k(z^k + z^{-k}), \tag{4.1}$$

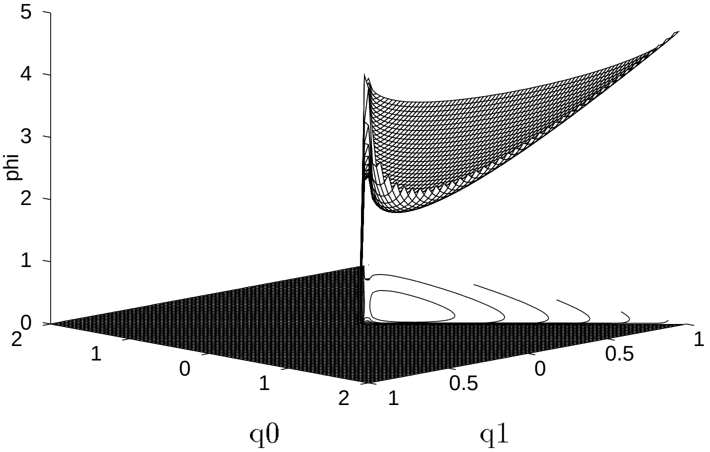


FIG. 3.1. *A typical cost function $\varphi(q)$ in the case $n = 1$.*

where

$$\hat{c}_k = c_k \quad \text{for } k = 1, 2, \dots, n. \tag{4.2}$$

The question now is: How do we find $a(z)$? In this section, we shall construct a nonlinear, strictly convex functional on a closed convex domain. In the next section, we shall show that this functional always has a unique minimum and that if such a minimum occurs as an interior point, it gives rise to $a(z)$.

As seen from (2.3), the interpolation condition (4.2) may be written

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\theta} \frac{|\sigma(e^{i\theta})|^2}{Q(e^{i\theta})} d\theta \quad \text{for } k = 0, 1, \dots, n, \tag{4.3}$$

where

$$Q(z) = a(z)a(z^{-1}), \tag{4.4}$$

so the problem is reduced to determining the variables

$$q = \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ q_n \end{bmatrix} \in \mathbb{R}^{n+1} \tag{4.5}$$

in the pseudopolynomial

$$Q(z) = q_0 + \frac{1}{2} q_1 (z + z^{-1}) + \frac{1}{2} q_2 (z^2 + z^{-2}) + \cdots + \frac{1}{2} q_n (z^n + z^{-n}) \tag{4.6}$$

so that the conditions (4.3) and

$$Q(e^{i\theta}) > 0 \quad \text{for all } \theta \in [-\pi, \pi] \tag{4.7}$$

are satisfied.

Now, consider the convex functional $\varphi(q) : \mathbb{R}^{n+1} \to \mathbb{R}$ defined by

$$\varphi(q) = c'q - \frac{1}{2\pi}\int_{-\pi}^{\pi} \log Q(e^{i\theta})|\sigma(e^{i\theta})|^2 d\theta. \tag{4.8}$$

Our motivation in defining $\varphi(q)$ comes in part from the desire to introduce a barrier-like term, as is done in interior point methods, and in part from our analysis of the maximum entropy method in the previous section. As it turns out, by a theorem of Szegö the logarithmic integrand is in fact integrable for nonzero Q having zeros on the boundary of the unit circle, so that $\varphi(q)$ does not become infinite on the boundary of the convex set. On the other hand, $\varphi(q)$ is a natural generalization of the functional (3.16) in section 3, since it specializes to (3.16) when $|\sigma(e^{i\theta})|^2 \equiv 1$ as for the maximum entropy solution. As we shall see, minimizing (4.8) yields precisely via (4.4) the unique $a(z)$ which corresponds to $\sigma(z)$.

It is clear that if $q \in \mathcal{D}_n^+$, where

$$\mathcal{D}_n^+ = \{q \in \mathbb{R}^{n+1} \mid Q(z) > 0 \quad \text{for } |z| = 1\}, \tag{4.9}$$

then $\varphi(q)$ is finite. Moreover, $\varphi(q)$ is also finite when $Q(z)$ has finitely many zeros on the unit circle, as can be seen from the following lemma.

LEMMA 4.1. *The functional $\varphi(q)$ is finite and continuous at any $q \in \overline{\mathcal{D}_n^+}$ except at zero. The functional is infinite, but continuous, at $q = 0$. Moreover, φ is a C^∞ function on $\mathcal{D}_n^+$.*

Proof. We want to prove that $\varphi(q)$ is finite when $q \neq 0$. Then the rest follows by inspection. Clearly, $\varphi(q)$ cannot take the value $-\infty$; hence, it remains to prove that $\varphi(q) < \infty$. Since $q \neq 0$,

$$\mu := \max_\theta Q(e^{i\theta}) > 0.$$

Then setting $P(z) := \mu^{-1}Q(z)$,

$$\log P(e^{i\theta}) \leq 0 \tag{4.10}$$

and

$$\varphi(q) = c'q - \frac{1}{2\pi}\log\mu \int_{-\pi}^{\pi} |\sigma(e^{i\theta})|^2 d\theta - \frac{1}{2\pi}\int_{-\pi}^{\pi} \log P(e^{i\theta})|\sigma(e^{i\theta})|^2 d\theta,$$

and hence, the question of whether $\varphi(q) < \infty$ is reduced to determining whether

$$-\int_{-\pi}^{\pi} \log P(e^{i\theta})|\sigma(e^{i\theta})|^2 d\theta < \infty.$$

However, since $|\sigma(e^{i\theta})|^2 \leq M$ for some bound M, this follows from

$$\int_{-\pi}^{\pi} \log P(e^{i\theta}) d\theta > -\infty, \tag{4.11}$$

which is the well-known Szegö condition: (4.11) is a necessary and sufficient condition for $P(e^{i\theta})$ to have a stable spectral factor [17]. However, since $P(z)$ is a symmetric pseudopolynomial which is nonnegative on the unit circle, there is a polynomial $\pi(z)$ such that $\pi(z)\pi(z^{-1}) = P(z)$. But then $w(z) = \frac{\pi(z)}{z^n}$ is a stable spectral factor, and hence (4.11) holds. $\square$

Lemma 4.2. *The functional $\varphi(q)$ is strictly convex and defined on a closed, convex domain.*

Proof. We first note that $q = 0$ is an extreme point, but it can never be a minimum of φ since $\varphi(0)$ is infinite. In particular, in order to check the strict inequality

$$\varphi(\lambda q^{(1)} + (1-\lambda)q^{(2)}) < \lambda\varphi(q^{(1)}) + (1-\lambda)\varphi(q^{(2)}), \tag{4.12}$$

where one of the arguments is zero, we need only consider the case that either $q^{(1)}$ or $q^{(2)}$ is zero, in which case the strict inequality holds. We can now assume that none of the arguments is zero, in which case the strict inequality in (4.12) follows from the strict concavity of the logarithm. Finally, it is clear that $\overline{\mathcal{D}_n^+}$ is a closed convex subset. □

Lemma 4.3. *Let $q \in \overline{\mathcal{D}_n^+}$, and suppose $q \neq 0$. Then $c'q > 0$.*

Proof. Consider an arbitrary covariance extension of c such as, for example, the maximum entropy extension, and let $\Phi(z)$ be the corresponding spectral density (2.9). Then c is given by (2.3), which may also be written

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

Therefore, in view of (4.6),

$$c'q = \frac{1}{2\pi}\int_{-\pi}^{\pi} Q(e^{ik\theta})\Phi(e^{i\theta})d\theta, \tag{4.13}$$

which is positive whenever $Q(z) \geq 0$ on the unit circle and $q \neq 0$. □

Proposition 4.4. *For all $r \in \mathbb{R}$, $\varphi^{-1}(-\infty, r]$ is compact. Thus φ is proper (i.e., $\varphi^{-1}(K)$ is compact whenever K is compact) and bounded from below.*

Proof. Suppose $q^{(k)}$ is a sequence in $M_r := \varphi^{-1}(-\infty, r]$. It suffices to show that $q^{(k)}$ has a convergent subsequence. Each $Q^{(k)}$ may be factored as

$$Q^{(k)}(z) = \lambda_k \bar{a}^{(k)}(z)\bar{a}^{(k)}(z^{-1}) = \lambda_k \bar{Q}^{(k)}(z),$$

where λ_k is positive and $\bar{a}^{(k)}(z)$ is a monic polynomial, all of whose roots lie in the closed unit disc. The corresponding sequence of the (unordered) set of n roots of each $\bar{a}^{(k)}(z)$ has a convergent subsequence, since all (unordered) sets of roots lie in the closed unit disc. Denote by $\bar{a}(z)$ the monic polynomial of degree n which vanishes at this limit set of roots. By reordering the sequence if necessary, we may assume the sequence $a^{(k)}(z)$ tends to $\bar{a}(z)$. Therefore, the sequence $q^{(k)}$ has a convergent subsequence if and only if the sequence λ_k does, which will be the case provided the sequence λ_k is bounded from above and from below away from zero. Before proving this, we note that the sequences $c'\bar{q}^{(k)}$, where $\bar{q}^{(k)}$ is the vector corresponding to the pseudopolynomial $\bar{Q}^{(k)}$, and

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \log \bar{Q}^{(k)}(e^{i\theta})|\sigma(e^{i\theta})|^2 d\theta \tag{4.14}$$

are both bounded from above and from below, respectively, away from zero and $-\infty$. The upper bounds come from the fact that $\{\bar{a}^{(k)}(z)\}$ are Schur polynomials and hence have their coefficients in the bounded Schur region. As for the lower bound of $c'\bar{q}^{(k)}$, note that $c'\bar{q}^{(k)} > 0$ for all k (Lemma 4.3) and $c'\bar{q}^{(k)} \to \alpha > 0$. In fact, $\bar{Q}^{(k)}(e^{i\theta}) \to |\bar{a}(e^{i\theta})|^2$, where $\bar{a}(z)$ has all its zeros in the closed unit disc, and hence it follows from (4.13) that $\alpha > 0$. Then, since $\varphi(q) < \infty$ for all $q \in \overline{\mathcal{D}_n^+}$ except $q = 0$ (Lemma 4.1), (4.14) is bounded away from $-\infty$. Next, observe that

$$\varphi(q^{(k)}) = \lambda_k c' \bar{q}^{(k)} - \frac{1}{2\pi} \log \lambda_k \int_{-\pi}^{\pi} |\sigma(e^{i\theta})|^2 d\theta - \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \bar{Q}^{(k)}(e^{i\theta}) |\sigma(e^{i\theta})|^2 d\theta.$$

From this we can see that if a subsequence of λ_k were to tend to zero, then $\varphi(q^{(k)})$ would exceed r. Likewise, if a subsequence of λ_k were to tend to infinity, φ would exceed r, since linear growth dominates logarithmic growth. □

5. Interior critical points and solutions of the rational covariance extension problem. In the previous section, we showed that φ has compact sublevel sets in $\overline{\mathcal{D}_n^+}$, so that φ achieves a minimum. Moreover, since φ is strictly convex and $\overline{\mathcal{D}_n^+}$ is convex, such a minimum is unique. We record these observations in the following statement.

PROPOSITION 5.1. *For each partial covariance sequence c and each Schur polynomial $\sigma(z)$, the functional φ has a unique minimum on $\overline{\mathcal{D}_n^+}$.*

In this paper we consider a question which is of independent interest: whether φ achieves its minimum at an interior point. The next result describes an interesting systems-theoretic consequence of the existence of such interior minima.

THEOREM 5.2. *Fix a partial covariance sequence c and a Schur polynomial $\sigma(z)$. If $\hat{q} \in \mathcal{D}_n^+$ is a minimum for φ, then*

$$\hat{Q}(z) = a(z)a(z^{-1}), \tag{5.1}$$

where $a(z)$ is the solution of the rational covariance extension problem.

Proof. Suppose that $\hat{q} \in \mathcal{D}_n^+$ is a minimum for φ. Then

$$\frac{\partial \varphi}{\partial q_k}(\hat{q}) = 0 \quad \text{for } k = 0, 1, 2, \ldots, n. \tag{5.2}$$

Differentiating inside the integral, which is allowed due to uniform convergence, (5.2) yields

$$c_k - \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{2}(e^{ik\theta} + e^{-ik\theta}) \frac{|\sigma(e^{i\theta})|^2}{\hat{Q}(e^{i\theta})} d\theta = 0 \quad \text{for } k = 0, 1, \ldots, n,$$

where $\hat{Q}(z)$ is the pseudopolynomial (4.6) corresponding to $\hat{q}$, or, equivalently,

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\theta} \frac{|\sigma(e^{i\theta})|^2}{\hat{Q}(e^{i\theta})} d\theta \quad \text{for } k = 0, 1, \ldots, n, \tag{5.3}$$

which is precisely the interpolation condition (4.3)–(4.4), provided (5.1) holds. □

As a corollary of this theorem, we have that the gradient of φ at *any* $\tilde{q} \in \mathcal{D}_n^+$ is given by

$$\frac{\partial \varphi}{\partial q_k}(\tilde{q}) = c_k - \tilde{c}_k, \tag{5.4}$$

where

$$\tilde{c}_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\theta} \frac{|\sigma(e^{i\theta})|^2}{\tilde{Q}(e^{i\theta})} d\theta, \quad k = 0, 1, 2, \ldots, n \tag{5.5}$$

is the partial covariance sequence corresponding to a process with spectral density

$$\tilde{\Phi}(e^{i\theta}) = \frac{|\sigma(e^{i\theta})|^2}{\tilde{Q}(e^{i\theta})},$$

where $\tilde{Q}(z)$ is the pseudopolynomial corresponding to $\tilde{q}$. The gradient is thus the difference between the true and calculated partial covariance sequences.

We now state the converse result, underscoring our interest in this particular convex optimization problem.

THEOREM 5.3. *For each partial covariance sequence c and each Schur polynomial $\sigma(z)$, suppose that $a(z)$ gives a solution to the rational covariance extension problem. If*

$$\hat{Q}(z) = a(z)a(z^{-1}), \tag{5.6}$$

then the corresponding $(n+1)$-vector $\hat{q}$ lies in $\mathcal{D}_n^+$ and is a unique minimum for φ.

Proof. Let $a(z)$ be the solution of the rational covariance extension problem corresponding to c and $\sigma(z)$, and let $\hat{Q}(z)$ be given by (5.6). Then c satisfies the interpolation condition (5.3), which is equivalent to (5.2), as seen from the proof of Theorem 5.2. However, since $a(z)$ is a Schur polynomial, $\hat{Q}(z) > 0$ on the unit circle, and thus $\hat{q} \in \mathcal{D}_n^+$. Since φ is strictly convex on $\mathcal{D}_n^+$, (5.3) implies that $\hat{q}$ is a unique minimum for φ. □

Since the existence of a solution to the rational covariance extension problem has been established in [14] (see also [7]), we do in fact know the existence of interior minima for this convex optimization problem. On the other hand, we know from Proposition 5.1 that φ has a minimum for some $\hat{q} \in \overline{\mathcal{D}_n^+}$, so to show that φ has a minimum in the interior $\mathcal{D}_n^+$ it remains to prove the following lemma.

LEMMA 5.4. *The functional φ never attains a minimum on the boundary $\partial\mathcal{D}_n^+$.*

Proof. Denoting by $D_p\varphi(q)$ the directional derivative of φ at q in the direction p, it is easy to see that

$$D_p\varphi(q) := \lim_{\epsilon\to 0} \frac{\varphi(q+\epsilon p) - \varphi(q)}{\epsilon} \tag{5.7}$$

$$= c'p - \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{P(e^{i\theta})}{Q(e^{i\theta})}|\sigma(e^{i\theta})|^2 d\theta, \tag{5.8}$$

where $P(z)$ is the pseudopolynomial

$$P(z) = p_0 + \frac{1}{2}p_1(z+z^{-1}) + \frac{1}{2}p_2(z^2+z^{-2}) + \cdots + \frac{1}{2}p_n(z^n+z^{-n})$$

corresponding to the vector $p \in \mathbb{R}^{n+1}$. In fact,

$$\frac{\log(Q+\epsilon P) - \log Q}{\epsilon} = \frac{P}{Q}\log\left[\left(1+\epsilon\frac{P}{Q}\right)^{\frac{1}{\epsilon}\frac{Q}{P}}\right] \to \frac{P}{Q}$$

as $\epsilon \to +0$, and hence (5.7) follows by dominated convergence.

Now, let $q \in \mathcal{D}_n^+$ and $\bar{q} \in \partial\mathcal{D}_n^+$ be arbitrary. Then the corresponding pseudopolynomials Q and $\bar{Q}$ have the properties

$$Q(e^{i\theta}) > 0 \quad \text{for all } \theta \in [-\pi, \pi]$$

and

$$\bar{Q}(e^{i\theta}) \geq 0 \quad \text{for all } \theta \text{ and } \bar{Q}(e^{i\theta_0}) = 0 \text{ for some } \theta_0.$$

Since $q_\lambda := \bar{q} + \lambda(q - \bar{q}) \in \mathcal{D}_n^+$ for $\lambda \in (0,1]$, we also have for $\lambda \in (0,1]$ that

$$Q_\lambda(e^{i\theta}) := \bar{Q}(e^{i\theta}) + \lambda[Q(e^{i\theta}) - \bar{Q}(e^{i\theta})] > 0 \quad \text{for all } \theta \in [-\pi, \pi],$$

and we may form the directional derivative

$$D_{\bar{q}-q}\varphi(q_\lambda) = c'(\bar{q} - q) + \frac{1}{2\pi}\int_{-\pi}^{\pi} h_\lambda(\theta)d\theta, \tag{5.9}$$

where

$$h_\lambda(\theta) = \frac{Q(e^{i\theta}) - \bar{Q}(e^{i\theta})}{Q_\lambda(e^{i\theta})}|\sigma(e^{i\theta})|^2.$$

Now,

$$\frac{d}{d\lambda}h_\lambda(\theta) = \frac{[Q(e^{i\theta}) - \bar{Q}(e^{i\theta})]^2}{Q_\lambda(e^{i\theta})^2}|\sigma(e^{i\theta})|^2 \geq 0,$$

and hence $h_\lambda(\theta)$ is a monotonically nondecreasing function of λ for all $\theta \in [-\pi, \pi]$. Consequently, h_λ tends pointwise to h_0 as $\lambda \to 0$. Therefore,

$$\int_{-\pi}^{\pi} h_\lambda(\theta)d\theta \to +\infty \quad \text{as } \lambda \to 0. \tag{5.10}$$

In fact, if

$$\int_{-\pi}^{\pi} h_\lambda(\theta)d\theta \to \alpha < \infty \quad \text{as } \lambda \to 0, \tag{5.11}$$

then $\{h_\lambda\}$ is a Cauchy sequence in $L^1(-\pi, \pi)$ and hence has a limit in $L^1(-\pi, \pi)$ which must equal h_0 almost everywhere. However, h_0, having poles in $[-\pi, \pi]$, is not summable and hence, as claimed, (5.11) cannot hold.

Consequently, by virtue of (5.9),

$$D_{q-\bar{q}}\varphi(q_\lambda) \to +\infty \quad \text{as } \lambda \to 0$$

for all $q \in \mathcal{D}_n^+$ and $\bar{q} \in \partial\mathcal{D}_n^+$, and hence, in view of Lemma 26.2 in [28], φ is essentially smooth. Then it follows from Theorem 26.3 in [28] that the subdifferential of φ is empty on the boundary of $\mathcal{D}_n^+$, and therefore φ cannot have a minimum there. □

Thus we have proven the following result.

THEOREM 5.5. *For each partial covariance sequence c and each Schur polynomial $\sigma(z)$, there exists an $(n+1)$-vector $\hat{q}$ in $\mathcal{D}_n^+$ which is a minimum for φ.*

Consequently, by virtue of Theorem 5.2, there does exist a solution to the rational covariance extension problem for each partial covariance sequence and zero polynomial $\sigma(z)$, and, in view of Theorem 5.3, this solution is unique.

These theorems have the following corollary.

COROLLARY 5.6 (Georgiou's conjecture). *For each partial covariance sequence c and each Schur polynomial $\sigma(z)$, there is a unique Schur polynomial $a(z)$ such that* (4.1) *and* (4.2) *hold.*

Hence, we have given an independent proof of the weaker version of Theorem 2.1 conjectured by Georgiou, but not of the stronger version of [7] which states that the problem is well posed in the sense that the one-to-one correspondence between $\sigma(z)$ and $a(z)$ is a diffeomorphism.

6. Some numerical examples. Given an arbitrary partial covariance sequence $c_0, c_1, \dots, c_n$ and an arbitrary zero polynomial $\sigma(z)$, the constructive proof of Georgiou's conjecture provides algorithmic procedures for computing the corresponding unique modeling filter, which are based on the convex optimization problem to minimize the functional (2.18) over all $q_0, q_1, \dots, q_n$ such that (2.19) holds.

In general such procedures will be based on the gradient of the cost functional φ, which, as we saw in section 5, is given by

$$\frac{\partial \varphi}{\partial q_k}(q_0, q_1, \dots, q_n) = c_k - \bar{c}_k, \tag{6.1}$$

where

$$\bar{c}_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\frac{|\sigma(e^{i\theta})|^2}{Q(e^{i\theta})}d\theta \quad \text{for } k = 0, 1, 2, \dots, n \tag{6.2}$$

are the covariances corresponding to a process with spectral density

$$\frac{|\sigma(e^{i\theta})|^2}{Q(e^{i\theta})} = \bar{c}_0 + 2\sum_{k=1}^{\infty} \bar{c}_k \cos(k\theta). \tag{6.3}$$

The gradient is thus the difference between the given partial covariance sequence $c_0, c_1, \dots, c_n$ and the partial covariance sequence corresponding to the choice of variables $q_0, q_1, \dots, q_n$ at which the gradient is calculated. The minimum is attained when this difference is zero.

The following simulations have been done by Per Enqvist, using Newton's method (see, e.g., [23, 26]), which of course also requires computing the Hessian (second-derivative matrix) in each iteration. A straightforward calculation shows that the Hessian is the sum of a Toeplitz and a Hankel matrix. More precisely,

$$H_{ij}(q_0, q_1, \dots, q_n) = \frac{1}{2}(d_{i+j} + d_{i-j}), \quad i, j = 0, 1, 2, \dots, n, \tag{6.4}$$

where

$$d_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\theta}\frac{|\sigma(e^{i\theta})|^2}{Q(e^{i\theta})^2}d\theta \quad \text{for } k = 0, 1, 2, \dots, 2n \tag{6.5}$$

and $d_{-k} = d_k$. Moreover, $d_0, d_1, d_2, \dots, d_{2n}$ are the $2n+1$ first Fourier coefficients of the spectral representation

$$\frac{|\sigma(e^{i\theta})|^2}{Q(e^{i\theta})^2} = d_0 + 2\sum_{k=1}^{\infty} d_k \cos(k\theta). \tag{6.6}$$

The gradient and the Hessian can be determined from (6.1) and (6.4), respectively, by applying the inverse Levinson algorithm (see, e.g., [27]) to the appropriate polynomial spectral factors of $Q(z)$ and $Q(z)^2$, respectively, and then solving the resulting linear equations for $\bar{c}_0, \bar{c}_1, \dots, \bar{c}_n$ and $d_0, d_1, d_2, \dots, d_{2n}$; see [12] for details.

To illustrate the procedure, let us again consider the sixth-order spectral envelopes of Figures 2.1 and 2.2 together with the corresponding zeros and poles. Hence, Figure 6.1 illustrates the periodogram for a section of speech data together with the corresponding sixth-order maximum entropy spectrum, which, since it lacks finite zeros,

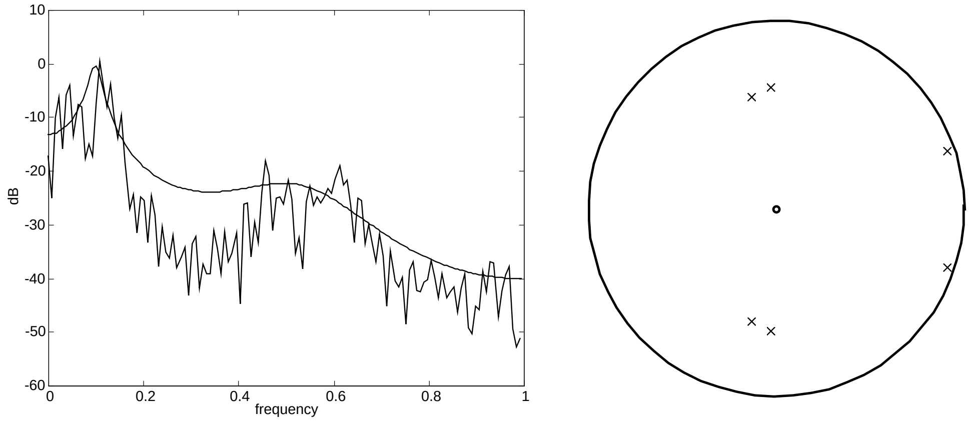


FIG. 6.1.

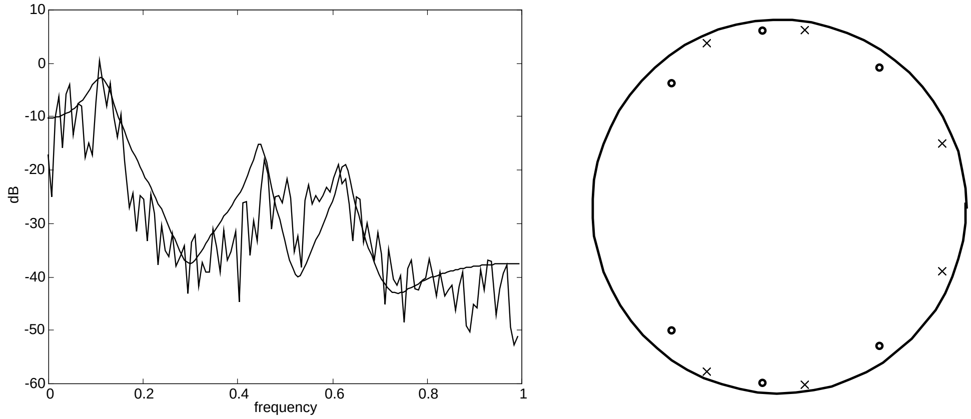


FIG. 6.2.

becomes rather “flat.” The location of the corresponding poles (marked by ×) in the unit circle is shown next to it. The zeros (marked by ∘) of course all lie at the origin.

Now, selecting the zeros appropriately as indicated to the right in Figure 6.2, we obtain the poles as marked, and the corresponding sixth-order modeling filter produces the spectral envelope to the left in Figure 6.2. We see that the second solution has a spectral density that is less flat and provides a better approximation, reflecting the fact that the filter is designed to have transmission zeros near the minima of the periodogram.

7. Conclusions. In [13, 14] Georgiou proved that to each choice of partial covariance sequence and numerator polynomial of the modeling filter there exists a rational covariance extension yielding a pole polynomial for the modeling filter, and he conjectured that this extension is unique so that it provides a complete parameterization of all rational covariance extensions. In [7] we proved this long-standing conjecture in the more general context of a duality between filtering and interpolation and showed that the problem is well posed in a very strong sense. In [6] we connected this solution to a certain Riccati-type matrix equation that sheds further light on the structure of this problem.

However, our proof in [7], as well as the existence proof of Georgiou [14], is nonconstructive. In this paper we presented a constructive proof of Georgiou's conjecture, which, although it is weaker than our result in [7], provides us for the first time with an algorithm for solving the problem of determining the unique pole polynomial corresponding to the given partial covariance sequence and the desired zeros.

This is done by means of a constrained convex optimization problem, which can be solved without explicitly computing the values of the cost function and which has the interesting property that the cost function is finite on the boundary but the gradient is not. In this context, Georgiou's conjecture is equivalent to establishing that there is a unique minimum in the *interior* of the feasible region. Specialized to the maximum entropy solution, this optimization problem was seen to be a dual to the well-known problem of maximizing the entropy gain.

Acknowledgments. We would like to thank Per Enqvist and Massoud Amin for their help with simulations.

REFERENCES

[1] N. I. Akhiezer, *The Classical Moment Problem and Some Related Questions in Analysis*, Hafner Publishing, New York, 1965.

[2] M. Aoki, *State Space Modeling of Time Series*, Springer-Verlag, Berlin, 1987.

[3] C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens, and A. S. House, *Reduction of speech spectra by analysis-by-synthesis techniques*, J. Acoust. Soc. Amer., 33 (1961), pp. 1725–1736.

[4] C. I. Byrnes and A. Lindquist, *On the geometry of the Kimura-Georgiou parameterization of modelling filter*, Internat. J. Control, 50 (1989), pp. 2301–2312.

[5] C. I. Byrnes and A. Lindquist, *Toward a solution of the minimal partial stochastic realization problem*, C. R. Acad. Sci. Paris Sér. I Math., 319 (1994), pp. 1231–1236.

[6] C. I. Byrnes and A. Lindquist, *On the partial stochastic realization problem*, IEEE Trans. Automat. Control, AC-42 (1997), pp. 1049–1070.

[7] C. I. Byrnes, A. Lindquist, S. V. Gusev, and A. S. Mateev, *A complete parametrization of all positive rational extensions of a covariance sequence*, IEEE Trans. Automat. Control, AC-40 (1995), pp. 1841–1857.

[8] J. A. Cadzow, *Spectral estimation: An overdetermined rational model equation approach*, Proc. IEEE, 70 (1982), pp. 907–939.

[9] C. Carathéodory, *Über den Variabilitätsbereich der Koeffizienten von Potenzreihen, die gegebene Werte nicht annehmen*, Math. Ann., 64 (1907), pp. 95–115.

[10] C. Carathéodory, *Über den Variabilitätsbereich der Fourierschen Konstanten von positiven harmonischen Functionen*, Rend. Circ. Mat. Palermo (2), 32 (1911), pp. 193–217.

[11] P. Delsarte, Y. Genin, Y. Kamp, and P. van Dooren, *Speech modelling and the trigonometric moment problem*, Philips J. Res., 37 (1982), pp. 277–292.

[12] P. Enqvist, forthcoming Ph.D. dissertation, Division of Optimization and Systems Theory, Royal Institute of Technology, Stockholm, Sweden.

[13] T. T. Georgiou, *Partial Realization of Covariance Sequences*, Ph.D. thesis, Center for Mathematical Systems Theory, University of Florida, Gainesville, FL, 1983.

[14] T. T. Georgiou, *Realization of power spectra from partial covariance sequences*, IEEE Trans. Acoust. Speech Signal Process., ASSP-35 (1987), pp. 438–449.

[15] Y. L. Geronimus, *Orthogonal Polynomials*, Consultants Bureau, New York, 1961.

[16] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, University of California Press, Berkeley, CA, 1958.

[17] U. Grenander and M. Rosenblatt, *Statistical Analysis of Stationary Time Series*, Almqvist & Wiksell, Stockholm, 1956.

[18] R. E. Kalman, *Realization of covariance sequences*, in Proc. Toeplitz Centennial, Tel Aviv, Israel, 1981, Oper. Theory Adv. Appl. 4, Birkhäuser, Basel, 1982, pp. 331–342.

[19] S. A. Kassam and H. V. Poor, *Robust techniques for signal processing*, Proc. IEEE, 73 (1985), pp. 433–481.

[20] S. M. Kay and S. L. Marple, Jr., *Spectrum analysis—A modern perspective*, Proc. IEEE, 69 (1981), pp. 1380–1419.

[21] H. Kimura, *Positive partial realization of covariance sequences*, in Modelling, Identification and Robust Control, C. I. Byrnes and A. Lindquist, eds., North–Holland, Amsterdam, 1987, pp. 499–513.

[22] A. Lindquist and G. Picci, *Canonical correlation analysis, approximate covariance extension, and identification of stationary time series*, Automatica J. IFAC, 32 (1996), pp. 709–733.

[23] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed., Addison–Wesley, Reading, MA, 1984.

[24] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, Springer-Verlag, Berlin, 1976.

[25] J. Makhoul, *Linear prediction: A tutorial review*, Proc. IEEE, 63 (1975), pp. 561–580.

[26] M. Minoux, Jr., *Mathematical Programming: Theory and Algorithms*, John Wiley, New York, 1986.

[27] B. Porat, *Digital Processing of Random Signals*, Prentice–Hall, Englewood Cliffs, NJ, 1994.

[28] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[29] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice–Hall, Englewood Cliffs, NJ, 1978.

[30] I. Schur, *On power series which are bounded in the interior of the unit circle* I *and* II, J. Reine Angew. Math., 148 (1918), pp. 122–145.

[31] O. Toeplitz, *Über die Fouriersche Entwicklung positiver Funktionen*, Rend. Circ. Mat. Palermo (2), 32 (1911), pp. 191–192.

[32] P. van Overschee and B. De Moor, *Subspace algorithms for stochastic identification problem*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 382–387.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 1, pp. 230–244

ROBUSTNESS WITH RESPECT TO DELAYS FOR EXPONENTIAL STABILITY OF DISTRIBUTED PARAMETER SYSTEMS*

RICHARD REBARBER[†] AND STUART TOWNLEY[‡]

Abstract. In this paper we address the question of whether the open-loop exponential growth rate of a linear system can be improved by a feedback in such a way that this improvement is robust with respect to small delays in the feedback loop. When the input operator is admissible, and the class of possible feedbacks consists of compact operators, we find that if a feedback can improve the exponential growth rate, then it can do so robustly. Furthermore, we find that if the control space is finite dimensional and a bounded feedback cannot be found to improve exponential stability, then a large class of *unbounded* feedbacks cannot improve the exponential growth rate robustly, even if such feedbacks can improve the exponential growth rate in the absence of delays.

Key words. robustness, delays, feedback, control, distributed parameter systems

AMS subject classifications. 93C20, 93C25, 93D09, 93D15, 93D25

PII. S0363012996312453

1. Introduction.

In this paper we are concerned with robustness of stability with respect to small delays in the feedback loop for distributed parameter systems. This issue has been analyzed a great deal when the stability in question is either input-output stability or modal stability. For a sample see Logemann, Rebarber, and Weiss [10] (for the frequency domain), Logemann and Rebarber [11], Datko, Lagnese, and Polis [3], Datko [4] (for modal stability of partial differential equations), and Logemann and Townley [12] (for modal stability of neutral systems). To our knowledge, there are no results available in the literature on robustness of *state space* stability for distributed parameter systems. If input-output stability is not robust, then in most situations it follows immediately that any state space stability is not robust. However, in general, robustness of input-output stability does not imply robustness of exponential stability, the most useful kind of state space stability. One reason for this is that, in general, input-output stability does not imply exponential stability. A more immediate difficulty is that the state space for the system with delay in the feedback is dependent on the delay.

In this paper we address the question of whether the exponential growth rate of a system can be improved by a feedback in such a way that this improvement is robust with respect to small delays in the feedback loop. When the input operator is admissible, and the class of feedbacks consists of compact operators, we find, in Theorem 3.5, that if a compact feedback can improve the exponential growth rate, then it can do so robustly when a natural state space is chosen for the delayed system. Sometimes it is possible to improve the exponential growth rate with an *unbounded* feedback when it is not possible with a bounded feedback. We find in Theorem 4.1 that if the control space is finite dimensional and a bounded feedback cannot be found

*Received by the editors November 25, 1996; accepted for publication (in revised form) December 30, 1997; published electronically October 7, 1998. This research was supported by NATO grant CRG 950179, the Human Capital and Mobility programme (project number CHRX-CT93-0402), and National Science Foundation grant DMS-9206986.

http://www.siam.org/journals/sicon/37-1/31245.html

[†]Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, NE 68588-0323 (rrebarbe@math.unl.edu).

[‡]Department of Mathematics and the Centre for Systems and Control Engineering, University of Exeter, Exeter EX4 4QE, UK (townley@maths.ex.ac.uk).

to improve exponential stability, then a large class of unbounded feedbacks cannot improve the exponential growth rate robustly, even if such feedbacks can improve the exponential growth when there are no delays present in the feedback loop.

We consider state space systems of the form

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{1.1}$$

where A is the infinitesimal generator of a strongly continuous semigroup $S(t)$ on X, a Hilbert space with norm $\|\cdot\|$. Let $R(s,A) = (sI-A)^{-1}$. Let

$$\mathbb{C}_\alpha = \{s \in \mathbb{C} \mid \mathrm{Re}\,(s) > \alpha\},$$

and for any Hilbert space X, let $H^\infty_\omega(X)$ be the set of all $\mathcal{B}(X)$-valued functions $f(s)$ which are holomorphic in $\mathbb{C}_\omega$ and for which $\sup_{s\in\mathbb{C}_\omega}\|f(s)\| < \infty$. The growth bound ω_0 of $S(t)$ is given by

$$\omega_0 = \lim_{t\to\infty} \frac{\ln(\|S(t)\|_{\mathcal{B}(X)})}{t} = \inf\{\omega \mid R(s,A) \in H^\infty_\omega(X)\}. \tag{1.2}$$

We let X_{-1} denote the closure of X in the norm $\|x\|_{-1} = \|R(\lambda,A)x\|$ for some λ in the resolvent set of A. The control space is a Hilbert space U with norm $\|\cdot\|_U$, and $B : U \to X_{-1}$ is an *admissible* input operator for $S(t)$; for details about admissibility see Weiss [20, 22]. Two consequences of the admissibility of B are that for every $F \in \mathcal{B}(X,U)$,

- $A - BF$ is the infinitesimal generator of a strongly continuous semigroup $S_F(t)$;
- the function $FR(s,A)B$ is analytic and bounded in $\mathbb{C}_\alpha$ for some $\alpha \in \mathbb{R}$.

For $F \in \mathcal{B}(X,U)$ we associate with (1.1) the observation

$$y(t) = Fx(t)\,. \tag{1.3}$$

Then (1.1) and (1.3) determine a linear open-loop, input-output system with transfer function $\mathbf{H}_F$ given by

$$\mathbf{H}_F(s) = FR(s,A)B\,. \tag{1.4}$$

Note that unity output feedback $u(t) = -y(t)$ for the input-output system (1.1) and (1.3) coincides with state feedback control $u(t) = -Fx(t)$ for (1.1).

We consider the closed-loop system with unity gain and delay ε in the feedback loop, described by Figure 1, equivalently (1.1), (1.3) with feedback control $u(t) = v(t) - Fx(t-\varepsilon)$. The block with transfer function $e^{-\varepsilon s}$ represents a delay by ε, where $\varepsilon \geq 0$. The transfer function $\mathbf{G}^\varepsilon$ for the closed-loop system (i.e., the transfer function from v to y) is given by

$$\mathbf{G}^\varepsilon(s) = \mathbf{H}_F(s)\left(I + e^{-\varepsilon s}\mathbf{H}_F(s)\right)^{-1}. \tag{1.5}$$

In the next section we obtain a sufficient condition for the existence of a feedback which robustly improves stability in a frequency domain sense; see Definition 2.6 and Proposition 2.7. In section 3 we show that the same condition is necessary and sufficient for the existence of a feedback which robustly improves stability in a natural state space setting; see Proposition 3.1, Definition 3.2, and Theorem 3.5. In section 4 we show that the necessary and sufficient condition of section 3 is also necessary and sufficient when a large class of unbounded feedbacks is used; see Theorem 4.1. In section 5 we give examples to illustrate our main results.

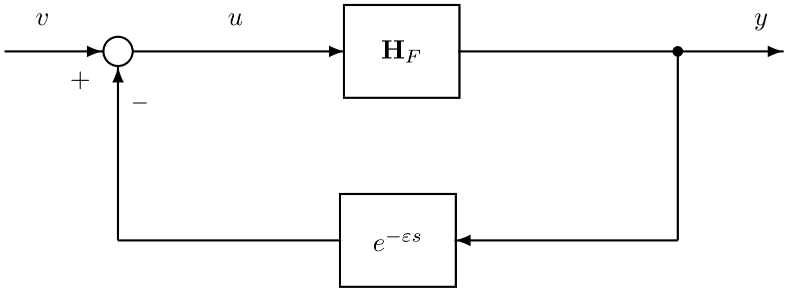


FIG. 1. *Feedback system with delay.*

2. Frequency domain robustness for state space systems. Let $\mathcal{K}(X,U)$ denote the set of compact operators from X into U. For systems of the form (1.1) we use the following definition of exponential stabilizability.

DEFINITION 2.1. *The system* (1.1) *is exponentially α-stabilizable if for every $\nu > \alpha$ there exists a feedback operator $F \in \mathcal{K}(X,U)$ and $M_\nu > 0$ such that*

$$\|S_F(t)\|_{\mathcal{B}(X)} \le M_\nu e^{\nu t}.$$

REMARK 2.2. *When the input space U is finite dimensional, so that $\mathcal{K}(X,U) = \mathcal{B}(X,U)$, Definition* 2.1 *coincides with the usual notion of closed-loop stabilizability* (*see*, *for instance*, *Curtain* [1] *or Jacobson and Nett* [8]). *Infinite dimensional input spaces U might arise in boundary control of partial differential equations in two or more spatial variables. In such cases*, *compact feedback operators arise naturally via integration over the boundary.*

We now summarize from Rebarber [14] necessary conditions for (1.1) to be exponentially α-stabilizable (see [1] or [8] for similar conditions in the case where the input space is finite dimensional). Let $\sigma(A)$ denote the spectrum of A, and suppose there exists $\nu \in \mathbb{R}$ such that

$$\sigma_u(A,\nu) := \{\lambda \mid \lambda \in \sigma(A) \text{ and } \operatorname{Re}(\lambda) \ge \nu\}$$

is bounded and can be isolated from the remaining part $\sigma_s(A,\nu)$ of the spectrum of A by a simple closed contour Γ. Following [9, Theorem 6.17], let

$$P = \int_\Gamma R(s,A)\,ds \qquad \text{and} \qquad N = I - P,$$

so that P is the projection corresponding to $\sigma_u(A,\nu)$ and N is the projection corresponding to $\sigma_s(A,\nu)$. Let $X_u = PX$ and $X_s = NX$, so that X can be represented by $X_u \oplus X_s$. Then (1.1) can be decomposed into the two systems

$$\dot{x}_u(t) = A_u x_u(t) + B_u u(t),$$

$$\dot{x}_s(t) = A_s x_s(t) + B_s u(t),$$

where $A_u = AP$, $A_s = AN$, $B_u = PB$, and $B_s = NB$.

LEMMA 2.3 (see [14, Theorem 1]). *If (A,B) is exponentially α-stabilizable, then for every $\nu > \alpha$, (A,B) has spectrum decomposition at ν in the sense that $\sigma_u(A,\nu)$ is bounded and can be isolated from $\sigma_s(A,\nu)$ by a simple closed contour, and*

(a) *$\sigma_u(A,\nu)$ consists of finitely many eigenvalues of finite multiplicity;*

(b) *A_s generates a strongly continuous semigroup $S_s(t)$ such that $\|S_s(t)\| \leq M_\nu e^{\nu t}$ for some $M_\nu \in \mathbb{R}$;*

(c) *(A_u, B_u) is a controllable pair.*

The following lemma shows that if $F \in \mathcal{K}(X,U)$, then the transfer function $\mathbf{H}_F$, for (1.1) and (1.3), is strictly proper in some right half-plane.

LEMMA 2.4. *Let A be the generator of a semigroup with growth bound ω_0, $B \in \mathcal{B}(U, X_{-1})$ be an admissible input operator for $S(t)$, and $F \in \mathcal{K}(X,U)$. Then for every $\gamma > \omega_0$,*

$$\lim_{\substack{|s|\to\infty \\ s\in\mathbb{C}_\gamma}} \|\mathbf{H}_F(s)\|_{\mathcal{B}(U)} = 0. \tag{2.1}$$

This result was proved in Curtain [1, Lemma 3.7] in the case where the input space U is finite dimensional.

Proof. F is compact if and only if F^* is compact (Schauder's theorem) and B^* is output admissible if B is input admissible (Salamon [17, Lemma 1.3.5]). Therefore, since the growth bound of the semigroup $S^*(t)$ generated by A^* is the same as that of $S(t)$,

$$\|\mathbf{H}_F(s)\|_{\mathcal{B}(U)} = \|FR(s,A)B\|_{\mathcal{B}(U)} = \|B^*R(\overline{s},A^*)F^*\|_{\mathcal{B}(U)}$$

holds for all $s \in \mathbb{C}_{\omega_0}$, where we make the identification between the Hilbert spaces X and U and their duals.

Step 1. We claim that for each $x \in X$ and $\gamma > \omega_0$,

$$\lim_{\substack{|s|\to\infty \\ s\in\mathbb{C}_\gamma}} \|B^*R(s,A^*)x\|_U = 0. \tag{2.2}$$

To see this, fix $x \in X$ and let $\{s_n\}_{n=1}^{\infty} \in \mathbb{C}_\gamma$ and $\lim_{n\to\infty}|s_n| = \infty$. If $\operatorname{Re}(s_n) \to \infty$, then it follows from [21, Proposition 2.3], that

$$\lim_{n\to\infty} \|B^*R(s_n,A^*)x\| = 0.$$

Therefore we assume that $\operatorname{Re}(s_n)$ is bounded, so there exists η such that $\gamma < \operatorname{Re}(s_n) \leq \eta$.

Write $s \in \mathbb{C}_\gamma$ as $a+ib$ for $a,b \in \mathbb{R}$. Then

$$R(s,A^*)x = \int_0^\infty e^{-at}e^{-ibt}S^*(t)x\,dt \tag{2.3}$$

$$= \int_0^{\pi/b} e^{-at}e^{-ibt}S^*(t)x\,dt + \int_{\pi/b}^\infty e^{-at}e^{-ibt}S^*(t)x\,dt. \tag{2.4}$$

By making a change of variable in the second integral in (2.4) and using the semigroup property of $S^*(t)$, we see that (2.4) can be written as

$$R(s,A^*)x = \int_0^{\pi/b} e^{-at}e^{-ibt}S^*(t)x\,dt - \int_0^\infty e^{-at}e^{-ibt}S^*(t)e^{-a\pi/b}S^*(\pi/b)x\,dt. \tag{2.5}$$

Multiplying both (2.3) and (2.5) on the left by $B^*/2$ and adding give

$$\begin{aligned} B^*R(s,A^*)x &= (1/2)\int_0^{\pi/b} e^{-at}e^{-ibt}B^*S^*(t)x\,dt \\ &\quad + (1/2)B^*\int_0^\infty e^{-at}e^{-ibt}S^*(t)[I - e^{-a\pi/b}S^*(\pi/b)]x\,dt. \end{aligned} \tag{2.6}$$

To analyze the first term in (2.6), note that by the output admissibility of B^*,

$$\|B^*S^*(\cdot)x\|_U \in L^2[0,\pi/b].$$

Hence, using Hölder's inequality, there exists M_1 such that

$$\left\|(1/2)\int_0^{\pi/b} e^{-at}e^{-ibt}B^*S^*(t)x\,dt\right\|_U \leq \frac{M_1}{\sqrt{b}} \tag{2.7}$$

for $s \in C_\gamma$.

We can rewrite the second term in (2.6) as

$$(1/2)B^*R(a+ib,A^*)[(I-e^{-a\pi/b}S^*(\pi/b))x].$$

By Proposition 2.3 in [21] we see that $\|B^*R(a+ib,A)\|_{\mathcal{B}(X,U)}$ is uniformly bounded for $a+ib \in \mathbb{C}_\gamma$, so there exists M_2 such that

$$\begin{aligned} &\left\|(1/2)B^*\int_0^{\infty} e^{-at}e^{-ibt}S^*(t)[I-e^{-a\pi/b}S^*(\pi/b)]x\,dt\right\|_U \\ &\leq M_2\|[I-e^{-a\pi/b}S^*(\pi/b)]x\| \end{aligned} \tag{2.8}$$

for $s \in \mathbb{C}_\gamma$. We see from the strong continuity of $S(t)$ that (2.8) → 0 as $|b| \to \infty$.

If we write $s_n = a_n + b_n$, then $\gamma < a_n < \eta$ and $|b_n| \to \infty$. We see from (2.6), (2.7), and (2.8) that $\|B^*R(s_n,A^*)x\|_U \to 0$ as $n \to \infty$.

Step 2. Using Step 1, it is clear that if F^* is of finite rank, then

$$\lim_{\substack{|s|\to\infty \\ s\in\mathbb{C}_\gamma}} \|B^*R(s,A^*)F^*\|_{\mathcal{B}(U)} = 0. \tag{2.9}$$

Step 3. Since any Hilbert space has the finite approximation property, it then follows, by using a standard finite rank approximation of the compact operator F^*, that (2.9) is true if B^* is admissible and F^* is compact. This completes the proof. □

Define

$$\alpha_{\min} = \inf\{\alpha \mid (A,B) \text{ is exponentially } \alpha\text{-stabilizable}\}.$$

The following result, which is of interest in its own right, will be used in the proof of Proposition 2.7.

PROPOSITION 2.5. *Suppose B is admissible and $F \in \mathcal{K}(X,U)$. If $\beta > \alpha_{\min}$, then*

$$\lim_{\substack{|s|\to\infty \\ s\in\mathbb{C}_\beta}} \|\mathbf{H}_F(s)\|_{\mathcal{B}(U)} = 0.$$

Proof. If $\alpha_{\min} < \nu_1 < \beta$, then (A,B) is exponentially ν_1-stabilizable. Let $\nu_1 < \nu_2 < \beta$. Using a spectrum decomposition for (A,B) at ν_2, we see that $\mathbf{H}_F(s) = \mathbf{H}_u(s) + \mathbf{H}_s(s)$, where $\mathbf{H}_u(s) = FR(s,A_u)B_u$ and $\mathbf{H}_s(s) = FR(s,A_s)B_s$. Since A_u has spectrum consisting of finitely many eigenvalues and F and B_u are bounded,

$$\lim_{\substack{|s|\to\infty \\ s\in\mathbb{C}_\beta}} \|\mathbf{H}_u(s)\|_{\mathcal{B}(U)} = 0.$$

Since $\beta > \nu_2$, we see from Lemma 2.4 that

$$\lim_{\substack{|s|\to\infty\\ s\in\mathbb{C}_\beta}} \|\mathbf{H}_s(s)\|_{\mathcal{B}(U)} = 0.$$

This completes the proof. □

We now formulate the notion of *robustly improving stability* in a frequency domain context. Let

$$\alpha_0(\mathbf{H}_F) = \inf\{\alpha \mid \mathbf{H}_F \in H^\infty_\alpha(U)\} \tag{2.10}$$

denote the input-output growth rate of $\mathbf{H}_F$. We are interested in using feedback to improve the stability of (1.1), that is, in finding a feedback F so that the growth rate of $\mathbf{G}^0$ is less than the growth rate of $\mathbf{H}_F$, and in determining whether this improved stability is robust with respect to delays. Because our interest is in improving stability, we emphasize the dependence on the growth rate α in our definitions.

Definition 2.6.

(a) $\mathbf{G}^\varepsilon$ *is* α-stable *if* $\mathbf{G}^\varepsilon \in H^\infty_\alpha(U)$. *It is well known that this property is equivalent to the property: if* $\hat{y} = \mathbf{G}^\varepsilon \hat{v}$ *and* $e^{-\alpha\cdot}v(\cdot) \in L^2([0,\infty),U)$, *then* $e^{-\alpha\cdot}y(\cdot) \in L^2([0,\infty),U)$.

(b) $\mathbf{G}^0$ *is* robustly α-stable with respect to delays *if there is an* $\varepsilon_0 > 0$ *such that for any* $\varepsilon \in [0,\varepsilon_0], \mathbf{G}^\varepsilon$ *is* α*-stable.*

(c) *We say that unity feedback* robustly improves the stability *of* $\mathbf{H}_F$ *if there exists* $\alpha < \alpha_0(\mathbf{H}_F)$ *such that* $\mathbf{G}^0$ *is robustly* α*-stable with respect to delays.*

We can now give a sufficient condition under which we can robustly improve stability in the sense of Definition 2.6.

Proposition 2.7. *If* $\alpha_{\min} < \omega_0$, *then there exists* $F \in \mathcal{K}(U,X)$ *for which unity feedback robustly improves the stability of* $\mathbf{H}_F(s)$.

Proof. By the definition of $\alpha_{\min}$, for any $\alpha \in (\alpha_{\min},\omega_0)$ we can find $F \in \mathcal{K}(X,U)$ so that the semigroup generated by $A - BF$ is α-stable. Hence $\mathbf{G}^0 \in H^\infty_\alpha(U)$, that is, $\mathbf{G}^0$ is α-stable. From Proposition 2.5 we see that for $\alpha \in (\alpha_{\min},\omega_0)$ and $F \in \mathcal{K}(X,U)$,

$$\limsup_{\substack{|s|\to\infty\\ s\in\mathbb{C}_\alpha}} \|\mathbf{H}_F(s)\|_{\mathcal{B}(U)} = 0.$$

Hence there exist $\delta \in (0,1)$ and $m_1 > 0$ such that

$$\|\mathbf{H}_F(s)\|_{\mathcal{B}(U)} < 1 - \delta, \qquad s \in \mathbb{C}_\alpha \cap \{|s| \geq m_1\}.$$

Therefore the Neumann series for $(I + e^{-\varepsilon s}\mathbf{H}_F(s))^{-1}$ converges uniformly in this region, so that $\mathbf{G}^\varepsilon(s)$ is bounded in $\mathbb{C}_\alpha \cap \{|s| \geq m_1\}$. Since $\mathbf{H}_F(s)(I + \mathbf{H}_F(s))^{-1}$ is bounded in the compact set $\mathbb{C}_\alpha \cap \{|s| \leq m_1\}$, a simple perturbation argument then shows that there exists $\varepsilon^* > 0$ such that for any $\varepsilon \in [0,\varepsilon^*]$, $\mathbf{H}_F(s)(I + e^{-\varepsilon s}\mathbf{H}_F(s))^{-1}$ is also bounded in $\mathbb{C}_\alpha \cap \{|s| < m_1\}$. It follows that $\mathbf{H}_F(I + e^{-\varepsilon\cdot}\mathbf{H}_F)^{-1}$ is α-stable for all $\varepsilon \in [0,\varepsilon^*]$ and therefore $\mathbf{G}^0$ is robustly α-stable with respect to delays.

Hence, to show that we have robustly improved the stability, all that remains is to show that we have in fact improved the stability. To do this it suffices to show that $\alpha_0(\mathbf{H}_F) = \omega_0$, which is not a priori guaranteed, since the input-output growth rate α_0 might be less than the exponential growth rate ω_0. Clearly $\alpha_0(\mathbf{H}_F) \leq \omega_0$. Moreover, since $(A - BF)$ is α-stable, $(A - \omega_0 I, B, F)$ is stabilizable and detectable in the sense discussed in Rebarber [15], and so results in [15] guarantee that input-output stability of $(A-\omega_0 I, B, F)$ is equivalent to exponential stability. Hence for $\varepsilon > 0$, exponential $(\omega_0 - \varepsilon)$-stability of $S(t)$ is equivalent to $(\omega_0 - \varepsilon)$-stability of $\mathbf{H}_F$. In particular, if $\mathbf{H}_F \in H^\infty_{\omega_0-\varepsilon}(U)$ for some $\varepsilon > 0$, then $S(t)$ is exponentially $(\omega_0-\varepsilon)$-stable. This would contradict the definition of ω_0, so $\alpha_0(\mathbf{H}_F)$ cannot be less than ω_0, and the proof is complete. □

3. State space robustness. In this section we consider robustness of exponential stability in a natural state space for the closed-loop system with delay in the feedback loop. We make the same assumptions on A, B, and F as in section 2, so that, in particular, B is admissible for $S(t)$ and F is compact. A closed-loop state space realization for the system described by Figure 1 with $v \equiv 0$ is given, formally, by the differential equation

$$\dot{x}(t) = Ax(t) - BFx(t-\varepsilon), \qquad y(t) = Fx(t), \tag{3.1}$$

where $x(t)$ takes values in the original state space X. In order to show that robustly improved stability in an input-output context can be strengthened to the robustly improved exponential stability for (3.1), we must first show that (3.1) is well-posed in some natural state space which takes account of the delay. A natural choice for this state space is $X \times L^2(-\varepsilon, 0; X)$. The well-posedness of (3.1) in this state space is summarized as follows.

PROPOSITION 3.1. *For each $x_0 \in X$ and $\phi_0 \in L^2(-\varepsilon, 0; X)$,*

(a) *there exists a unique $x(\cdot)$ evolving continuously in X which satisfies*

$$x(t) = S(t)x_0 - \int_0^t S(t-\tau)BFx(\tau-\varepsilon)d\tau, \qquad t \geq 0, \tag{3.2}$$

with $x(t) = \phi_0(t)$ for $t \in [-\varepsilon, 0]$. Furthermore, for each $t \geq 0$ there exists $M_1 > 0$ such that

$$\|x(t)\|^2 \leq M_1(\|x_0\|^2 + \|\phi\|^2_{L^2(-\varepsilon,0;X)}); \tag{3.3}$$

(b)

$$\dot{x}(t) = Ax(t) - BFx(t-\varepsilon) \quad \textit{for almost every (a.e.)} \quad t \geq 0, \tag{3.4}$$

holds as an equation in X_{-1};

(c) *let $\Phi(t)$ be defined for $t \geq 0$ by*

$$\Phi(t)(r) = x(t+r), \qquad r \in [-\varepsilon, 0]. \tag{3.5}$$

Then for each $t \geq 0$, $\Phi(t) \in L^2(-\varepsilon, 0; X)$ and there exists $M_2 > 0$ such that

$$\|\Phi(t)\|^2_{L^2(-\varepsilon,0;X)} \leq M_2(\|x_0\|^2_X + \|\phi\|^2_{L^2(-\varepsilon,0;X)}); \tag{3.6}$$

(d) *the formula*

$$\mathcal{T}_{F,\varepsilon}(t)\begin{pmatrix} x_0 \\ \phi_0 \end{pmatrix} = \begin{pmatrix} x(t) \\ \Phi(t) \end{pmatrix}$$

defines a strongly continuous semigroup on $X \times L^2(-\varepsilon, 0; X)$.

Proof. (a) The main step in the proof is the construction of a continuous solution trajectory $x(\cdot)$. This is achieved by piecing together segments on the intervals $[n\varepsilon, (n+1)\varepsilon)$, for $n \in \mathbb{N}$. Indeed, given that

$$x(t) = \phi_0(t) \quad \text{for} \quad t \in [-\varepsilon, 0],$$

we can define

$$x(t) = S(t-n\varepsilon)x(n\varepsilon) - \int_0^{t-n\varepsilon} S(t-n\varepsilon-\tau)BFx(\tau+n\varepsilon-\varepsilon)d\tau, \qquad n\varepsilon \le t \le (n+1)\varepsilon$$

recursively for $n \in \mathbb{N}$. By the admissibility of B, for every $n \in \mathbb{N}$, $x(\cdot)$ is continuous on $[0, (n+1)\varepsilon)$ with values in X and $Fx(\cdot) \in L^2(n\varepsilon, (n+1)\varepsilon; U)$. Using induction, it is easy to verify that $x(\cdot)$ satisfies (3.2) and (3.3) for $t \in [0, n\varepsilon)$ and all $n \in N$, thus proving (a).

(b) Since $x_0 \in X$ and $Fx(\cdot - \tau) \in L^2(0, T; U)$ for all T, this follows immediately from Theorem 3.9 in [20].

(c) This follows from the construction of Φ from $x(\cdot)$.

The proof of (d) is similar to the case when B is bounded (see for example Curtain and Zwart [2]). Indeed, given x_0, ϕ_0, and s, let $g(t) = x(t+s)$. Then

$$g(t) = x(t+s) = S(t+s)x_0 - \int_0^{t+s} S(t+s-\tau)BFx(\tau-\varepsilon)d\tau$$

$$= S(t)x(s) - \int_0^{t} S(t-\sigma)BFx(s+\sigma-\varepsilon)d\sigma\,.$$

Hence $g(\cdot)$ is the solution of (3.2) corresponding to initial conditions $g(0) = x(s)$ and $g(\theta) = x(s+\theta)$ for $\theta \in [-\varepsilon, 0]$. It now follows from the definition of $\mathcal{T}_{F,\varepsilon}(t)$ that

$$\mathcal{T}_{F,\varepsilon}(t+s)\begin{pmatrix} x_0 \\ \phi_0 \end{pmatrix} = \begin{pmatrix} g(t) \\ g(t+\cdot) \end{pmatrix} = \mathcal{T}_{F,\varepsilon}(t)\begin{pmatrix} x(s) \\ x(s+\cdot) \end{pmatrix} = \mathcal{T}_{F,\varepsilon}(t)\mathcal{T}_{F,\varepsilon}(s)\begin{pmatrix} x_0 \\ \phi_0 \end{pmatrix}$$

which is the semigroup property for $\mathcal{T}_{F,\varepsilon}(t)$. Strong continuity of the semigroup $\mathcal{T}_{F,\varepsilon}$ is an easy consequence of continuity of $x(\cdot)$ and the estimates (3.3) and (3.6). □

The following definition is a state space analogue of parts (b) and (c) of Definition 2.6.

Definition 3.2.

(a) *$F \in \mathcal{K}(X, U)$ robustly exponentially α-stabilizes* (1.1) *if there exists $\varepsilon_0 > 0$ such that for any $\varepsilon \in [0, \varepsilon_0]$ the semigroup $\mathcal{T}_{F,\varepsilon}$ is α-stable.*

(b) *F robustly improves the exponential stability of* (1.1) *if F robustly exponentially α-stabilizes* (1.1) *for some $\alpha < \omega_0$ (see* (1.2)*).*

We will show that exponential stabilization of the delay-free system by compact feedback is robust if the corresponding input-output stabilization is robust, in particular if $\alpha_{\min} < \omega_0$. Our approach is to show that the resolvent of the generator of $\mathcal{T}_{F,\varepsilon}(t)$ is bounded in $\overline{\mathbb{C}_\alpha}$, which shows, using Prüss [13, Proposition 2], that the semigroup $\mathcal{T}_{F,\varepsilon}(t)$ is α-stable. Instead of determining the generator of $\mathcal{T}_{F,\varepsilon}$, we use the fact that the resolvent of the generator is the Laplace transform of $\mathcal{T}_{F,\varepsilon}(t)$. To this end we take Laplace transforms of (3.2) and (3.5) which, after some rearrangement, gives

$$\hat{x}(s) = (I + e^{-\varepsilon s}R(s,A)BF)^{-1}R(s,A)\left[x_0 - e^{-\varepsilon s}BF\int_{-\varepsilon}^{0} e^{-st}\phi_0(t)dt\right] \tag{3.7}$$

and, for $r \in [-\varepsilon, 0]$,

$$\hat{\Phi}(s)(r) = e^{-sr}\left(\hat{x}(s) + \int_r^0 e^{-st}\phi_0(t)dt\right). \tag{3.8}$$

If we show that

$$\|(I + e^{-\varepsilon s}R(s,A)BF)^{-1}R(s,A)\|_{\mathcal{B}(X)}$$

and

$$\|(I + e^{-\varepsilon s}R(s,A)BF)^{-1}R(s,A)B\|_{\mathcal{B}(U,X)}$$

are bounded in $\overline{\mathbb{C}_\alpha}$, then it will follow that

$$\left\|\begin{pmatrix} \hat{x}(s) \\ \hat{\Phi}(s) \end{pmatrix}\|_{X\otimes L^2(-\varepsilon,0;X)} \leq M \| \begin{pmatrix} x_0 \\ \phi_0 \end{pmatrix}\right\|_{X\otimes L^2(-\varepsilon,0;X)} \tag{3.9}$$

for some $M > 0$ and all $s \in \overline{\mathbb{C}_\alpha}$ so that $\mathcal{T}_{F,\varepsilon}(t)$ is α-stable.

Lemma 3.3. *If there exists $M > 0$ such that for all $s \in \overline{\mathbb{C}_\alpha}$*

$$\|(I + R(s,A)BF)^{-1}R(s,A)\|_{\mathcal{B}(X)} \leq M, \tag{3.10a}$$

$$\|(I + e^{-\epsilon s}\mathbf{H}_F(s))^{-1}\|_{\mathcal{B}(U)} \leq M, \tag{3.10b}$$

$$\|\mathbf{H}_F(s)(I + e^{-\epsilon s}\mathbf{H}_F(s))^{-1}\|_{\mathcal{B}(U)} \leq M, \tag{3.10c}$$

then there exists $\tilde{M} > 0$ such that

$$\|(I + e^{-\varepsilon s}R(s,A)BF)^{-1}R(s,A)\|_{\mathcal{B}(X)} \leq \tilde{M}, \tag{3.11}$$

$$\|(I + e^{-\varepsilon s}R(s,A)BF)^{-1}R(s,A)B\|_{\mathcal{B}(U,X)} \leq \tilde{M} \tag{3.12}$$

for all $s \in \overline{\mathbb{C}_\alpha}$.

Remark 3.4. *Note that*

$$R(s, A - BF) = (I + R(s,A)BF)^{-1}R(s,A),$$

so (3.10a) *is equivalent to exponential α-stability of the closed-loop delay-free state space system, whilst* (3.10b) *and* (3.10c) *are equivalent to input-output α-stability of the closed-loop system with delay.*

Proof. Since B is admissible for the semigroup generated by A, it is also admissible for the semigroup generated by $A-BF$ (see Salamon [17]), which means that $R(s, A-BF)B$ is bounded in the same half-plane as $R(s, A - BF)$; see [21, Proposition 2.3]. Hence

$$\|(I + R(s,A)BF)^{-1}R(s,A)B\|_{\mathcal{B}(U,X)} \leq M. \tag{3.13}$$

Let

$$\mathbf{I}(s) = (I + R(s,A)BF)^{-1}R(s,A)B\Delta(s)S(s) + (I + R(s,A)BF)^{-1}R(s,A), \tag{3.14}$$

where

$$S(s) = (I + \mathbf{H}_F(s))(I + \mathbf{H}_F(s)e^{-\epsilon s})^{-1}F(I + R(s,A)BF)^{-1}R(s,A)$$

and

$$\Delta(s) = (1 - e^{-\epsilon s}).$$

We will show that

$$(I + e^{-\varepsilon s} R(s, A)BF)^{-1} R(s, A) = \mathbf{I}(s). \tag{3.15}$$

This would verify (3.11), since all the terms on the right side of (3.14) are, by hypothesis, bounded in $\overline{\mathbb{C}_\alpha}$. (3.12) would then follow because

$$(I + e^{-\varepsilon s} R(s, A)BF)^{-1} R(s, A)B = \mathbf{I}(s)B\,,$$

and, using (3.14), $\mathbf{I}(s)B$ is clearly bounded in $\overline{\mathbb{C}_\alpha}$, since the only difficult term is $S(s)B$, which is bounded in $\overline{\mathbb{C}_\alpha}$ by (3.13).

For brevity, we write $R(s, A) = R(s)$. Then

$$(I + e^{-\varepsilon s} R(s)BF)^{-1} R(s) - \mathbf{I}(s)$$

$$= (I + e^{-\varepsilon s} R(s)BF)^{-1}(R(s) - (I + e^{-\varepsilon s} R(s)BF))\mathbf{I}(s).$$

Substituting $\mathbf{I}(s)$ from (3.14) into this expression, we obtain, after a number of straightforward manipulations,

$$\begin{aligned} &(R(s) - (I + e^{-\varepsilon s} R(s)BF))\mathbf{I}(s) \\ &\quad = \big[(I + R(s)BF) - \{(I + e^{-\varepsilon s} R(s)BF) \\ &\qquad \times [(I + R(s)BF)^{-1} R(s)B\Delta(s)(I + \mathbf{H}_F(s))(I + \mathbf{H}_F(s)e^{-\epsilon s})^{-1}F + I]\}\big] \\ &\qquad \times (I + R(s)BF)^{-1} R(s) \\ &\quad = \Delta(s)\{R(s)BF - \big[(I + e^{-\varepsilon s} R(s)BF)(I + R(s)BF)^{-1} R(s)B \\ &\qquad \times\ (I + \mathbf{H}_F(s))(I + \mathbf{H}_F(s)e^{-\epsilon s})^{-1}F\big]\}(I + R(s)BF)^{-1} R(s). \end{aligned} \tag{3.16}$$

Recall that

$$\mathbf{H}_F(s) = FR(s)B\,,$$

$$F(I + R(s)BF)^{-1} R(s)B = \mathbf{H}_F(s)(I - \mathbf{H}_F(s))^{-1},$$

and

$$(I + R(s)BF)^{-1} R(s)B = R(s)B(I + \mathbf{H}_F(s))^{-1}\,.$$

Using these inside the second pair of square brackets in (3.16), expanding, and simplifying, we see that (3.16) is zero. Therefore

$$(I + e^{-\varepsilon s} R(s, A)BF)^{-1} R(s, A) = \mathbf{I}(s)$$

as claimed, and the proof is complete. □

THEOREM 3.5. *It is possible to robustly improve the exponential stability of* (1.1) *in the sense of Definition* 3.2 *if and only if* $\alpha_{\min} < \omega_0$.

Proof. The “only if” part follows immediately from the definition of $\alpha_{\min}$.

For the "if" part, we first show that for any $\alpha \in (\alpha_{\min}, \omega_0)$, there exists ε_1 such that (3.10a), (3.10b), and (3.10c) are true for $\varepsilon \in (0, \varepsilon_1)$. By the definition of $\alpha_{\min}$, for any $\alpha \in (\alpha_{\min}, \omega_0)$ we know that there exists $F \in \mathcal{K}(X, U)$ such that the delay-free closed-loop system is exponentially α-stable, i.e., so that (3.10a) holds for $s \in \overline{\mathbb{C}_\alpha}$. Since $\alpha < \omega_0$, the proof of Proposition 2.7 shows that unity feedback robustly improves the stability of $\mathbf{H}_F$. This establishes the existence of ε_1 such that (3.10b) and (3.10c) are bounded in $\overline{\mathbb{C}_\alpha}$ for any $\varepsilon \in (0, \varepsilon_1)$. Lemma 3.3 now implies that (3.11) and (3.12) are true in $\overline{\mathbb{C}_\alpha}$ for any $\varepsilon \in (0, \varepsilon_1)$. Therefore, (3.9) holds in $\overline{\mathbb{C}_\alpha}$ for some M and for any $\varepsilon \in (0, \varepsilon_1)$. This shows that the resolvent of the semigroup $\mathcal{T}_{F,\varepsilon}(t)$ is bounded in $\overline{\mathbb{C}_\alpha}$, so that $\mathcal{T}_{F,\varepsilon}(t)$ is α-stable. Hence F robustly improves the exponential stability of (1.1) in the sense of Definition 3.2. □

4. Unbounded feedback. We have seen that if B is an admissible input operator for $S(t)$, then we can robustly improve the exponential stability of (1.1) by compact feedback if and only if $\alpha_{\min} < \omega_0$. In [15] a definition of exponential stabilizability (which we call *regular stabilizability*) is given which allows F to be unbounded provided that (A, B, F) is a *regular system* and unity gain is an *admissible feedback*. We refer the reader to Weiss [22] for details about regular systems, and to Weiss [23] for details about admissible feedbacks. We mention here that if the feedback loop is closed with unity gain and unity gain is an admissible feedback for (A, B, F), then the closed-loop generator is $A - BF_L$, where F_L is the Lebesgue extension of F (see [22]), and that the formulas (1.4) and (1.5) hold when F is replaced by F_L. Roughly speaking, we say that (1.1) is regularly stabilizable if $A + BF_L$ generates an exponentially stable semigroup. With this kind of stabilizability it can be possible to improve the exponential stability of (1.1) even if $\alpha_{\min} = \omega_0$. (We should note here that in the definition of $\alpha_{\min}$ we still use the definition of stabilizability by compact F.) However, we will show that whenever the input space U is finite dimensional, any such improvement of stability is not robust with respect to delays.

THEOREM 4.1. *Suppose U is finite dimensional and $\alpha_{\min} = \omega_0$. If (1.1) is regularly μ-stabilizable, where $\mu < \omega_0$, then there exists sequences (ε_n) and (p_n) with*

$$\varepsilon_n > 0\,, \quad \lim_{n\to\infty} \varepsilon_n = 0\,, \quad p_n \in \mathbb{C}_{\omega_0}\,, \quad \lim_{n\to\infty} |\operatorname{Im} p_n| = \infty\,,$$

and such that for any $n \in \mathbb{N}$, p_n is a pole of $\mathbf{G}^{\varepsilon_n}$.

Proof. Suppose (1.1) is regularly μ-stabilizable with $\mu < \omega_0$ with the admissible feedback F. We break the proof up into two cases.

Case 1. Suppose

$$\limsup_{\substack{|s|\to\infty \\ s\in\mathbb{C}_{\omega_0}}} \|R(s, A)\|_{\mathcal{B}(X)} = \infty,$$

or, equivalently,

$$\limsup_{\substack{|s|\to\infty \\ s\in\mathbb{C}_0}} \|R(s, A - \omega_0 I)\|_{\mathcal{B}(X)} = \infty. \tag{4.1}$$

Since $A - BF_L$ is the generator of a semigroup with growth rate μ, $(A - \omega_0 I, B, F)$ is stabilizable and detectable in the sense given in [15]. Combining this with (4.1), we can conclude from (4.1) and Theorem 1.6 in [15] that

$$\limsup_{\substack{|s|\to\infty \\ s\in\mathbb{C}_0}} \|\mathbf{H}_F(s + \omega_0)\|_{\mathcal{B}(U)} = \limsup_{\substack{|s|\to\infty \\ s\in\mathbb{C}_{\omega_0}}} \|\mathbf{H}_F(s)\|_{\mathcal{B}(U)} = \infty. \tag{4.2}$$

By hypothesis, there exists F such that (A, B, F) is a regular system and $A - BF_L$ is μ-exponentially stable. From Proposition 4.1 in Weiss [22] it follows that the closed-loop transfer function $\mathbf{G}^0$ for (A, B, F) with unity gain is μ-stable. Hence, in the terminology of [10], the identity operator I stabilizes $\mathbf{H}_F(s + \omega_0)$, and the conclusion of Theorem 4.1 follows immediately from Lemma 8.5 in [10].

Case 2. Suppose

$$\limsup_{\substack{|s|\to\infty \\ s\in\mathbb{C}_{\omega_0}}} \|R(s, A)\|_{\mathcal{B}(X)} < \infty. \tag{4.3}$$

Since $u(t) = -F_L x(t)$ is such that for any $\alpha \in (\mu, \omega_0)$,

$$\int_0^\infty e^{-\alpha t} \left(\|u(t)\|_U^2 + \|x(t)\|^2\right) dt < \infty,$$

we see that (1.1) is *open-loop α-stabilizable* as defined in Rebarber and Zwart [16]. By Theorem 2.11 in [16], for any $\alpha \in (\mu, \omega_0)$ we have that $\{\lambda \in \sigma(A) \mid \operatorname{Re}\lambda \geq \alpha\}$ contains no finite accumulation point (although ∞ might be an accumulation point) and consists only of point spectrum with finite multiplicity; see also Theorem 3.5 in Zwart [24] for a related result. Using this fact along with (4.3), we see that

$$\sigma_1(A) = \{\lambda \in \sigma(A) \mid \operatorname{Re}\lambda = \omega_0\}$$

consists of finitely many eigenvalues of finite multiplicity. Since $\{\lambda \in \sigma(A) \mid \operatorname{Re}\lambda \geq \alpha\}$ contains no finite accumulation point, $\sigma_1(A)$ can be isolated from the remaining part, $\sigma_2(A)$, of $\sigma(A)$ by a simple closed contour Γ. We proceed as in [9], Theorem 6.17 (see also the development before Lemma 2.3 in this paper) to decompose (1.5) into

$$\dot{x}_1(t) = A_1 x_1(t) + B_1 u(t),$$

$$\dot{x}_2(t) = A_2 x_2(t) + B_2 u(t),$$

where $\sigma(A_1) = \sigma_1(A)$, $\sigma(A_2) = \sigma_2(A)$, and the x_1-subsystem is finite-dimensional. By (4.3) and the fact that A_2 does not have spectrum in $\overline{\mathbb{C}_{\omega_0}}$, $\|R(s, A_2)\|$ is bounded on $\overline{\mathbb{C}_{\omega_0}}$. Using [13, Proposition 2], this shows that A_2 generates an $(\omega_0 - \varepsilon)$-stable semigroup for some $\varepsilon > 0$. It follows that exponential α-stabilizability of (1.1) by bounded feedback is equivalent to α-stabilizability of the pair (A_1, B_1), which in turn is equivalent to controllability of the pair (A_1, B_1). Since $\alpha_{\min} = \omega_0$, (1.1) is not stabilizable by bounded feedback, we conclude that (A_1, B_1) is not controllable. In particular, the (A_1, B_1)-subsystem has an uncontrollable mode λ, with $\operatorname{Re}\lambda = \omega_0$. This implies that (1.5) cannot be regularly α-stabilized for any $\alpha < \omega_0$, which contradicts our hypotheses. Therefore Case 2 cannot hold, so Case 1 holds and the proof is complete. $\square$

5. Examples.

Example 1. To illustrate Theorem 3.5, we consider a heat equation with Neumann boundary control on part of the boundary. Let Ω be a bounded open domain in $\mathbb{R}^n$, $n = 2$ or 3, with Lipschitz boundary Γ, and let Γ_0 be a nonempty simply connected subset of Γ. Let $\langle \cdot, \cdot \rangle$ denote the real inner product in $L^2(\Omega)$, $\dot{w}$ denote differentiation with respect to time $t \geq 0$, $\partial w/\partial\nu$ denote the normal derivative of w on Γ, and Δ be the Laplacian in Ω. Unless otherwise stated, ξ will denote a variable in Ω and ζ will denote a variable in Γ.

We now consider the following heat equation with Neumann boundary control on Γ_0 and Dirichlet data on $\Gamma \setminus \Gamma_0$:

$$\dot{w}(\xi,t) = \Delta w(\xi,t), \qquad \xi \in \Omega, \quad t \geq 0, \tag{4.1a}$$

$$w(\zeta,t) = 0, \qquad \zeta \in \Gamma \setminus \Gamma_0, \quad t \geq 0, \tag{4.1b}$$

$$\frac{\partial w(\zeta,t)}{\partial \nu} = u(\zeta,t), \qquad \zeta \in \Gamma_0, \quad t \geq 0, \tag{4.1c}$$

where $u \in L^2[0,T;U]$, with $U = L^2(\Gamma_0)$.

We wish to represent (4.1a)–(4.1c) as a state space equation in $X = L^2(\Omega)$. Let the operator $A : \mathcal{D}(A) \to L^2(\Omega)$ be defined by

$$Aw = \Delta w, \qquad \mathcal{D}(A) = \left\{ z \in L^2(\Omega) \mid \Delta z \in L^2(\Omega), \ z|_{\Gamma\setminus\Gamma_0} = 0, \ \frac{\partial z}{\partial \nu}|_{\Gamma_0} = 0 \right\}.$$

It is well known that A is negative self-adjoint and has compact resolvent. In particular, for any $\beta \in \mathbb{R}$, $\sigma_u(A,\beta)$ consists of finitely many eigenvalues, each with finite multiplicity.

We define the Neumann map $\mathcal{N}$ as follows: If f is defined on Γ_0, then $\mathcal{N} f = R$ if R is the (distributional) solution to

$$\Delta R(\xi) = 0, \qquad \xi \in \Omega,$$

with boundary conditions

$$R(\zeta) = 0, \qquad \zeta \in \Gamma \setminus \Gamma_0,$$

$$R(\zeta) = f(\zeta), \qquad \zeta \in \Gamma_0,$$

where the boundary values are to be understood in the sense of trace.

It is a standard exercise to show that (4.1a)–(4.1c) can be put into the form (1.1) in X_{-1} (where w is now the dependent variable instead of x) with

$$B = -A\mathcal{N}.$$

In order to apply the results of sections 2 and 3, we need to show that B is admissible for the semigroup $S(t)$ generated by A. Let λ be in the resolvent set of A, and define $X_{-1/2}$ as the completion of X in the norm $\|(\lambda I - A)^{-1/2} \cdot \|$. It follows from Weiss [21] and Hansen and Weiss [7] that if the range of B is contained in $X_{-1/2}$, then B is admissible. From Grisvard [6],

$$\mathcal{D}((-A)^{1/2}) = \{ z \in H^1(\Omega) \mid z = 0 \text{ on } \Gamma \setminus \Gamma_0 \},$$

so from elliptic theory

$$\mathcal{N} \in \mathcal{B}(L^2(\Gamma), \mathcal{D}((-A)^{1/2})).$$

Hence we see that the range of $B = -A\mathcal{N}$ is contained in $X_{-1/2}$, so B is admissible. Therefore the results in sections 2 and 3 show that if $F \in \mathcal{K}(X,U)$ is such that $u(\zeta,t) = Fw(\xi,t)$ is an exponentially α-stabilizing feedback, then this stability is robust in the natural state spaces described in section 3.

As a specific example, let $\Omega = \{(x,y) \mid 0 < x < \pi,\ 0 < y < \pi\} \subset \mathbb{R}_2$ and Γ_0 be $\{0 \leq x \leq \pi\}$. The eigenvalues of A are easily computed to be

$$\lambda_{j,k} = -\left(\left(j + \frac{1}{2}\right)^2 + k^2\right), \qquad j \in \mathbb{N},\ k \in \mathbb{Z}^+,$$

with associated eigenvectors

$$\Phi_{j,k}(x,y) = \sin(kx)\cos((j+1/2)y), \qquad j \in \mathbb{N},\ k \in \mathbb{Z}^+.$$

Since the eigenvalue with the largest real part is $\lambda_{0,1} = -5/4$, the open loop system is exponentially $(-5/4)$-stable. The eigenvalue with the second largest real part is $\lambda_{1,1}$, which is equal to $-13/4$. In order to improve stability to $(-13/4)$-stability, we use a one dimensional bounded feedback of the form

$$F : X \to U, \quad (Fw) = \langle \eta, w \rangle g$$

with $\eta \in X$ and $g \in U$. With this feedback (4.1a)–(4.1c) are equivalent to

$$\dot{w}(t) = Aw(t) + Bgu(t), \qquad u(t) = \langle \eta, w(\cdot, t) \rangle.$$

We can find η so that F of this form exponentially $-13/4$-stabilizes (4.1a)–(4.1c) if and only if (A_u, Bg) is controllable on X_u, which is the closed span of $\Phi_{0,1}$ in X (see Theorem 6.1 in Triggiani [18]). In particular, suppose the trace $\Phi_{0,1}|_{\Gamma_0}$ (which is $\sin x$) is not orthogonal to g in $L^2(\Gamma)$. Then we can use the characterization of the adjoint of $A\mathcal{N}$ found in Triggiani [19] to show that $-A\mathcal{N}g$ is not orthogonal to $\Phi_{0,1}$ in $L^2(\Omega)$, and so (A_u, Bg) is controllable on X_u as required.

Example 2. Let X be a Hilbert space, and let A be a nonnegative self-adjoint operator on X with domain $\mathcal{D}(A)$. To illustrate Theorem 4.1 we consider a large class of hyperbolic systems of the form

$$\ddot{x}(t) + Ax(t) = Bu(t). \tag{4.2}$$

We assume that $B : U \to X_{-1}$, where U is finite dimensional, is admissible for this system in the sense that $\mathcal{B} = [0, B]^T$, is admissible for the semigroup on $H = \mathcal{D}(A^{1/2}) \oplus X$ generated by

$$\mathcal{A} = \begin{bmatrix} 0 & I \\ -A & 0 \end{bmatrix}.$$

The first-order state space equation for (4.2) is

$$\dot{z}(t) = \mathcal{A}x(t) + \mathcal{B}u(t). \tag{4.3}$$

It is easy to see that $\omega_0 = 0$ for this system. Since $\mathcal{A}$ does not have the spectrum decomposition described in Lemma 2.3, (4.3) is not stabilizable by bounded feedback, so $\alpha_{\min} = \omega_0$. Many systems of this form can be exponentially stabilized by regular feedback, for instance, many wave and beam equations in spatial dimension one. However, by Theorem 4.1 this stabilization cannot be robust with respect to small delays in the feedback loop. Related nonrobustness results for this system can be found in Datko and You [5].

Acknowledgments. We would like to thank Uwe Stroinski, George Weiss, and Mohammad Rammaha for several helpful discussions related to the work in this paper.

REFERENCES

[1] R.F. Curtain, *Equivalence of input-output stability and exponential stability for infinite dimensional systems*, Math. Systems Theory, 21 (1988), pp. 19–48.

[2] R. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Texts Appl. Math. 21, Springer-Verlag, New York, 1995.

[3] R. Datko, J. Lagnese, and M.P. Polis, *An example of the effect of time delays in boundary feedback stabilization of wave equations*, SIAM J. Control Optim., 24 (1986), pp. 152–156.

[4] R. Datko, *Not all feedback stabilized hyperbolic systems are robust with respect to small time delays in their feedbacks*, SIAM J. Control Optim., 26 (1988), pp. 697–713.

[5] R. Datko and Y.C. You, *Some second-order vibrating systems cannot tolerate small delays in their damping*, J. Optim. Theory Appl., 70 (1991), pp. 521–537.

[6] P. Grisvard, *Equations differentielles abstraites*, Ann. Sci. École Norm. Sup., 2 (1969), pp. 311–395.

[7] S. Hansen and G. Weiss, *The operator Carleson measure criterion for admissiblity of control operators for diagonal semigroups on l^2*, Systems Control Lett., 16 (1991), pp. 219–227.

[8] C.A. Jacobson and C.N. Nett, *Linear state-space systems in infinite-dimensional space: The role and characterization of joint stabilizability/detectability*, IEEE Trans. Automat. Control, 38 (1993) pp. 994–998.

[9] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, New York, 1980.

[10] H. Logemann, R. Rebarber, and G. Weiss, *Conditions for robustness and nonrobustness of the stability of feedback systems with respect to small delays in the feedback loop*, SIAM J. Control Optim., 34 (1996), pp. 572–600.

[11] H. Logemann and R. Rebarber, *The effect of small time-delays on the closed-loop stability of boundary control systems*, Math. Control Signals Systems, 9 (1996), PP. 123–151.

[12] H. Logemann and S. Townley, *The effect of small delays in the feedback loop on the stability of neutral systems*, Systems Control Lett., 27 (1996), pp. 267–274.

[13] J. Prüss, *On the spectrum of C_0-semigroups*, Trans. Amer. Math. Soc., 24 (1984), pp. 847–857.

[14] R. Rebarber, *Necessary conditions for exponential stabilizability of distributed parameter systems with infinite dimensional unbounded feedback*, Systems Control Lett., 14 (1990), pp. 241–248.

[15] R. Rebarber, *Conditions for the equivalence of internal and external stability for distributed parameter systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 994–998.

[16] R. Rebarber and H. Zwart, *Open loop stabilizability of infinite dimensional systems*, Math. Control Signals Systems, to appear.

[17] D. Salamon, *Control and Observation of Neutral Systems*, Res. Notes Math. 91, Pitman, Boston, MA, 1984.

[18] R. Triggiani, *On the stabilizability problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383–403.

[19] R. Triggiani, *Wave equation on a bounded domain with boundary dissipation: An operator approach*, J. Math. Anal. Appl., 137 (1989), pp. 438–461.

[20] G. Weiss, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.

[21] G. Weiss, *Two conjectures on the admissibility of control operators*, in Estimation and Control of Distributed Parameter Systems, W. Desch, F. Kappel, and K. Kunisch, eds., Birkhäuser-Verlag, Basel, 1991, pp. 367–378.

[22] G. Weiss, *Transfer functions of regular linear systems, part* I: *Characterization of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.

[23] G. Weiss, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.

[24] H.J. Zwart, *Some remarks on open and closed loop stabilizability for infinite dimensional systems*, in Control and Estimation of Distributed Parameter Systems, Internat. Ser. Numer. Math. 91, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser-Verlag, Basel, 1989, pp. 425–434.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 1, pp. 245–277

DETERMINATION OF THE INTERFACE BETWEEN THE FLUIDS IN A HALL–HÉROULT CELL FROM MEASUREMENTS OF ELECTRIC POTENTIALS AND CURRENTS ON THE ELECTRODES[*]

D. J. CEDIO-FENGYA[†], M. V. ROMERIO[‡], AND M. VOGELIUS[§]

Abstract. In this paper we apply the method of electrical impedance tomography to the problem of determining the interface between two fluid layers in a shallow electrolytic cell. We show that any nontrivial set of boundary currents and corresponding boundary voltages suffices to uniquely identify the interface. We propose a very simple reconstruction method that in an essential way uses the fact that the cell is of very small height. Estimates of the accuracy of this method are also provided. Finally we perform a number of computational experiments to demonstrate the efficiency of the method we propose. For the numerical reconstructions we use synthetic data generated from a discretized boundary integral formulation of the underlying conductivity problem.

Key words. electrical impedance tomography, Hall–Héroult cells, numerical reconstruction methods

AMS subject classifications. 35R30, 65N99

PII. S0363012996306327

1. Introduction.

Hall and Héroult are two names commonly associated with the electrolytic cells used for industrial aluminum production. The main components of a Hall–Héroult cell are the electrolytic bath and the liquid aluminum. At the top and the bottom of a cell are the arrays of anode and cathode blocks. The electrolytic bath contains, among other things, cryolite, and it is situated on top of the liquid aluminum layer. There are other less important layers which we disregard. The efficiency of the cell (a measure of the aluminum production) depends strongly on the location of the interface between the electrolytic bath and the aluminum layer. It is thus practically important to be able to control the location of the interface. However, before we can proceed to control the interface location, we need to be able to determine fairly accurately its current state. In industrial plants, cells are operating at a temperature of the order of 970°C. At this temperature the cryolite contained in the bath is chemically very active. Under such conditions measurements of the interface location by mechanical means encounter severe difficulties. In fact, the location cannot be determined accurately with such procedures. It is, on the other hand, very simple to make measurements of the voltages and currents at the top and bottom of the cell (at the anode and cathode blocks). These are exactly the kinds of measurements that are used in so-called electrical impedance tomography to determine the internal conductivity distribution. Since the conductivity properties of the electrolytic bath and the aluminum are fairly well known (and very different), the determination of the

[*] Received by the editors July 8, 1996; accepted for publication (in revised form) August 14, 1997; published electronically October 7, 1998.
http://www.siam.org/journals/sicon/37-1/30632.html

[†] Department of Mathematics, James Madison University, Harrisonburg, VA 22807 (fengyadj @math,jmu.edu).

[‡] Départemente de Mathématiques, Ecole Polytechnique Federale de Lausanne, CH 1015, Lausanne, Switzerland (romerio@dma.epfl.ch).

[§] Department of Mathematics, Rutgers University, New Brunswick, NJ 08903 (vogelius@ math.rutgers.edu). The research of this author was partially supported by National Science Foundation grant NSF DMS-9202042, Air Force Office of Scientific Research grant 89NM605, and URI-RIP grant F49620-93-I-0500.

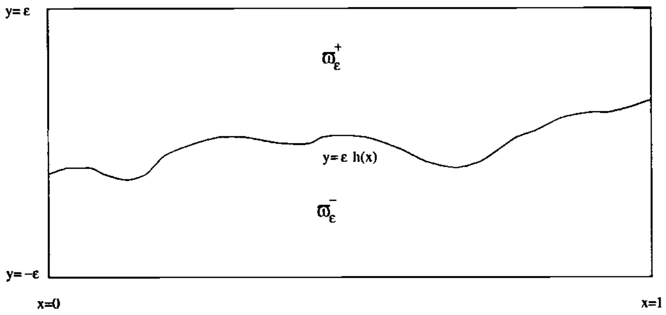


FIG. 1. *The domain Ω_ϵ representing the electrolytic cell.*

conductivity distribution is really equivalent to the determination of the location of the interface.

The purpose of this work is to demonstrate that electrical impedance tomography can be effectively applied to determine the location of an interface between two horizontal layers of different (but known) conductivity. In doing so we shall strongly make use of one geometric characteristic of a Hall–Héroult cell. Such a cell typically has a width of about 1/3 of its length and a height of about 1/50 of its length. Consequently we shall restrict our attention to domains of very small height. In this initial study we also suppose that the properties of the cell are independent of the width variable. We therefore model the cell by the two-dimensional rectangular domain $\Omega_\epsilon = (0,1) \times (-\epsilon, \epsilon)$ of length 1 and height 2ϵ.

As shown in Figure 1 we suppose that the domain Ω_ϵ is separated into two parts, ω_ϵ^+, and ω_ϵ^-, by the curve $y = \epsilon h(x)$. The subdomains $\omega_\epsilon^\pm$ are given by $\omega_\epsilon^+ = \{(x,y) :\ 0 < x < 1,\ \epsilon h(x) < y < \epsilon\}$ and $\omega_\epsilon^- = \{(x,y) :\ 0 < x < 1,\ -\epsilon < y < \epsilon h(x)\}$, respectively. The subdomain ω_ϵ^+ corresponds to the electrolytic bath and the subdomain ω_ϵ^- corresponds to the liquid aluminum layer; these have *different*, known, constant conductivities $a_+ > 0$ and $a_- > 0$ (we note that our theoretical results could very easily be generalized to the case in which a_+ and a_- are known, variable conductivities, with a proper jump across the curve $y = \epsilon h(x)$). We assume that the function $h(\cdot)$ satisfies $-1 < h(x) < 1$, $0 \le x \le 1$; for simplicity we also always (for our theoretical results) assume that $h(\cdot)$ is sufficiently smooth. We model the imposed boundary currents as distributed currents at the top and bottom boundaries. The two vertical boundaries $x = 0, 1$ are assumed to be insulated. The boundary value problem for the voltage potential u_ϵ therefore becomes

$$\nabla \cdot (a_+ \nabla u_\epsilon) = 0 \quad \text{in} \quad \omega_\epsilon^+, \tag{1a}$$

$$\nabla \cdot (a_- \nabla u_\epsilon) = 0 \quad \text{in} \quad \omega_\epsilon^-, \tag{1b}$$

$$\frac{\partial u_\epsilon}{\partial x} = 0 \quad \text{on} \quad x = 0, 1, \tag{1c}$$

$$a_+ \frac{\partial u_\epsilon}{\partial y} = f \text{ on } \ y = \epsilon, \qquad a_- \frac{\partial u_\epsilon}{\partial y} = g \text{ on } \ y = -\epsilon, \tag{1d}$$

with the continuity condition

$$u_\epsilon \text{ is continuous across the interface } y = \epsilon h(x) \tag{1e}$$

and the jump condition

$$\text{(1f)} \qquad a_+ \left(\frac{\partial}{\partial n}\right)_+ u_\epsilon = a_- \left(\frac{\partial}{\partial n}\right)_- u_\epsilon \quad \text{on the interface } y = \epsilon h(x).$$

Here n is a fixed normal direction to the interface $\{y = \epsilon h(x)\}$; $\left(\frac{\partial}{\partial n}\right)_+$ and $\left(\frac{\partial}{\partial n}\right)_-$ denote the derivatives in the direction n as we approach the interface from the top and the bottom, respectively. This boundary value problem has a corresponding weak formulation: find $u_\epsilon \in H^1(\Omega_\epsilon)$ such that

$$\text{(2)} \int_{\Omega_\epsilon} a_\epsilon \nabla u_\epsilon \nabla v \; dxdy = \int_{\partial\Omega_\epsilon \cap \{y=\epsilon\}} f \; v \; dx - \int_{\partial\Omega_\epsilon \cap \{y=-\epsilon\}} g \; v \; dx \quad \forall \; v \in H^1(\Omega_\epsilon).$$

Here the (piecewise-constant) conductivity distribution $a_\epsilon(x,y)$ is given by

$$a_\epsilon(x,y) = a_+ \;\text{ for } \epsilon h(x) < y < \epsilon \;, \;\; a_\epsilon(x,y) = a_- \;\text{ for } -\epsilon < y < \epsilon h(x).$$

Given $f, \, g \in H^{1/2}(0,1)$, with $\int_0^1 f \; dx = \int_0^1 g \; dx$, the weakly formulated boundary value problem (2) has a solution, and this solution is unique up to an additive constant. If the current distributions f and g are in $C^{2,\alpha}([0,1])$ for some $\alpha > 0$, then elliptic regularity theory ensures that u_ϵ is C^∞ inside each of the subdomains $\omega_\epsilon^\pm$ and is twice continuously differentiable up to the smooth parts of the boundaries $\partial\omega_\epsilon^\pm$. Such a smooth, weak solution u_ϵ satisfies (1) in a classical sense.

For our identification problem we suppose that measurements of the voltage potential u_ϵ are available all along the top and bottom boundaries $\{y = \pm\epsilon\}$. The main objective of this paper is to show that the interface $y = \epsilon h(x)$ may be very effectively determined from knowledge of these measurements. The present study may thus be seen as another piece of evidence that electrical impedance tomography is a quite effective method to determine special features of an otherwise known conductor; see also [1, 7, 8, 4, 15] and references therein. The setting we study here is as already mentioned very simplified when compared to the industrial, three-dimensional problem. We do however believe that the simplified setting exhibits many of the key characteristics. The added difficulties of the industrial, three-dimensional problem will be the focus of future work.

2. A two-dimensional identification result. The first question that naturally arises is whether knowledge of u_ϵ at the top and bottom boundaries suffices to identify the location of the interface between the constant conductivities a_+ and $a_-(\neq a_+)$. This question is somewhat analogous to the question of identifiability of an inaccessible boundary part (cf. [2]) and it is therefore not surprising that it is answered in the affirmative by the following theorem.

THEOREM 2.1. *Suppose $h_i \in C^{2,\alpha}([0,1])$, $i = 1,2$, and suppose $f, g \in C^{2,\alpha}([0,1])$ with at least one of these boundary currents not identically zero. Let u_ϵ^i, $i = 1,2$ denote solutions to* (2) (*or* (1)) *corresponding to the two interfaces $y = \epsilon h_i(x)$, $i = 1,2$, but with the same fixed set of boundary currents f, g. Then*

$$u_\epsilon^1 - u_\epsilon^2 = c \text{ (a single constant) } \; on \; \partial\Omega_\epsilon \cap (\{y = \epsilon\} \cup \{y = -\epsilon\})$$
$$\Rightarrow \;\; h_1 = h_2 \;\; on \; [0,1].$$

Proof. By adding the constant c to u_ϵ^2 we get another solution to (2). It therefore suffices to prove that $h_1 = h_2$ on $[0,1]$ whenever we have two solutions with $u_\epsilon^1 = u_\epsilon^2$

on $\partial\Omega_\epsilon \cap (\{y=\epsilon\} \cup \{y=-\epsilon\})$. Let $\tilde{\omega}_\epsilon^+$, $\tilde{\omega}_\epsilon^{int}$, and $\tilde{\omega}_\epsilon^-$ denote the sets

$$\begin{aligned}\tilde{\omega}_\epsilon^+ &= \{(x,y)\ :\ 0<x<1\ ,\ \ \epsilon\max(h_1(x),h_2(x)) < y < \epsilon\},\\ \tilde{\omega}_\epsilon^{int} &= \{(x,y)\ :\ 0<x<1\ ,\ \ \epsilon\min(h_1(x),h_2(x)) < y < \epsilon\max(h_1(x),h_2(x))\},\\ \tilde{\omega}_\epsilon^- &= \{(x,y)\ :\ 0<x<1\ ,\ \ -\epsilon < y < \epsilon\min(h_1(x),h_2(x))\}.\end{aligned}$$

Due to elliptic regularity theory the u_ϵ^i are both smooth solutions to $\triangle u_\epsilon^i = 0$ inside the three sets $\tilde{\omega}_\epsilon^+$, $\tilde{\omega}_\epsilon^{int}$, and $\tilde{\omega}_\epsilon^-$. Due to elliptic regularity theory and the regularity assumptions on f and g it also follows that the u_ϵ^i are C^2 up to (the interior of) the top and bottom boundaries $\{(x,y)\ :\ 0<x<1\ ,\ \ y=\pm\epsilon\}$ and that they satisfy the boundary conditions in a classical sense.

Now consider $v_\epsilon = u_\epsilon^1 - u_\epsilon^2$. Since, by assumption, v_ϵ has zero Cauchy data (v_ϵ and $\frac{\partial}{\partial y}v_\epsilon$ vanish) on the top and bottom boundaries $\{(x,y)\ :\ 0<x<1\ ,\ \ y=\pm\epsilon\}$, and since v_ϵ satisfies $\triangle v_\epsilon = 0$ in $\tilde{\omega}_\epsilon^+ \cup \tilde{\omega}_\epsilon^-$, it follows from unique continuation (Holmgren's uniqueness theorem) that $v_\epsilon \equiv 0$ in $\tilde{\omega}_\epsilon^+ \cup \tilde{\omega}_\epsilon^-$. It is well known that the u_ϵ^i, and thus v_ϵ, are continuous on $\overline{\Omega}_\epsilon$. We therefore get

$$v_\epsilon \equiv 0 \ \text{ on } \overline{\tilde{\omega}_\epsilon^+ \cup \tilde{\omega}_\epsilon^-}\,. \tag{3}$$

Let $\tilde{\omega}$ denote any nonempty, connected component of the open set $\tilde{\omega}_\epsilon^{int}$ (if one exists). Corresponding to $\tilde{\omega}$ we either have $h_1(x) < h_2(x)\ \forall\ (x,y) \in \tilde{\omega}$ or $h_2(x) < h_1(x)\ \forall\ (x,y)\in\tilde{\omega}$. Without loss of generality we suppose that

$$h_1(x) < h_2(x)\ \ \forall\ (x,y)\in\tilde{\omega}\ \ .$$

The set $\tilde{\omega}$ has an open, connected projection on the x-axis, hence there exists an open interval $I \subset (0,1)$ such that

$$\begin{gathered} h_1(x) < h_2(x)\ \ \forall\ x\in I \quad\text{and}\\ \tilde{\omega} = \{(x,y)\ :\ x\in I\ ,\ \ \epsilon h_1(x) < y < \epsilon h_2(x)\}.\end{gathered}$$

The following argument is slightly different depending on whether or not the endpoints of I fall in the interior of the interval $(0,1)$. Let us first consider the case where both endpoints fall in the interior. Afterwards we shall outline the changes necessary if this is not the case. The boundary of $\tilde{\omega}$ is now given by

$$\partial\tilde{\omega} = \{(x,y)\ :\ x\in\overline{I}\ ,\ \ y=\epsilon h_2(x)\} \cup \{(x,y)\ :\ x\in\overline{I}\ ,\ \ y=\epsilon h_1(x)\}. \tag{4}$$

Because of the identity (3) it follows immediately that $v_\epsilon = 0$ on $\partial\tilde{\omega}$. We also have that v_ϵ is continuous on the closure of $\tilde{\omega}$, and satisfies $\triangle v_\epsilon = 0$ in $\tilde{\omega}$. From the maximum principle it now follows that $v_\epsilon \equiv 0$ in $\tilde{\omega}$.

If one (or both) of the endpoints of I lies on the boundary of $[0,1]$, then $\partial\tilde{\omega}$ will generally contain at least one vertical line segment in addition to the curves $y=\epsilon h_i(x)$. If (by chance) there happens to be no such vertical boundary part(s) we proceed just as before to show that $v_\epsilon \equiv 0$ in $\tilde{\omega}$. Now suppose vertical lines form part of the boundary. The classical version of the maximum principle asserts that $v_\epsilon \equiv 0$ in $\tilde{\omega}$ or a nonzero extremal value is attained (in the interior) of one of the vertical boundary parts. As the normal derivative of v_ϵ vanishes along the vertical boundary part(s) the strong version of the maximum principle (Hopf's lemma) implies that we necessarily must have $v_\epsilon \equiv 0$ in $\tilde{\omega}$. In summary we have now verified that v_ϵ always vanishes in $\tilde{\omega}$. Since $\tilde{\omega}$ is any nonempty, connected component of $\tilde{\omega}_\epsilon^{int}$ this immediately shows that one of the following two statements must be true: 1) $v_\epsilon \equiv 0$ on the closure of $\tilde{\omega}_\epsilon^{int}$, or 2) $\tilde{\omega}_\epsilon^{int}$ is the empty set.

Let us now suppose that $\tilde{\omega}_\epsilon^{int} \neq \varnothing$, and therefore $v_\epsilon \equiv 0$ on the closure of $\tilde{\omega}_\epsilon^{int}$. By a combination with (3) this implies that

$$v_\epsilon = u_\epsilon^1 - u_\epsilon^2 = 0 \ \text{ on } \overline{\Omega}_\epsilon.$$

We consider a single, nonempty connected component $\tilde{\omega}$ of the set $\tilde{\omega}_\epsilon^{int}$ (such a component exists since we suppose that $\tilde{\omega}_\epsilon^{int}$ itself is nonempty). We shall analyze the case where the boundary of $\tilde{\omega}$ has the form (4). The case in which there are vertical boundary parts may be treated in a completely analogous manner due to the fact that the normal derivatives of u_ϵ^i, $i = 1, 2$ vanish on these vertical parts. Along the boundary curve $\{(x, y) \ : \ x \in I \ , \ \ y = \epsilon h_2(x)\}$ we have that $[\frac{\partial}{\partial n} u_\epsilon^1] = \left(\frac{\partial}{\partial n}\right)_+ u_\epsilon^1 - \left(\frac{\partial}{\partial n}\right)_- u_\epsilon^1 = 0$ (since u_ϵ^1 satisfies $\triangle u_\epsilon^1 = 0$ across this curve). Along this curve we also have that the conormal derivative of u_ϵ^2 satisfies the jump condition

$$\left[a_\epsilon \frac{\partial}{\partial n} u_\epsilon^2\right] = a_+ \left(\frac{\partial}{\partial n}\right)_+ u_\epsilon^2 - a_- \left(\frac{\partial}{\partial n}\right)_- u_\epsilon^2 = 0. \tag{5}$$

Here n denotes a fixed normal direction to the curve $y = \epsilon h_2(x)$; $\left(\frac{\partial}{\partial n}\right)_+$ and $\left(\frac{\partial}{\partial n}\right)_-$ denote the derivatives in the direction n as we approach the curve from the top and the bottom, respectively. Since $u_\epsilon^1 = u_\epsilon^2$ in all of Ω_ϵ we in particular have that $\left(\frac{\partial}{\partial n}\right)_\pm u_\epsilon^2 = \left(\frac{\partial}{\partial n}\right)_\pm u_\epsilon^1 = \frac{\partial}{\partial n} u_\epsilon^1$ along the curve $\{(x, y) \ : \ x \in I \ , \ \ y = \epsilon h_2(x)\}$. Since $a_+ \neq a_-$, the jump condition (5) therefore yields

$$\left(\frac{\partial}{\partial n}\right)_\pm u_\epsilon^2 = \frac{\partial}{\partial n} u_\epsilon^1 = 0$$

along the curve $\{(x, y) \ : \ x \in I \ , \ \ y = \epsilon h_2(x)\}$. By a completely identical argument we get that

$$\left(\frac{\partial}{\partial n}\right)_\pm u_\epsilon^1 = \frac{\partial}{\partial n} u_\epsilon^2 = 0$$

along the curve $\{(x, y) \ : \ x \in I \ , \ \ y = \epsilon h_1(x)\}$. Consider now the function u_ϵ^1. From the previous two identities we get that

$$\frac{\partial}{\partial n} u_\epsilon^1 = 0$$

along the boundary curves $\{(x, y) \ : \ x \in I \ , \ \ y = \epsilon h_i(x)\}$, $i = 1, 2$, of $\tilde{\omega}$. Since the function u_ϵ^1 also satisfies $\triangle u_\epsilon^1 = 0$ in $\tilde{\omega}$ and since it has finite energy (it is in $H^1(\tilde{\omega})$), it follows that $u_\epsilon^1 =$ constant in $\tilde{\omega}$. Unique continuation (Holmgren's uniqueness theorem) now implies that $u_\epsilon^1 =$ constant in all of Ω_ϵ, which contradicts the fact that the boundary currents f and g are not both identically zero. The assumption that the set $\tilde{\omega}_\epsilon^{int}$ is nonempty has thus led to a contradiction, and as a consequence it must necessarily be empty. This implies that $h_1(x) = h_2(x)$ for all $x \in [0, 1]$, exactly as desired. □

Remark. We note that the proof of Theorem 2.1 carries through virtually without any changes when the h_i are only assumed to be piecewise $C^{2,\alpha}$ and globally Lipschitz. This observation is relevant for interfaces such as the piecewise-linear ones we consider in connection with our numerical experiments. We also note that there is some similarity between the above proof and the proof of Lemma 2.1 in [7].

Theorem 2.1 guarantees the identifiability of h from the overdetermined data $u_\epsilon|_{y=\pm\epsilon}$, but it does not provide an explicit reconstruction formula. The derivation of such a reconstruction formula, in the case when the height of the cell (2ϵ) approaches zero, is the focus of the next two sections.

3. The asymptotic limit as $\boldsymbol{\epsilon \longrightarrow 0}$. In this section we shall study the limit of the voltage potential u_ϵ as $\epsilon \to 0$. Studies of such "thin domain" limits are quite common in continuum mechanics—for instance in connection with the derivation of "dimensionally reduced" equations for beams, plates, or shells (see [5, 14] and references therein). These studies have incorporated sandwich-like structures [13, 14] and they have extended to the modeling of electrolytic cells [3], but to the best of our knowledge distributions of "material properties" such as those encountered here have not previously been rigorously analyzed. Since this analysis is furthermore quite short and elementary we shall, for the convenience of the reader, include it here.

As is frequently the case it is also here preferable to "stretch" the domain Ω_ϵ, by introduction of the new variables $(x,z) = (x, y/\epsilon)$. In these new coordinates the domain becomes $\Omega = (0,1) \times (-1,1)$, and the corresponding voltage potential $v_\epsilon(x,z) = u_\epsilon(x,\epsilon z)$ satisfies the boundary value problem

$$\frac{\partial}{\partial x}\left(a(x,z)\frac{\partial v_\epsilon}{\partial x}\right) + \epsilon^{-2}\frac{\partial}{\partial z}\left(a(x,z)\frac{\partial v_\epsilon}{\partial z}\right) = 0 \quad \text{in} \quad \Omega, \tag{6a}$$

$$\frac{\partial v_\epsilon}{\partial x} = 0 \quad \text{at} \quad x = 0,1, \tag{6b}$$

$$a_+\frac{\partial v_\epsilon}{\partial z} = \epsilon f \ \text{ at } z = 1, \quad a_-\frac{\partial v_\epsilon}{\partial z} = \epsilon g \ \text{ at } z = -1. \tag{6c}$$

Here the conductivity $a(x,z)$ is independent of ϵ and is given by

$$a(x,z) = a_+ \ \text{ for } \ h(x) < z < 1, \quad a(x,z) = a_- \ \text{ for } \ -1 < z < h(x).$$

(6a) is the (customary) shorthand for

$$\frac{\partial}{\partial x}a_+\frac{\partial v_\epsilon}{\partial x} + \epsilon^{-2}\frac{\partial}{\partial z}a_+\frac{\partial v_\epsilon}{\partial z} = 0 \quad \text{for } h(x) < z < 1,$$

$$\frac{\partial}{\partial x}a_-\frac{\partial v_\epsilon}{\partial x} + \epsilon^{-2}\frac{\partial}{\partial z}a_-\frac{\partial v_\epsilon}{\partial z} = 0 \quad \text{for } -1 < z < h(x),$$

with the continuity condition

$$v_\epsilon \text{ is continuous across the interface } z = h(x),$$

and the jump condition

$$\left[a\frac{\partial}{\partial x}v_\epsilon\right]n_x + \epsilon^{-2}\left[a\frac{\partial}{\partial z}v_\epsilon\right]n_z = 0 \quad \text{on the interface } z = h(x).$$

By formally expanding $v_\epsilon(x,z)$ in powers of ϵ and matching terms of equal powers in (6) we obtain

$$v_\epsilon(x,z) = \epsilon^{-1}v^{(-1)}(x,z) + \epsilon v^{(1)}(x,z) + \epsilon^3 v^{(3)}(x,z) + \cdots,$$

where the functions $v^{(i)}$ satisfy the equations

$$\frac{\partial}{\partial x}a\frac{\partial}{\partial x}v^{(i-2)}+\frac{\partial}{\partial z}a\frac{\partial}{\partial z}v^{(i)}=0 \quad \text{in } \Omega, \tag{7}$$

the (vertical) boundary conditions $\frac{\partial}{\partial x}v^{(i)}=0$ at $x=0,1$, and the (horizontal) boundary conditions

$$a_+\frac{\partial}{\partial z}v^{(1)}=f \quad \text{at } z=1, \qquad a_-\frac{\partial}{\partial z}v^{(1)}=g \quad \text{at } z=-1, \tag{8a}$$

$$a_\pm\frac{\partial}{\partial z}v^{(i)}=0 \quad \text{at } z=\pm 1 \quad \text{for } i\neq 1. \tag{8b}$$

We note that as before (7) is to be interpreted in the strong sense in each of the subdomains $\omega^+=\{(x,z)\ :\ 0<x<1,\ h(x)<z<1\}$ and $\omega^-=\{(x,z)\ :\ 0<x<1,\ -1<z<h(x)\}$, but in a weak sense across their interface, i.e., all the $v^{(i)}$ are continuous, and furthermore

$$\left[a\frac{\partial}{\partial x}v^{(i-2)}\right]n_x+\left[a\frac{\partial}{\partial z}v^{(i)}\right]n_z=0 \tag{9}$$

along the interface $z=h(x)$. Here $n=(n_x,n_z)$ denotes a vector field normal to the interface, and $[\cdot]$ indicates, as before, the difference between the values as we approach the interface from above and below. As it turns out, these equations imply that all the v's corresponding to even superscripts vanish and that the first nonvanishing v corresponding to an odd superscript is $v^{(-1)}$. (7) and the boundary condition (8b) corresponding to $i=-1$ now read

$$\frac{\partial}{\partial z}a\frac{\partial}{\partial z}v^{(-1)}=0 \ \text{ in } \Omega, \qquad a_\pm\frac{\partial}{\partial z}v^{(-1)}=0 \quad \text{at } z=\pm 1.$$

This immediately implies that $v^{(-1)}(x,z)=v^{(-1)}(x)$ (a function of x only). We now multiply (7) with $i=1$ by any smooth $\phi(x)$ (a function of x only) and integrate over Ω. After integration by parts (and use of (8a) and (9) for $i=1$) this yields

$$\int_0^1 ((1-h(x))a_+ +(1+h(x))a_-)\,\frac{d}{dx}v^{(-1)}\frac{d}{dx}\phi\ dx=\int_0^1 (f-g)\phi\ dx.$$

In other words, $v^{(-1)}(x)$ is the solution to the two-point boundary value problem

$$\begin{aligned}\frac{d}{dx}\left(((1-h(x))a_+ +(1+h(x))a_-)\,\frac{d}{dx}v^{(-1)}\right)&=-f+g \quad \text{in } \ (0,1),\\ \frac{d}{dx}v^{(-1)}(x)&=0 \quad \text{at } \ x=0,1.\end{aligned} \tag{10}$$

The following proposition translates the previous formal analysis into a completely rigorous convergence statement. The proof of this proposition is based on a variational technique somewhat similar to that used in [12] to derive the beam equation in the constant coefficient case.

PROPOSITION 3.1. *Suppose the function h as well as the boundary currents f,g are in $C^{2,\alpha}([0,1])$ with $\int_0^1 f\ dx=\int_0^1 g\ dx$. Let $u_\epsilon(x,y)$ be a solution to* (1) (*or* (2)) *and let* $v^{(-1)}(x)$ *be a solution to* (10). *There exists a constant* C, *dependent on* f, g, h, *and* $a_{\pm}$ *but independent of the (small) positive parameter* ϵ, *such that*

$$\int_{\Omega_\epsilon} a_\epsilon(x,y)|\nabla(\epsilon u_\epsilon - v^{(-1)})|^2 \, dxdy \leq C\epsilon^3.$$

Proof. The solution u_ϵ has a very simple variational characterization, related to (2). It is a minimizer of the energy expression

$$\frac{1}{2}\int_{\Omega_\epsilon} a_\epsilon |\nabla u|^2 \, dxdy - \int_{\partial\Omega_\epsilon \cap \{y=\epsilon\}} fu \, dx + \int_{\partial\Omega_\epsilon \cap \{y=-\epsilon\}} gu \, dx$$

in the space $H^1(\Omega_\epsilon)$. Selecting as a test function $u = \epsilon^{-1}v^{(-1)}$ we therefore get

$$\begin{aligned}
&\frac{1}{2}\int_{\Omega_\epsilon} a_\epsilon |\nabla u_\epsilon|^2 \, dxdy - \int_{\partial\Omega_\epsilon \cap \{y=\epsilon\}} fu_\epsilon \, dx + \int_{\partial\Omega_\epsilon \cap \{y=-\epsilon\}} gu_\epsilon \, dx \\
&\qquad \leq \frac{1}{2}\epsilon^{-2}\int_{\Omega_\epsilon} a_\epsilon \left|\frac{d}{dx}v^{(-1)}\right|^2 \, dxdy - \epsilon^{-1}\int_0^1 (f-g)v^{(-1)} \, dx \\
&\qquad = \frac{1}{2}\epsilon^{-1}\int_0^1 ((1-h(x))a_+ + (1+h(x))a_-)\left|\frac{d}{dx}v^{(-1)}\right|^2 \, dx \qquad (11)\\
&\qquad -\epsilon^{-1}\int_0^1 (f-g)v^{(-1)} \, dx.
\end{aligned}$$

We note that the last expression is simply ϵ^{-1} times the energy (minimum) associated with the two-point boundary value problem (10). From the natural dual variational principle we get that

$$\begin{aligned}
&\frac{1}{2}\int_{\Omega_\epsilon} a_\epsilon |\nabla u_\epsilon|^2 \, dxdy - \int_{\partial\Omega \cap \{y=\epsilon\}} fu_\epsilon \, dx + \int_{\partial\Omega \cap \{y=-\epsilon\}} gu_\epsilon \, dx \\
&\qquad = \max_{\sigma \in V} -\frac{1}{2}\int_{\Omega_\epsilon} a_\epsilon^{-1}|\sigma|^2 \, dxdy, \qquad (12)
\end{aligned}$$

where the set V is characterized by

$$\begin{aligned}
V = \Bigg\{\sigma = (\sigma_x, \sigma_y) \in \left(L^2(\Omega_\epsilon)\right)^2 \ : \ &\int_{\Omega_\epsilon} \sigma \cdot \nabla v \, dxdy \\
&= \int_{\partial\Omega_\epsilon \cap \{y=\epsilon\}} fv \, dx - \int_{\partial\Omega_\epsilon \cap \{y=-\epsilon\}} gv \, dx \ \ \forall v \in H^1(\Omega_\epsilon)\Bigg\}.
\end{aligned}$$

In terms of the subsets ω_ϵ^+ and ω_ϵ^-, the elements of V are those vector fields that satisfy (in a "weak" distributional sense) the divergence constraints

$$\nabla \cdot \sigma = 0 \text{ in } \omega_\epsilon^+, \ \nabla \cdot \sigma = 0 \text{ in } \omega_\epsilon^-, \qquad (13a)$$

$$[\sigma_x]n_x + [\sigma_y]n_y = [\sigma \cdot n] = 0 \text{ on the interface } y = \epsilon h(x), \qquad (13b)$$

and the boundary conditions

$$\sigma_y = f \text{ at } y = \epsilon, \quad \sigma_y = g \text{ at } y = -\epsilon, \quad \text{and } \sigma_x = 0 \text{ at } x = 0, 1.$$

We note that (13) is just a "piecewise" formulation of the constraint $\nabla \cdot \sigma = 0$ in Ω_ϵ. Now define the test field $\tau = (\tau_x, \tau_y)$ as follows:

$$\tau_x = \epsilon^{-1} a_\epsilon(x,y) \frac{d}{dx} v^{(-1)}(x), \tag{14a}$$

$$\tau_y = -\epsilon^{-1} \frac{\partial}{\partial x} \left(\int_{-\epsilon}^{y} a_\epsilon(x,t)\ dt\ \frac{d}{dx} v^{(-1)}(x) \right) + g(x). \tag{14b}$$

We may alternatively express τ_y in terms of its formulas in the subdomains $\omega_\epsilon^{\pm}$

$$\begin{aligned} \tau_y(x,y) &= h'(x)(a_+ - a_-) \frac{d}{dx} v^{(-1)}(x) \\ &\quad -\epsilon^{-1} \left((y - \epsilon h(x)) a_+ + (\epsilon h(x) + \epsilon) a_- \right) \frac{d^2}{dx^2} v^{(-1)}(x) + g(x) \ \text{ in } \omega_\epsilon^+, \end{aligned} \tag{15a}$$

$$\tau_y(x,y) = -\epsilon^{-1}(y+\epsilon) a_- \frac{d^2}{dx^2} v^{(-1)}(x) + g(x) \ \text{ in } \omega_\epsilon^-. \tag{15b}$$

Due to the regularity assumptions about f, g, and h it follows that $v^{(-1)} \in C^3([0,1])$; τ is thus piecewise C^1 and certainly globally in $(L^2(\Omega_\epsilon))^2$. From the piecewise version of the definition of τ_y in combination with the definition of τ_x we get

$$\begin{aligned} \nabla \cdot \tau = \frac{\partial}{\partial x}\tau_x + \frac{\partial}{\partial y}\tau_y &= \epsilon^{-1} a_\epsilon \frac{d^2}{dx^2} v^{(-1)}(x) - \epsilon^{-1} a_\epsilon \frac{d^2}{dx^2} v^{(-1)}(x) \\ &= 0 \text{ in } \omega_\epsilon^+ \text{ and } \omega_\epsilon^-. \end{aligned} \tag{16}$$

For the same reason it also follows that

$$\begin{aligned} [\tau_x] n_x &= (\tau_x^+ - \tau_x^-) \frac{-\epsilon h'(x)}{\sqrt{1 + \epsilon^2 h'(x)^2}} \\ &= -(a_+ - a_-) \frac{d}{dx} v^{(-1)}(x) \frac{h'(x)}{\sqrt{1 + \epsilon^2 h'(x)^2}} \quad \text{on } y = \epsilon h(x) \text{ and} \\ [\tau_y]\, n_y &= (\tau_y^+ - \tau_y^-) \frac{1}{\sqrt{1 + \epsilon^2 h'(x)^2}} \\ &= \Bigg(h'(x)(a_+ - a_-) \frac{d}{dx} v^{(-1)}(x) \\ &\qquad -\epsilon^{-1} \left((y - \epsilon h(x)) a_+ + (\epsilon h(x) + \epsilon) a_- \right) \frac{d^2}{dx^2} v^{(-1)}(x) + g(x) \\ &\qquad +\epsilon^{-1}(y+\epsilon) a_- \frac{d^2}{dx^2} v^{(-1)}(x) - g(x) \Bigg) \frac{1}{\sqrt{1 + \epsilon^2 h'(x)^2}} \\ &= h'(x)(a_+ - a_-) \frac{d}{dx} v^{(-1)}(x)\ \frac{1}{\sqrt{1 + \epsilon^2 h'(x)^2}} \quad \text{on } y = \epsilon h(x), \end{aligned}$$

so that

$$[\tau_x] n_x + [\tau_y] n_y = 0 \text{ on the interface } y = \epsilon h(x). \tag{17}$$

It is also immediately clear from (10) and (14a) that the field τ satisfies the boundary condition

$$\begin{aligned} \tau_y &= -\frac{d}{dx} \left(((1 - h(x)) a_+ + (1 + h(x)) a_-) \frac{d}{dx} v^{(-1)} \right) + g(x) \\ &= f(x) \ \text{ at } y = \epsilon, \end{aligned} \tag{18a}$$

as well as

$$\tau_y = g(x) \ \text{ at } y=-\epsilon \ , \quad \text{and } \tau_x = 0 \ \text{ at } x=0,1. \tag{18b}$$

The statements (16)–(18) together assert that the field τ, defined by (14), is indeed an element of V. From (15) and the facts that g and h are in $C^{2,\alpha}$ and $v^{(-1)}$ is in C^3, we immediately get

$$\int_{\Omega_\epsilon} a_\epsilon^{-1}|\tau_y|^2 \ dxdy \leq C\epsilon. \tag{19}$$

By inserting τ into the dual variational expression, and using (12) and (19), we now get that

$$\begin{aligned}
&\frac{1}{2}\int_{\Omega_\epsilon} a_\epsilon|\nabla u_\epsilon|^2 \ dxdy - \int_{\partial\Omega_\epsilon\cap\{y=\epsilon\}} fu_\epsilon \ dx + \int_{\partial\Omega_\epsilon\cap\{y=-\epsilon\}} gu_\epsilon \ dx \\
&\quad \geq -\frac{1}{2}\int_{\Omega_\epsilon} a_\epsilon^{-1}|\tau|^2 \ dxdy \\
&\quad = -\frac{1}{2}\int_{\Omega_\epsilon} a_\epsilon^{-1}|\tau_x|^2 \ dxdy + O(\epsilon) \\
&\quad = -\frac{1}{2}\epsilon^{-1}\int_0^1 ((1-h(x))a_+ + (1+h(x))a_-)\left|\frac{d}{dx}v^{(-1)}\right|^2 \ dx + O(\epsilon) \\
&\quad = \frac{1}{2}\epsilon^{-1}\int_0^1 ((1-h(x))a_+ + (1+h(x))a_-)\left|\frac{d}{dx}v^{(-1)}\right|^2 \ dx \\
&\qquad\quad -\epsilon^{-1}\int_0^1 (f-g)v^{(-1)} \ dx \quad + O(\epsilon).
\end{aligned} \tag{20}$$

A combination of the upper bound (11) and the lower bound (20) gives that

$$\begin{aligned}
&\frac{1}{2}\int_{\Omega_\epsilon} a_\epsilon|\nabla u_\epsilon|^2 \ dxdy - \int_{\partial\Omega_\epsilon\cap\{y=\epsilon\}} fu_\epsilon \ dx + \int_{\partial\Omega_\epsilon\cap\{y=-\epsilon\}} gu_\epsilon \ dx \\
&\quad = \frac{1}{2}\epsilon^{-1}\int_0^1 ((1-h(x))a_+ + (1+h(x))a_-)\left|\frac{d}{dx}v^{(-1)}\right|^2 \ dx \\
&\qquad\quad -\epsilon^{-1}\int_0^1 (f-g)v^{(-1)} \ dx + O(\epsilon) \\
&\quad = \frac{1}{2}\int_{\Omega_\epsilon} a_\epsilon|\epsilon^{-1}\nabla v^{(-1)}|^2 \ dxdy \\
&\qquad\quad -\int_{\partial\Omega_\epsilon\cap\{y=\epsilon\}} f\epsilon^{-1}v^{(-1)} \ dx + \int_{\partial\Omega_\epsilon\cap\{y=-\epsilon\}} g\epsilon^{-1}v^{(-1)} \ dx + O(\epsilon).
\end{aligned} \tag{21}$$

It is well known that

$$\begin{aligned}
&\frac{1}{2}\int_{\Omega_\epsilon} a_\epsilon|\nabla(u_\epsilon - v)|^2 \ \ dxdy \\
&\quad = \frac{1}{2}\int_{\Omega_\epsilon} a_\epsilon|\nabla v|^2 \ dxdy - \int_{\partial\Omega_\epsilon\cap\{y=\epsilon\}} fv \ dx + \int_{\partial\Omega_\epsilon\cap\{y=-\epsilon\}} gv \ dx \\
&\quad -\left(\frac{1}{2}\int_{\Omega_\epsilon} a_\epsilon|\nabla u_\epsilon|^2 \ dxdy - \int_{\partial\Omega_\epsilon\cap\{y=\epsilon\}} fu_\epsilon \ dx + \int_{\partial\Omega_\epsilon\cap\{y=-\epsilon\}} gu_\epsilon \ dx\right)
\end{aligned}$$

for any $v \in H^1(\Omega_\epsilon)$. By a combination of this identity and the estimate (21) it follows immediately that

$$\frac{1}{2}\int_{\Omega_\epsilon} a_\epsilon |\nabla(u_\epsilon - \epsilon^{-1}v^{(-1)})|^2 \;\; dxdy = O(\epsilon)$$

or

$$\int_{\Omega_\epsilon} a_\epsilon |\nabla(\epsilon u_\epsilon - v^{(-1)})|^2 \;\; dxdy = O(\epsilon^3),$$

exactly as desired. □

Returning to the "stretched" domain $\Omega = (0,1) \times (-1,1)$ and the function $v_\epsilon(x,z) = u_\epsilon(x,\epsilon z)$, Proposition 3.1 asserts that

$$\text{(22a)} \int_\Omega a(x,z)\left(\left(\frac{\partial}{\partial x}(\epsilon v_\epsilon - v^{(-1)})\right)^2 + \epsilon^{-2}\left(\frac{\partial}{\partial z}(\epsilon v_\epsilon - v^{(-1)})\right)^2\right) \; dxdz \le C\epsilon^2$$

(recall $dxdy = \epsilon dxdz$) from which it also immediately follows that

$$\text{(22b)} \int_\Omega a(x,z)\left(\frac{\partial}{\partial z}v_\epsilon\right)^2 \; dxdz = \epsilon^{-2}\int_\Omega a(x,z)\left(\frac{\partial}{\partial z}(\epsilon v_\epsilon - v^{(-1)})\right)^2 \; dxdz \le C\epsilon^2.$$

The functions u_ϵ, v_ϵ, and $v^{(-1)}$ have so far been defined only up to additive constants. It will be convenient to make a specific choice for these constants. This can for instance be done by assuming that

$$\int_{\partial\Omega} v_\epsilon \; ds = \int_{\partial\Omega} v^{(-1)} \; ds = 0.$$

From (22a) it now follows that

$$\text{(23)} \qquad \int_\Omega \left(\epsilon v_\epsilon - v^{(-1)}\right)^2 \; dxdz \le C\epsilon^2,$$

and therefore

$$\|\epsilon v_\epsilon - v^{(-1)}\|_{H^1(\Omega)} \le C\epsilon.$$

Based on a standard trace estimate we conclude that

$$\text{(24)} \qquad \begin{aligned} \|\epsilon u_\epsilon(\cdot, \pm\epsilon) - v^{(-1)}(\cdot)\|_{H^{1/2}(0,1)} &= \|\epsilon v_\epsilon(\cdot, \pm 1) - v^{(-1)}(\cdot)\|_{H^{1/2}(0,1)} \\ &\le C\|\epsilon v_\epsilon - v^{(-1)}\|_{H^1(\Omega)} \\ &\le C\epsilon. \end{aligned}$$

This estimate of the approximation error in the boundary data, however, is not sufficient for our later analysis. We shall need an estimate of the H^1 approximation error. In order to establish such an estimate we need two a priori estimates of $\epsilon v_\epsilon - v^{(-1)}$ in higher-order norms. These estimates require stricter regularity assumptions about the boundary currents and the "interface" function $h(x)$. From now on we shall assume that $h \in C^{3,\alpha}([0,1])$ and that $f,\ g \in C^{4,\alpha}([0,1])$ with

$$\text{(25)} \qquad f'(x) = f'''(x) = 0 \quad \text{at} \;\; x = 0, 1.$$

Let c be a constant such that $\max_{x\in[0,1]} h(x) < c < 1$. This guarantees that the domain $\omega_c = \{(x,z) \ : \ 0 < x < 1, \quad c < z < 1\}$ lies inside $\omega^+ = \{(x,z) \ : \ 0 < x < 1, \quad h(x) < z < 1\}$. Select a smooth cutoff function $\phi(z)$ with the properties that $\phi \equiv 1$ near $z = 1$, $\phi \equiv 0$ near $z = c$, and $0 \le \phi(z) \le 1$, $c \le z \le 1$, and let w_ϵ denote the function

$$w_\epsilon(x,z) = (\epsilon v_\epsilon(x,z) - v^{(-1)}(x))\phi(z). \tag{26}$$

Our first a priori estimate is the following lemma.

LEMMA 3.2. *Suppose that h is in $C^{3,\alpha}([0,1])$ and suppose that the boundary currents f, g are in $C^{4,\alpha}([0,1])$ with f satisfying* (25). *Let w_ϵ denote the function defined by* (26). *There exists a constant C, dependent on the boundary currents f, g, the function h, and the conductivities $a_\pm$, but independent of $0 < \epsilon < 1$, such that*

$$\int_{\omega_c} \left(\left(\frac{\partial^2}{\partial x^2} w_\epsilon \right)^2 + \epsilon^{-4} \left(\frac{\partial^2}{\partial z^2} w_\epsilon \right)^2 + \epsilon^{-2} \left(\frac{\partial^2}{\partial x \partial z} w_\epsilon \right)^2 \right) \, dxdz \le C\epsilon^{-2}.$$

Proof. Let $\tilde{v}_\epsilon$ denote the even extension of v_ϵ across the boundary line $x = 0$. Since $\frac{\partial}{\partial x} v_\epsilon = 0$ at $x = 0$ it follows immediately that $\tilde{v}_\epsilon$ remains a solution to

$$\left(\frac{\partial^2}{\partial x^2} + \epsilon^{-2} \frac{\partial^2}{\partial z^2} \right) \tilde{v}_\epsilon = 0 \quad \text{in } (-1,1) \times (\max h(x), 1),$$

with

$$a_+ \frac{\partial}{\partial z} \tilde{v}_\epsilon|_{z=1} = \epsilon \tilde{f} \quad \text{on } z = 1, \ -1 < x < 1.$$

Here $\tilde{f}$ denotes the even extension of f across the point $x = 0$. Since $f \in C^{4,\alpha}([0,1])$ with $f'(0) = f'''(0) = 0$ it follows that $\tilde{f}$ is a $C^{4,\alpha}$ function, and consequently elliptic regularity theory gives that $\tilde{v}_\epsilon$ is C^4 on $(-1,1) \times [c,1]$. It immediately follows that v_ϵ is C^4 on the domain $[0,1) \times [c,1]$. By a similar extension argument across the boundary line $x = 1$ we get that v_ϵ is C^4 on the domain $(0,1] \times [c,1]$, and thus

$$v_\epsilon \in C^4([0,1] \times [c,1]) = C^4(\overline{\omega}_c).$$

Since f, g are $C^{4,\alpha}$ and h is $C^{3,\alpha}$, we have

$$v^{(-1)} \in C^4([0,1]).$$

Altogether we therefore conclude that

$$w_\epsilon \in C^4([0,1] \times [c,1]) = C^4(\overline{\omega}_c). \tag{27}$$

Using the definition (26) of w_ϵ we calculate

$$\begin{aligned} &\left(\frac{\partial^2}{\partial x^2} + \epsilon^{-2} \frac{\partial^2}{\partial z^2} \right) w_\epsilon(x,z) \\ &\quad = -\frac{d^2}{dx^2} v^{(-1)}(x)\phi(z) + 2\epsilon^{-1} \frac{\partial}{\partial z} v_\epsilon(x,z) \frac{d}{dz}\phi(z) \\ &\qquad + \epsilon^{-2}(\epsilon v_\epsilon(x,z) - v^{(-1)}(x)) \frac{d^2}{dz^2}\phi(z). \end{aligned}$$

From estimates (22b) and (23) it now immediately follows that, for $0 < \epsilon < 1$,

$$\left\| \left(\frac{\partial^2}{\partial x^2} + \epsilon^{-2} \frac{\partial^2}{\partial z^2} \right) w_\epsilon \right\|_{L^2(\omega_c)} \leq C\epsilon^{-1},$$

which is equivalent to

$$\int_{\omega_c} \left(\left(\frac{\partial^2}{\partial x^2} w_\epsilon \right)^2 + \epsilon^{-4} \left(\frac{\partial^2}{\partial z^2} w_\epsilon \right)^2 + 2\epsilon^{-2} \frac{\partial^2}{\partial x^2} w_\epsilon \frac{\partial^2}{\partial z^2} w_\epsilon \right) \, dxdz = O(\epsilon^{-2}). \tag{28}$$

Let us now for a moment consider the last term of this integral:

$$\begin{aligned} \int_{\omega_c} \frac{\partial^2}{\partial x^2} w_\epsilon \frac{\partial^2}{\partial z^2} w_\epsilon \, dxdz &= -\int_{\omega_c} \frac{\partial}{\partial x} w_\epsilon \frac{\partial}{\partial z} \left(\frac{\partial^2}{\partial x \partial z} w_\epsilon \right) \, dxdz \\ &= \int_{\omega_c} \left(\frac{\partial^2}{\partial x \partial z} w_\epsilon \right)^2 \, dxdz - \int_{\partial\Omega \cap \{z=1\}} \frac{\partial}{\partial x} w_\epsilon \frac{\partial^2}{\partial x \partial z} w_\epsilon \, dx \\ &= \int_{\omega_c} \left(\frac{\partial^2}{\partial x \partial z} w_\epsilon \right)^2 \, dxdz - \epsilon^2 a_+^{-1} \int_{\partial\Omega \cap \{z=1\}} \frac{\partial}{\partial x} w_\epsilon \, f'(x) \, dx. \end{aligned} \tag{29}$$

In the first integration by parts there are no boundary terms coming from the vertical boundaries, since $\frac{\partial}{\partial x} w_\epsilon$ vanishes at $x = 0, 1$. In the second integration by parts there is no boundary term coming from the lower boundary $\{0 < x < 1, \ z = c\}$, due to the fact that ϕ vanishes in a neighborhood of $z = c$. In the last identity we have used that

$$\frac{\partial^2}{\partial x \partial z} w_\epsilon = \frac{\partial^2}{\partial x \partial z} (\epsilon v_\epsilon - v^{(-1)})\phi = \epsilon^2 a_+^{-1} f'(x) \quad \text{at } z = 1.$$

A combination of (28) and (29) gives that

$$\begin{aligned} &\int_{\omega_c} \left(\left(\frac{\partial^2}{\partial x^2} w_\epsilon \right)^2 + \epsilon^{-4} \left(\frac{\partial^2}{\partial z^2} w_\epsilon \right)^2 + 2\epsilon^{-2} \left(\frac{\partial^2}{\partial x \partial z} w_\epsilon \right)^2 \right) \, dxdz \\ &\qquad = 2a_+^{-1} \int_{\partial\Omega \cap \{z=1\}} \frac{\partial}{\partial x} w_\epsilon \, f'(x) \, dx + O(\epsilon^{-2}). \end{aligned} \tag{30}$$

Since $\frac{\partial}{\partial x} w_\epsilon = 0$ at $z = c$,

$$\frac{\partial}{\partial x} w_\epsilon(x, 1) = \int_c^1 \frac{\partial^2}{\partial x \partial z} w_\epsilon(x, z) \, dz,$$

and therefore

$$\int_{\partial\Omega \cap \{z=1\}} \left(\frac{\partial}{\partial x} w_\epsilon \right)^2 \, dx \leq (1 - c) \int_{\omega_c} \left(\frac{\partial^2}{\partial x \partial z} w_\epsilon \right)^2 \, dxdz. \tag{31}$$

This last estimate together with (30) immediately implies that

$$\int_{\omega_c} \left(\left(\frac{\partial^2}{\partial x^2} w_\epsilon \right)^2 + \epsilon^{-4} \left(\frac{\partial^2}{\partial z^2} w_\epsilon \right)^2 + \epsilon^{-2} \left(\frac{\partial^2}{\partial x \partial z} w_\epsilon \right)^2 \right) \, dxdz \leq C\epsilon^{-2},$$

$0 < \epsilon < 1$, as desired. □

Let $c < c_2 < 1$ be chosen sufficiently close to 1 so that $\phi(z) \equiv 1$ for $z \in [c_2, 1]$, and let ϕ_2 denote a second cutoff function with the properties that $\phi_2 \equiv 1$ near $z = 1$, $\phi_2 \equiv 0$ near $z = c_2$, and $0 \leq \phi_2(z) \leq 1$ for $z \in [c_2, 1]$. Let $w_{2,\epsilon}$ denote the function defined by

$$w_{2,\epsilon}(x,z) = w_\epsilon(x,z)\phi_2(z) = (\epsilon v_\epsilon(x,z) - v^{(-1)}(x))\phi_2(z). \tag{32}$$

Our last a priori estimate is the following lemma.

LEMMA 3.3. *Suppose that h is in $C^{3,\alpha}([0,1])$ and suppose that f, g are in $C^{4,\alpha}([0,1])$ with f satisfying* (25). *Let $w_{2,\epsilon}$ denote the function defined by* (32) *and define $Dw_{2,\epsilon} = \frac{\partial}{\partial z} w_{2,\epsilon}$. There exists a constant C, dependent on the boundary currents f, g, the function h, and the conductivities $a_\pm$, but independent of $0 < \epsilon < 1$, such that*

$$\int_{\omega_{c_2}} \left(\left(\frac{\partial^2}{\partial x^2} Dw_{2,\epsilon} \right)^2 + \epsilon^{-4} \left(\frac{\partial^2}{\partial z^2} Dw_{2,\epsilon} \right)^2 + \epsilon^{-2} \left(\frac{\partial^2}{\partial x \partial z} Dw_{2,\epsilon} \right)^2 \right) dxdz \leq C\epsilon^{-2}.$$

Proof. A simple calculation, just as in the previous proof, yields that

$$\begin{aligned} &\left(\frac{\partial^2}{\partial x^2} + \epsilon^{-2} \frac{\partial^2}{\partial z^2} \right) Dw_{2,\epsilon}(x,z) \\ &\quad = -\frac{d^2}{dx^2} v^{(-1)}(x) \frac{d}{dz}\phi_2(z) + 2\epsilon^{-1} \frac{\partial^2}{\partial z^2} v_\epsilon(x,z) \frac{d}{dz}\phi_2(z) \\ &\quad\quad + 3\epsilon^{-1} \frac{\partial}{\partial z} v_\epsilon(x,z) \frac{d^2}{dz^2}\phi_2(z) \\ &\quad\quad + \epsilon^{-2}(\epsilon v_\epsilon(x,z) - v^{(-1)}(x)) \frac{d^3}{dz^3}\phi_2(z). \end{aligned} \tag{33}$$

From the estimate in Lemma 3.2 we have

$$\int_{\omega_c} \left(\frac{\partial^2}{\partial z^2} w_\epsilon \right)^2 dxdz \leq C\epsilon^2.$$

Since $\phi \equiv 1$ on $\omega_{c_2} \subset \omega_c$,

$$\frac{\partial^2}{\partial z^2} w_\epsilon = \frac{\partial^2}{\partial z^2}(\epsilon v_\epsilon - v^{(-1)}) = \epsilon \frac{\partial^2}{\partial z^2} v_\epsilon \quad \text{in } \omega_{c_2}.$$

The previous integral estimate therefore immediately gives

$$\int_{\omega_{c_2}} \left(\frac{\partial^2}{\partial z^2} v_\epsilon \right)^2 dxdz \leq C. \tag{34}$$

A combination of (33) with the estimates (22b), (23), and (34) now leads to

$$\left\| \left(\frac{\partial^2}{\partial x^2} + \epsilon^{-2} \frac{\partial^2}{\partial z^2} \right) Dw_{2,\epsilon} \right\|_{L^2(\omega_{c_2})} \leq C\epsilon^{-1},$$

or equivalently

$$\begin{aligned} \int_{\omega_{c_2}} &\left(\left(\frac{\partial^2}{\partial x^2} Dw_{2,\epsilon} \right)^2 + \epsilon^{-4} \left(\frac{\partial^2}{\partial z^2} Dw_{2,\epsilon} \right)^2 + 2\epsilon^{-2} \frac{\partial^2}{\partial x^2} Dw_{2,\epsilon} \frac{\partial^2}{\partial z^2} Dw_{2,\epsilon} \right) dxdz \\ &= O(\epsilon^{-2}). \end{aligned} \tag{35}$$

As in the proof of Lemma 3.2 let us now for a moment consider the last term of this integral:

$$\begin{aligned}
\int_{\omega_{c_2}} \frac{\partial^2}{\partial x^2} Dw_{2,\epsilon} \frac{\partial^2}{\partial z^2} Dw_{2,\epsilon}\, dxdz &= -\int_{\omega_{c_2}} \frac{\partial}{\partial x} Dw_{2,\epsilon} \frac{\partial}{\partial z}\left(\frac{\partial^2}{\partial x \partial z} Dw_{2,\epsilon}\right) dxdz \\
&= \int_{\omega_{c_2}} \left(\frac{\partial^2}{\partial x \partial z} Dw_{2,\epsilon}\right)^2 dxdz \\
&\quad - \int_{\partial\Omega \cap \{z=1\}} \frac{\partial}{\partial x} Dw_{2,\epsilon} \frac{\partial^2}{\partial x \partial z} Dw_{2,\epsilon}\, dx \\
&= \int_{\omega_{c_2}} \left(\frac{\partial^2}{\partial x \partial z} Dw_{2,\epsilon}\right)^2 dxdz \\
&\quad + \epsilon^5 \int_{\partial\Omega \cap \{z=1\}} f'(x) \frac{\partial^3}{\partial x^3} v_\epsilon\, dx.
\end{aligned} \tag{36}$$

In the first integration by parts there are no boundary terms coming from the vertical boundaries since $\frac{\partial}{\partial x} Dw_{2,\epsilon} = \frac{\partial}{\partial z}\frac{\partial}{\partial x} w_{2,\epsilon}$ vanishes at $x = 0, 1$. In the second integration by parts there is no boundary term coming from the lower boundary $\{0 < x < 1,\ z = c_2\}$, due to the fact that ϕ_2 vanishes in a neighborhood of $z = c_2$. In the last identity we have used that

$$\begin{aligned}
\frac{\partial}{\partial x} Dw_{2,\epsilon} &= \epsilon \frac{\partial}{\partial x}\frac{\partial}{\partial z} v_\epsilon = \epsilon^2 f' \quad \text{at } z = 1 \quad \text{and} \\
\frac{\partial^2}{\partial x \partial z} Dw_{2,\epsilon} &= \epsilon \frac{\partial}{\partial x}\frac{\partial^2}{\partial z^2} v_\epsilon = -\epsilon^3 \frac{\partial^3}{\partial x^3} v_\epsilon \quad \text{at } z = 1.
\end{aligned}$$

The last term in the right-hand side of (36) may be estimated by

$$\begin{aligned}
\left| \epsilon^5 \int_{\partial\Omega \cap \{z=1\}} f'(x) \frac{\partial^3}{\partial x^3} v_\epsilon\, dx \right| &= \left| \epsilon^5 \int_{\partial\Omega \cap \{z=1\}} f'''(x) \frac{\partial}{\partial x} v_\epsilon\, dx \right| \\
&= \left| \epsilon^4 \int_{\partial\Omega \cap \{z=1\}} f'''(x) \frac{\partial}{\partial x} (w_\epsilon + v^{(-1)})\, dx \right| \\
&\leq C\epsilon^4 \left(1 + \left(\int_{\partial\Omega \cap \{z=1\}} \left[\frac{\partial}{\partial x} w_\epsilon\right]^2 dx \right)^{1/2} \right).
\end{aligned}$$

If we now apply the inequality (31) we get

$$\left| \epsilon^5 \int_{\partial\Omega \cap \{z=1\}} f'(x) \frac{\partial^3}{\partial x^3} v_\epsilon\, dx \right| \leq C\epsilon^4 \left(1 + \left(\int_{\omega_c} \left[\frac{\partial^2}{\partial x \partial z} w_\epsilon\right]^2 dxdz \right)^{1/2} \right),$$

which, because of Lemma 3.2, immediately leads to

$$\left| \epsilon^5 \int_{\partial\Omega \cap \{z=1\}} f'(x) \frac{\partial^3}{\partial x^3} v_\epsilon\, dx \right| \leq C\epsilon^4. \tag{37}$$

A combination of (35), (36), and (37) gives

$$\int_{\omega_{c_2}} \left(\left(\frac{\partial^2}{\partial x^2} Dw_{2,\epsilon}\right)^2 + \epsilon^{-4} \left(\frac{\partial^2}{\partial z^2} Dw_{2,\epsilon}\right)^2 + 2\epsilon^{-2} \left(\frac{\partial^2}{\partial x \partial z} Dw_{2,\epsilon}\right)^2 \right) dxdz = O(\epsilon^{-2}),$$

exactly as desired. □

The estimate in Lemma 3.3 immediately implies that

$$\int_{\omega_{c_2}} \left(\frac{\partial}{\partial z}\frac{\partial^2}{\partial x^2} w_{2,\epsilon}\right)^2 \, dxdz \le C\epsilon^{-2}.$$

Proceeding along the lines of the argument that led to (31) we obtain

$$\begin{aligned}\int_0^1 \left(\frac{\partial^2}{\partial x^2}(\epsilon v_\epsilon(x,1) - v^{(-1)}(x))\right)^2 \, dx &= \int_0^1 \left(\frac{\partial^2}{\partial x^2} w_{2,\epsilon}(x,1)\right)^2 \, dx \\ &\le C\int_{\omega_{c_2}} \left(\frac{\partial}{\partial z}\frac{\partial^2}{\partial x^2} w_{2,\epsilon}\right)^2 \, dxdz \\ &\le C\epsilon^{-2}.\end{aligned}$$

Consequently

$$\begin{aligned}\int_0^1 \left(\frac{\partial^2}{\partial x^2}(\epsilon u_\epsilon(x,\epsilon) - v^{(-1)}(x))\right)^2 \, dx &= \int_0^1 \left(\frac{\partial^2}{\partial x^2}(\epsilon v_\epsilon(x,1) - v^{(-1)}(x))\right)^2 \, dx \\ &\le C\epsilon^{-2},\end{aligned}$$

or in combination with the estimate (24)

$$\|\epsilon u_\epsilon(\cdot,\epsilon) - v^{(-1)}(\cdot)\|_{H^2(0,1)} \le C\epsilon^{-1}.$$

If we again combine this estimate with (24) and use the fact that the Sobolev norms are logarithmically convex (with respect to the smoothness index), it follows that

$$\|\epsilon u_\epsilon(\cdot,\epsilon) - v^{(-1)}(\cdot)\|_{H^1(0,1)} \le C\epsilon^{1/3},$$

or in particular

$$\left\|\frac{\partial}{\partial x}(\epsilon u_\epsilon(\cdot,\epsilon) - v^{(-1)}(\cdot))\right\|_{L^2(0,1)} \le C\epsilon^{1/3}.$$

A completely similar argument with the boundary $\partial\Omega \cap \{z = 1\}$ $(\partial\Omega_\epsilon \cap \{y = \epsilon\})$ replaced by $\partial\Omega \cap \{z = -1\}$ $(\partial\Omega_\epsilon \cap \{y = -\epsilon\})$, and f replaced by g would lead to

$$\left\|\frac{\partial}{\partial x}(\epsilon u_\epsilon(\cdot,-\epsilon) - v^{(-1)}(\cdot))\right\|_{L^2(0,1)} \le C\epsilon^{1/3}.$$

In summary we have therefore proven the following result.

THEOREM 3.4. *Suppose that h is in $C^{3,\alpha}([0,1])$ and suppose that the boundary currents f, g are in $C^{4,\alpha}([0,1])$ with both f and g satisfying* (25). *There exists a constant C, dependent on the boundary currents f, g, the function h, and the conductivities $a_\pm$, but independent of $0 < \epsilon < 1$, such that*

$$\left\|\frac{\partial}{\partial x}(\epsilon u_\epsilon(\cdot,\pm\epsilon) - v^{(-1)}(\cdot))\right\|_{L^2(0,1)} \le C\epsilon^{1/3}.$$

By an extension of the arguments above it may be proven that $\frac{\partial}{\partial x}(\epsilon u_\epsilon)(\cdot,\pm\epsilon)$ converge to $\frac{d}{dx}v^{(-1)}(\cdot)$ uniformly. However, for reasons of space we do not present the details here. The above theorem plays a significant role in the derivation of a very accurate and robust reconstruction formula, valid for ϵ sufficiently small. This formula, which is derived in the next section, is conceptually somewhat similar to the linearized formula derived in [9], [10] for the reconstruction of a corroded (inaccessible) surface of a thin structure. However, the formula here is not based on any linearization and furthermore we provide estimates of its accuracy.

4. Reconstruction of the interface. Based on (10) we get that

$$(1-h(x))a_+ + (1+h(x))a_- = \frac{-\int_0^x (f-g)(t)\ dt}{\frac{d}{dx}v^{(-1)}(x)} \tag{38}$$

whenever $\int_0^x (f-g)(t)\ dt \neq 0$. On the right-hand side we know the numerator $-\int_0^1 (f-g)(t)\ dt$, however, we do not know the denominator $\frac{d}{dx}v^{(-1)}(x)$. What we do have access to are the boundary voltage derivatives $\frac{\partial}{\partial x}(\epsilon u_\epsilon)(x, \pm\epsilon)$. Let δ be an arbitrarily small positive number, and let I_δ denote the set

$$I_\delta = \left\{ x \in [0,1] \ : \ \left| \int_0^x (f-g)(t)\ dt \right| \geq 2\delta \max\{a_-, a_+\} \right\}.$$

From (38) we see that

$$2\min\{a_-, a_+\} \leq \frac{-\int_0^x (f-g)(t)\ dt}{\frac{d}{dx}v^{(-1)}(x)} \leq 2\max\{a_-, a_+\}.$$

It therefore immediately follows that

$$\left| \frac{d}{dx}v^{(-1)}(x) \right| \geq \delta \quad \text{for } x \in I_\delta.$$

From (38) we also see that

$$\operatorname{sign}\left(\frac{d}{dx}v^{(-1)}(x) \right) = \operatorname{sign}\left(-\int_0^x (f-g)(t)\ dt \right).$$

We now define approximations to $\frac{d}{dx}v^{(-1)}$:

$$DV_{\pm\epsilon}(x) = \begin{cases} \epsilon \frac{\partial}{\partial x} u_\epsilon(x, \pm\epsilon) & \text{when } |\epsilon \frac{\partial}{\partial x} u_\epsilon(x, \pm\epsilon)| \geq \delta, \\ \delta \ \operatorname{sign}\left(-\int_0^x (f-g)(t)\ dt \right) & \text{otherwise.} \end{cases} \tag{39}$$

We note that since the data $\epsilon \frac{\partial}{\partial x} u_\epsilon(x, \pm\epsilon)$ are at our disposal, the functions $DV_{\pm\epsilon}$ may easily be computed. The two corresponding reconstructed interfaces $h_{\pm\epsilon}$ are then given by

$$(1-h_{\pm\epsilon}(x))a_+ + (1+h_{\pm\epsilon}(x))a_- = \frac{-\int_0^x (f-g)(t)\ dt}{DV_{\pm\epsilon}(x)}, \tag{40}$$

or more explicitly

$$h_{\pm\epsilon}(x) = \frac{-\int_0^x (f-g)(t)\ dt}{(a_- - a_+)DV_{\pm\epsilon}(x)} - \frac{a_- + a_+}{a_- - a_+}. \tag{41}$$

That these formulas do indeed provide good approximations to the true "interface" function on I_δ is guaranteed by the following theorem.

THEOREM 4.1. *Suppose that h is in $C^{3,\alpha}([0,1])$ and that the boundary currents f, g are in $C^{4,\alpha}([0,1])$ with both f and g satisfying* (25). *Let $h_{\pm\epsilon}$ be as defined by* (39) *and* (41) *for some $\delta > 0$. There exists a constant C, dependent on f, g, h, and $a_\pm$ but independent of $0 < \epsilon < 1$ and δ such that*

$$\|h - h_{\pm\epsilon}\|_{L^2(I_\delta)} \leq C\delta^{-2}\epsilon^{1/3}.$$

If the function $x \to \int_0^x (f-g)(t)\ dt$ has only finitely many zeros on $[0,1]$ and if all these zeros are simple, then there exists a constant K such that $|[0,1] \setminus I_\delta| \leq K\delta$.

Proof. For $x \in I_\delta$ we have that

$$|\frac{d}{dx}v^{(-1)}(x)| \geq \delta \quad \text{and} \quad |DV_{\pm\epsilon}(x)| \geq \delta.$$

By subtraction of (40) from (38) it therefore follows that

$$\begin{aligned}
&|h(x) - h_{\pm\epsilon}(x)| \\
&\quad \leq \frac{1}{|a_- - a_+|}\left|\frac{-\int_0^x (f-g)(t)\ dt}{\frac{d}{dx}v^{(-1)}(x)} - \frac{-\int_0^x (f-g)(t)\ dt}{DV_{\pm\epsilon}(x)}\right| \\
&\quad = \frac{\left|\int_0^x (f-g)(t)\ dt\right|}{|a_- - a_+|\left|\frac{d}{dx}v^{(-1)}(x)\right| |DV_{\pm\epsilon}(x)|}\left|DV_{\pm\epsilon}(x) - \frac{d}{dx}v^{(-1)}(x)\right| \qquad (42)\\
&\quad \leq \frac{\|f\|_{L^2(0,1)} + \|g\|_{L^2(0,1)}}{\delta^2|a_- - a_+|}\left|DV_{\pm\epsilon}(x) - \frac{d}{dx}v^{(-1)}(x)\right| \\
&\quad = C\delta^{-2}\left|DV_{\pm\epsilon}(x) - \frac{d}{dx}v^{(-1)}(x)\right| \qquad \text{for } x \in I_\delta.
\end{aligned}$$

The definition of $DV_{\pm\epsilon}$, and the fact that $\frac{d}{dx}v^{(-1)}$ and $-\int_0^x (f-g)(t)\ dt$ have the same sign, guarantee that

$$\left|DV_{\pm\epsilon}(x) - \frac{d}{dx}v^{(-1)}(x)\right| \leq \left|\epsilon\frac{\partial}{\partial x}u_\epsilon(x,\pm\epsilon) - \frac{d}{dx}v^{(-1)}(x)\right| \quad \text{for } x \in I_\delta.$$

From a combination of this estimate with (42) it now immediately follows that

$$|h(x) - h_{\pm\epsilon}(x)|^2 \leq C\delta^{-4}\left|\epsilon\frac{\partial}{\partial x}u_\epsilon(x,\pm\epsilon) - \frac{d}{dx}v^{(-1)}(x)\right|^2 \quad \text{for } x \in I_\delta.$$

Integration over I_δ, extraction of the square root, and use of the estimate in Theorem 3.4 lead to the desired estimate for $\|h - h_{\pm\epsilon}\|_{L^2(I_\delta)}$.

The estimate on the measure of the set $[0,1] \setminus I_\delta$ is obvious, given that $\int_0^x (f - g)(t)\ dt$ is continuously differentiable with only finitely many simple zeros. □

If the function $x \to \int_0^x (f-g)(t)\ dt$ has only finitely many zeros which are all simple, and if we were to change the definition of $h_{\pm\epsilon}$ a little just to make sure that it stays bounded:

$$\tilde{h}_{\pm\epsilon}(x) = \begin{cases} h_{\pm\epsilon}(x) & \text{when } |h_{\pm\epsilon}(x)| \leq 1, \\ \text{sign}(h_{\pm\epsilon}(x)) & \text{otherwise ,} \end{cases}$$

then Theorem 4.1, for the choice $\delta = \epsilon^{2/15}$, immediately implies that

$$\|h - \tilde{h}_{\pm\epsilon}\|_{L^2(0,1)} \leq C(\delta^{-2}\epsilon^{1/3} + \delta^{1/2}) = 2C\epsilon^{1/15}.$$

The above analysis gives no indication of how to proceed if the function $\int_0^x (f - g)(t)\, dt$ vanishes in an entire interval (c, d). We shall now formally find an appropriate approximation to $h(x)$ in this case as well. Since $\int_0^x (f-g)(t)\, dt = 0$ in (c, d) it follows from (10) that $\frac{d}{dx}v^{(-1)} = 0$ in (c, d). We return to the asymptotic expansion at the beginning of section 3 in order to calculate the second (nontrivial) term, corresponding to the function $v^{(1)}(x, z)$. This function satisfies the weak formulation of the equation

$$\frac{\partial}{\partial x} a \frac{\partial}{\partial x} v^{(-1)} + \frac{\partial}{\partial z} a \frac{\partial}{\partial z} v^{(1)} = 0 \quad \text{in } \Omega,$$

with the boundary conditions

$$a_+ \frac{\partial}{\partial z} v^{(1)} = f \quad \text{at } z = 1 \ , \quad a_- \frac{\partial}{\partial z} v^{(1)} = g \quad \text{at } z = -1.$$

By integration of these equations and use of the fact that $\frac{d}{dx}v^{(-1)}(x) = 0$ (and $f(x) = g(x)$) for $x \in (c, d)$ we get that $v^{(1)}(x, z)$, for $x \in (c, d)$, has the form

$$v^{(1)}(x, z) = \begin{cases} \frac{1}{a_+} f(x)(z - h(x)) + b(x) & \text{for } h(x) < z < 1, \\ \frac{1}{a_-} f(x)(z - h(x)) + b(x) & \text{for } -1 < z < h(x), \end{cases}$$

where $b(x)$ is some function of x alone. We now consider the expression $\epsilon^{-1}(u_\epsilon(x, \epsilon) - u_\epsilon(x, -\epsilon))$; from the asymptotic expansion at the beginning of section 3 it formally follows that

$$\begin{aligned} \epsilon^{-1}(u_\epsilon(x, \epsilon) - u_\epsilon(x, -\epsilon)) &= \epsilon^{-1}(v_\epsilon(x, 1) - v_\epsilon(x, -1)) \\ &= \frac{1}{a_+} f(x)(1 - h(x)) + \frac{1}{a_-} f(x)(1 + h(x)) + O(\epsilon^2) \\ &= \left(\frac{1}{a_-} - \frac{1}{a_+}\right) f(x)h(x) + \left(\frac{1}{a_-} + \frac{1}{a_+}\right) f(x) + O(\epsilon^2) \end{aligned}$$

for $x \in (c, d)$. Provided the boundary current f has no zeros in the interval (c, d), we thus arrive at a formula for an approximation to h:

$$h_{1,\epsilon}(x) = \frac{u_\epsilon(x, \epsilon) - u_\epsilon(x, -\epsilon)}{\epsilon f(x) \left(\frac{1}{a_-} - \frac{1}{a_+}\right)} - \frac{a_+ + a_-}{a_+ - a_-}. \tag{43}$$

If f has simple isolated zeros, then a procedure similar to that before could be applied; on the other hand, if f has an entire interval of zeros inside (c, d), then the determination of higher-order terms in the asymptotic expansion for v_ϵ would be required in order to obtain a formula for an approximation to h (in terms of the boundary data).

The derivation of the formula (43) is entirely formal. In order to provide a rigorous justification we would need estimates of how well $\epsilon^{-1}(u_\epsilon(x, \epsilon) - u_\epsilon(x, -\epsilon))$ approximates $v^{(1)}(x, 1) - v^{(1)}(x, -1)$. We are confident that such estimates can be proven along the lines of our proofs in section 3, but we have not carried out this analysis.

Remarks. So far we have assumed that a_+ and a_- are fixed positive constants (i.e., they are bounded away from 0 and ∞). There is one degenerate case that is of high interest in the context of Hall–Héroult cells, namely the case in which a_+ is very small (nearly zero) relative to a_-. The formula we derived for $h_{-\epsilon}$ makes sense even for $a_+/a_- = 0$. In this extreme case it reads

$$h^0_{-\epsilon}(x) = \frac{-\int_0^x (f-g)(t)\ dt}{a_- DV^0_{-\epsilon}(x)} - 1,$$

with

$$a_- DV^0_{-\epsilon}(x) = \begin{cases} a_-\epsilon\frac{\partial}{\partial x}u_\epsilon(x,-\epsilon) & \text{when } |a_-\epsilon\frac{\partial}{\partial x}u_\epsilon(x,-\epsilon)| \geq \delta', \\ \delta' \text{ sign}\left(-\int_0^x (f-g)(t)\ dt\right) & \text{otherwise.} \end{cases}$$

The small parameter δ' has taken the role of δa_- in our earlier notation. We expect this formula to provide an acceptable approximation to $h(x)$, at least for $x \in I^0_\delta = \{x \ :\ |\int_0^x (f-g)(t)\ dt| \geq 2\delta'\}$, an expectation which is confirmed by our numerical experiments. The situation is completely different when it comes to the formula for $h_{+\epsilon}$, since this refers to boundary data of u_ϵ along $y = \epsilon$. In the extreme case, $a_+/a_- = 0$, the top boundary of the domain Ω_ϵ has been completely isolated from the bottom subdomain, ω^+_ϵ, where the limit of the rescaled solutions $(a_- u_\epsilon)$ appears to be well defined. On the top boundary, indications are that $a_- DV_+\epsilon$ becomes infinitely large as a_+/a_- tends to zero, and therefore the formula $h_{+\epsilon}$ does not provide a good approximation to h, rather it converges to -1.

If we multiply by a_+ in both the numerator and the denominator (of the first term) of the right-hand side of (43), then we may formally insert $a_+/a_- = 0$ to get

$$h^0_{1,\epsilon}(x) = -\frac{a_+\epsilon^{-1}(u_\epsilon(x,\epsilon) - u_\epsilon(x,-\epsilon))}{f(x)} + 1.$$

It appears that the rescaled top boundary data $a_+\epsilon^{-1}u_\epsilon(x,\epsilon)$ has a well-defined limit as a_+/a_- tends to zero and that the rescaled bottom data $a_+\epsilon^{-1}u_\epsilon(x,-\epsilon)$ tends to zero as a_+/a_- tends to zero. Our numerical experiments indicate that the limiting formula provides a reliable approximation to h (recall that this corresponds to intervals in which $\int_0^x (f-g)(t)\ dt \equiv 0$).

5. Integral equation formulation. In this section we describe how we numerically approximate the solution u_ϵ of the boundary value problem (1) (the forward problem) by solving a system of integral equations. We use this solution to generate the data needed for the numerical reconstruction of the interface h.

As in the calculation of the asymptotic limit (section 3), it is preferable to introduce the new coordinate $z = y/\epsilon$ and solve the following rescaled problem (the same as (6)) in the ϵ-independent domain $\Omega = (0,1) \times (-1,1)$:

$$\frac{\partial}{\partial x}\left(a(x,z)\frac{\partial v_\epsilon}{\partial x}\right) + \epsilon^{-2}\frac{\partial}{\partial z}\left(a(x,z)\frac{\partial v_\epsilon}{\partial z}\right) = 0 \quad \text{in} \quad \Omega,$$
$$\frac{\partial v_\epsilon}{\partial x} = 0 \quad \text{at} \quad x = 0, 1,$$
$$a_+\frac{\partial v_\epsilon}{\partial z} = \epsilon f \ \text{ at } z = 1, \quad a_-\frac{\partial v_\epsilon}{\partial z} = \epsilon g \ \text{ at } z = -1.$$

Here the conductivity $a(x,z)$ is given by

$$a(x,z) = \begin{cases} a_+, & (x,z) \in \omega^+, \\ a_-, & (x,z) \in \omega^-, \end{cases}$$

with

$$\omega^+ = \{(x,z) :\ 0 < x < 1,\ h(x) < z < 1\},$$
$$\omega^- = \{(x,z) :\ 0 < x < 1,\ -1 < z < h(x)\}.$$

The physical voltage potential u_ϵ is related to v_ϵ by $u_\epsilon(x,y) = v_\epsilon(x, y/\epsilon)$. We note that v_ϵ satisfies

$$\frac{\partial^2}{\partial x^2} v_\epsilon + \frac{1}{\epsilon^2}\frac{\partial^2}{\partial z^2} v_\epsilon = 0 \tag{44}$$

in each of the subdomains ω^+ and ω^-. We also note that the fundamental solution corresponding to the differential operator in (44) is given by

$$N_\epsilon((x,z);(\tilde{x},\tilde{z})) = \frac{\epsilon}{4\pi}\log((\tilde{x}-x)^2 + \epsilon^2(\tilde{z}-z)^2). \tag{45}$$

Unless otherwise specified $n = (n_x, n_z)$ denotes the exterior unit normal associated with each of the subdomains ω^+ and ω^-. Γ_I denotes the interface between ω^+ and ω^-, that is,

$$\Gamma_I = \{(x, h(x)) : 0 \le x \le 1\}.$$

We now derive an integral formulation for the values of v_ϵ along the boundary $\partial\omega^+ \cup \partial\omega^- = \partial\Omega \cup \Gamma_I$. It is convenient to introduce the following operators L and D:

$$Lv = \frac{\partial^2 v}{\partial x^2} + \frac{1}{\epsilon^2}\frac{\partial^2 v}{\partial z^2}, \qquad Dv = \frac{\partial v}{\partial x} n_x + \frac{1}{\epsilon^2}\frac{\partial v}{\partial z} n_z,$$

corresponding to each of the two subdomains ω^+ and ω^-. Then, for $(\tilde{x},\tilde{z}) \in \omega^+$ we have

$$\begin{aligned} v_\epsilon(\tilde{x},\tilde{z}) &= \int_{\omega^+} Lv_\epsilon\ N_\epsilon(\cdot;(\tilde{x},\tilde{z}))\ dxdz - \int_{\omega^+} v_\epsilon\ LN_\epsilon(\cdot;(\tilde{x},\tilde{z}))\ dxdz \\ &= \int_{\partial\omega^+} Dv_\epsilon N_\epsilon(\cdot;(\tilde{x},\tilde{z}))\ ds - \int_{\partial\omega^+} v_\epsilon\ DN_\epsilon(\cdot;(\tilde{x},\tilde{z}))\ ds \end{aligned} \tag{46}$$

and

$$0 \ = \ \int_{\partial\omega^-} Dv_\epsilon\ N_\epsilon(\cdot;(\tilde{x},\tilde{z}))\ ds - \int_{\partial\omega^-} v_\epsilon\ DN_\epsilon(\cdot;(\tilde{x},\tilde{z}))\ ds. \tag{47}$$

Multiplying (46) by a_+ and (47) by a_- and adding we obtain

$$\begin{aligned} a_+ v_\epsilon(\tilde{x},\tilde{z}) + a_+ \int_{\partial\omega^+} v_\epsilon\ DN_\epsilon(\cdot\ ;(\tilde{x},\tilde{z}))\ ds + a_- \int_{\partial\omega^-} v_\epsilon\ DN_\epsilon(\cdot\ ;(\tilde{x},\tilde{z}))\ ds \\ = a_+ \int_{\Gamma_+} Dv_\epsilon\ N_\epsilon(\cdot\ ;(\tilde{x},\tilde{z}))\ ds\ +\ a_- \int_{\Gamma_-} Dv_\epsilon\ N_\epsilon(\cdot\ ;(\tilde{x},\tilde{z}))\ ds \end{aligned} \tag{48}$$

for $(\tilde{x},\tilde{z}) \in \omega^+$. Here Γ_+ and Γ_- are the top and bottom boundaries given by

$$\Gamma_+ = \{(x,1) : 0 \le x \le 1\},$$
$$\Gamma_- = \{(x,-1) : 0 \le x \le 1\}.$$

We have used the conditions

$$\begin{aligned} a_+ \left(Dv_\epsilon\right)^+ &= -a_- \left(Dv_\epsilon\right)^- \quad \text{along } \ \Gamma_I, \\ Dv_\epsilon \ &= 0 \quad \text{along } x = 0 \text{ and } x = 1, \end{aligned}$$

where $(Dv_\epsilon)^+$ and $(Dv_\epsilon)^-$ correspond to the situations in which Γ_I is interpreted as a boundary part of ω^+ and ω^-, respectively (recall that the normals are exterior to the subdomains). Now let $(\tilde{x}, \tilde{z})$ approach $\partial\omega^+ \setminus \Gamma_I$, excluding the top two corner points. (48) then becomes

$$\begin{aligned} a_+ v_\epsilon(\tilde{x},\tilde{z}) + a_+ \int_{\partial\omega^+} v_\epsilon \ DN_\epsilon(\cdot;(\tilde{x},\tilde{z}))ds - \frac{a_+ v_\epsilon(\tilde{x},\tilde{z})}{2} \ + a_- \int_{\partial\omega^-} v_\epsilon \ DN_\epsilon(\cdot;(\tilde{x},\tilde{z}))ds \\ = a_+ \int_{\Gamma_+} N_\epsilon(\cdot;(\tilde{x},\tilde{z})) \ Dv_\epsilon ds \ + \ a_- \int_{\Gamma_-} N_\epsilon(\cdot;(\tilde{x},\tilde{z})) \ Dv_\epsilon ds. \end{aligned}$$

The additional term on the left-hand side reflects the jump associated with the double-layer potential. For $(\tilde{x}, \tilde{z}) \in \partial\omega^+ \setminus \Gamma_I$, excluding the top two corners, we thus have

$$\begin{aligned} & v_\epsilon(\tilde{x},\tilde{z}) + 2\int_{\partial\omega^+} v_\epsilon \ DN_\epsilon(\cdot;(\tilde{x},\tilde{z}))ds + \frac{2a_-}{a_+}\int_{\partial\omega^-} v_\epsilon \ DN_\epsilon(\cdot;(\tilde{x},\tilde{z}))ds \\ & \quad = \frac{2}{a_+}\left(a_+\int_{\Gamma_+} N_\epsilon(\cdot;(\tilde{x},\tilde{z})) \ Dv_\epsilon \ ds \ + \ a_-\int_{\Gamma_-} N_\epsilon(\cdot;(\tilde{x},\tilde{z})) \ Dv_\epsilon ds\right). \end{aligned} \tag{49}$$

Similarly, for $(\tilde{x}, \tilde{z}) \in \omega^- \setminus \Gamma_I$, excluding the bottom two corners, we have

$$\begin{aligned} & v_\epsilon(\tilde{x},\tilde{z}) + \frac{2a_+}{a_-}\int_{\partial\omega^+} v_\epsilon \ DN_\epsilon(\cdot;(\tilde{x},\tilde{z})) \ ds + 2\int_{\partial\omega^-} v_\epsilon \ DN_\epsilon(\cdot;(\tilde{x},\tilde{z})) \ ds \\ & \quad = \frac{2}{a_-}\left(a_+\int_{\Gamma_+} N_\epsilon(\cdot;(\tilde{x},\tilde{z})) \ Dv_\epsilon \ ds \ + \ a_-\int_{\Gamma_-} N_\epsilon(\cdot;(\tilde{x},\tilde{z})) \ Dv_\epsilon \ ds\right). \end{aligned} \tag{50}$$

Next, consider the interior (smooth) points of Γ_I. For such points we arrive at

$$\begin{aligned} & v_\epsilon(\tilde{x},\tilde{z}) + \frac{2a_+}{a_+ + a_-}\int_{\partial\omega^+} v_\epsilon \ DN_\epsilon(\cdot\ ;(\tilde{x},\tilde{z})) \ ds + \frac{2a_-}{a_+ + a_-}\int_{\partial\omega^-} v_\epsilon \ DN_\epsilon(\cdot\ ;(\tilde{x},\tilde{z})) \ ds \\ & \quad = \frac{2}{a_+ + a_-}\left(a_+\int_{\Gamma_+} N_\epsilon(\cdot\ ;(\tilde{x},\tilde{z})) \ Dv_\epsilon \ ds + a_-\int_{\Gamma_-} N_\epsilon(\cdot\ ;(\tilde{x},\tilde{z})) \ Dv_\epsilon \ ds\right). \end{aligned} \tag{51}$$

This equation can be obtained by letting $(\tilde{x}, \tilde{z})$ approach Γ_I from inside ω^+ in (48). It can also be obtained by letting $(\tilde{x}, \tilde{z})$ approach Γ_I from inside ω^- in the equation analogous to (48) for the subdomain ω^-.

We note that at a corner the jump associated with a double-layer potential must be modified. The jump now becomes

$$\frac{1}{2}\delta_{out} \ v_\epsilon \quad \text{and} \quad -\frac{1}{2}\delta_{in} \ v_\epsilon$$

as $(\tilde{x}, \tilde{z})$ approaches the corner, (x_c, z_c), from outside and from inside the domain, respectively. Here

$$\delta_{out} = \frac{\gamma}{\pi} \quad \text{and} \quad \delta_{in} = 2 - \frac{\gamma}{\pi},$$

and γ is the interior angle of the corner. Therefore, letting $(\tilde{x}, \tilde{z}) \to (x_c, z_c)$ in (48) and simplifying we obtain

$$\begin{aligned} v_\epsilon(x_c, z_c) + 4 \int_{\partial\omega^+} v_\epsilon \; DN_\epsilon(\cdot\,; (x_c, z_c)) \; ds + \frac{4a_-}{a_+} \int_{\partial\omega^-} v_\epsilon \; DN_\epsilon(\cdot\,; (x_c, z_c)) \; ds \\ = \frac{4}{a_+} \left(a_+ \int_{\Gamma_+} N_\epsilon(\cdot\,; (x_c, z_c)) \; Dv_\epsilon \; ds + a_- \int_{\Gamma_-} N_\epsilon(\cdot\,; (x_c, z_c)) \; Dv_\epsilon \; ds \right), \end{aligned} \tag{52}$$

where (x_c, z_c) is one of the top two corners of ω^+. Similarly, we can establish the following formula for the bottom two corners of ω^-:

$$\begin{aligned} v_\epsilon(x_c, z_c) + \frac{4a_+}{a_-} \int_{\partial\omega^+} v_\epsilon \; DN_\epsilon(\cdot\,; (x_c, z_c)) \; ds + 4 \int_{\partial\omega^-} v_\epsilon \; DN_\epsilon(\cdot\,; (x_c, z_c)) \; ds \\ = \frac{4}{a_-} \left(a_+ \int_{\Gamma_+} N_\epsilon(\cdot\,; (x_c, z_c)) \; Dv_\epsilon \; ds \; + \; a_- \int_{\Gamma_-} N_\epsilon(\cdot\,; (x_c, z_c)) \; Dv_\epsilon \; ds \right). \end{aligned} \tag{53}$$

We now consider the two endpoints of Γ_I. Letting $(\tilde{x}, \tilde{z}) \to (x_c, z_c)$ in (48) and applying the jump condition above we obtain

$$\begin{aligned} v_\epsilon(x_c, z_c) + \frac{2\pi a_+}{\gamma a_+ + (\pi - \gamma) a_-} \int_{\partial\omega^+} v_\epsilon \; DN_\epsilon(\cdot\,; (x_c, z_c)) \; ds \\ + \frac{2\pi a_-}{\gamma a_+ + (\pi - \gamma) a_-} \int_{\partial\omega^-} v_\epsilon \; DN_\epsilon(\cdot\,; (x_c, z_c)) \; ds \\ = \frac{2\pi}{\gamma a_+ + (\pi - \gamma) a_-} \left(a_+ \int_{\Gamma_+} N_\epsilon(\cdot\,; (x_c, z_c)) \; Dv_\epsilon \; ds \right. \\ \left. + a_- \int_{\Gamma_-} N_\epsilon(\cdot\,; (x_c, z_c)) \; Dv_\epsilon \; ds \right), \end{aligned} \tag{54}$$

where (x_c, z_c) is now one of the endpoints of Γ_I (and γ is the angle interior to ω^+). While we have initially for (most of) our theoretical analysis assumed that h is globally smooth we use a piecewise-linear representation for h in our numerical computations. Since we deliberately avoid placing any collocation points at interior corners of Γ_I it is, however, not necessary to derive the special form of the integral formulation corresponding to such interior corner points. Summarizing (49) through (54) we have

$$\begin{aligned} v_\epsilon(\tilde{x}, \tilde{z}) + k^+(\tilde{x}, \tilde{z}) \int_{\partial\omega^+} v_\epsilon \; DN_\epsilon(\cdot\,; (\tilde{x}, \tilde{z})) \; ds \\ + k^-(\tilde{x}, \tilde{z}) \int_{\partial\omega^-} v_\epsilon \; DN_\epsilon(\cdot\,; (\tilde{x}, \tilde{z})) \; ds \\ = \frac{k^+(\tilde{x}, \tilde{z})}{a_+} \left(a_+ \int_{\Gamma_+} N_\epsilon(\cdot\,; (\tilde{x}, \tilde{z})) \; Dv_\epsilon \; ds \right. \\ \left. + a_- \int_{\Gamma_-} N_\epsilon(\cdot\,; (\tilde{x}, \tilde{z})) \; Dv_\epsilon \; ds \right), \end{aligned}$$

where $(\tilde{x}, \tilde{z})$ denotes any point on the boundary. The coefficients $k^+(\tilde{x}, \tilde{z})$ and $k^-(\tilde{x}, \tilde{z})$ are related by $k^+(\tilde{x}, \tilde{z}) = \frac{a_+}{a_-} k^-(\tilde{x}, \tilde{z})$, and their exact dependence on $(\tilde{x}, \tilde{z})$ is given by (49)–(54). The previous equation may be rewritten

$$v_\epsilon(\tilde{x}, \tilde{z}) + \int_{\partial\Omega \cup \Gamma_I} K(\cdot\,; (\tilde{x}, \tilde{z})) \left(n_x \frac{\partial}{\partial x} + n_z \frac{1}{\epsilon^2} \frac{\partial}{\partial z} \right) N_\epsilon(\cdot\,; (\tilde{x}, \tilde{z})) \; v_\epsilon \; ds \tag{55}$$

$$= \frac{k^+(\tilde{x},\tilde{z})}{\epsilon a_+}\left(\int_{\Gamma_+} f\ N_\epsilon(\cdot\ ; (\tilde{x},\tilde{z}))\ ds - \int_{\Gamma_-} g\ N_\epsilon(\cdot\ ; (\tilde{x},\tilde{z}))\ ds\right),$$

with

$$K((x,z);(\tilde{x},\tilde{z})) = \begin{cases} k^+(\tilde{x},\tilde{z}), & (x,z) \in \partial\omega^+ \setminus \Gamma_I, \\ k^-(\tilde{x},\tilde{z}), & (x,z) \in \partial\omega^- \setminus \Gamma_I, \\ k^+(\tilde{x},\tilde{z}) - k^-(\tilde{x},\tilde{z}), & (x,z) \in \Gamma_I. \end{cases}$$

In the above formula n still denotes the exterior unit normal on $\partial\Omega$. However, along the interface Γ_I, n is now taken to be the exterior unit normal to ω^+. The change in the right-hand side has been obtained by insertion of the formulas

$$a_+ Dv_\epsilon|_{\Gamma_+} = \frac{f}{\epsilon}, \quad a_- Dv_\epsilon|_{\Gamma_-} = -\frac{g}{\epsilon}.$$

The solution v_ϵ (being the solution to a Neumann problem) is only determined up to an additive constant. To remedy this situation we impose the additional condition

$$\int_{\partial\Omega} v_\epsilon\ ds = 0. \tag{56}$$

The equations (55) and (56) are now uniquely solvable.

We shall now briefly describe the collocation method which we use to discretize the system of equations (55)–(56); for more details including an analysis of its convergence properties, see Chapter 13 in [11]. Let $s(t) = (x(t), z(t))\ :\ [0,T] \to \partial\Omega \cup \Gamma_I$ be a parametrization by arclength of the cell boundary and the interface. In our implementation we take $s(0)$ to be the point where Γ_I meets the left vertical boundary, $s(t)$, $t > 0$, then traces the boundary $\partial\Omega$ clockwise until it returns to $s(0)$, at which point it continues along Γ_I finally reaching the point where Γ_I meets the right vertical boundary at $t = T$. We select $p = 7N - 1$ collocation points along $\partial\Omega \cup \Gamma_I$. It is always required that the four corners of $\partial\Omega$ and the endpoints of Γ_I be collocation points. We select $N + 1$ equidistant points along the top side of Ω as well as along the bottom side. $2N + 1$ points are chosen on each of the vertical sides, such as to be approximately equidistant while satisfying the restriction that the endpoints of Γ_I be amongst them. Finally we select $N + 1$ points along the interface Γ_I. These points are again evenly spaced with the following exception. Given an equidistant set of points

$$(x(t_1), y(t_1)), \ldots, (x(t_{N+1}), y(t_{N+1}))$$

along Γ_I we minimally shift those points (if any) that fall at interior corners of Γ_I. This is done to avoid explicitly incorporating the special integral formulation corresponding to such corners. The total number of collocation points is $7N-1$ (as each of the corners of Ω and the endpoints of Γ_I are included twice above). Given the collocation points

$$s(t_j) = (x(t_j), y(t_j)), \qquad 0 < t_1 < t_2 \cdots < t_p < T,$$

we seek an approximate solution $v_\epsilon^{(p)}$ from a finite-dimensional subspace X_p, with $\dim X_p = p = 7N - 1$, by requiring that (55) be satisfied at these points. For our computations, X_p consists of the continuous piecewise-linear functions with nodes at the given collocation points. For convenience we shall sometimes think of $v_\epsilon^{(p)}$ as a function of t through the following natural identification:

$$v_\epsilon^{(p)}(t) = v_\epsilon^{(p)}(s(t)).$$

Let $c = \{c_j\}_{j=1}^{p}$ denote the nodal values of $v_\epsilon^{(p)}$. Then for $t \in [t_j, t_{j+1}]$,

$$v_\epsilon^{(p)}(t) = \frac{c_j(t_{j+1} - t) + c_{j+1}(t - t_j)}{t_{j+1} - t_j}.$$

$v_\epsilon^{(p)}$ also has the representation

$$v_\epsilon^{(p)}(t) = \sum_{j=1}^{p} c_j L_j(t), \quad t_1 < t < t_p, \tag{57}$$

where L_j is the standard piecewise-linear "hat" function with support in $[t_{j-1}, t_{j+1}]$. Upon substitution of (57) into (55) the coefficients $c = \{c_j\}_{j=1}^{p}$ solve the linear system

$$Mc = \varphi, \tag{58}$$

with the "generic" elements of the matrix M defined by

$$m_{ij} = \begin{cases} L_j(t_i) + \int_{t_{j-1}}^{tj+1} K(s(t); s(t_i)) \; L_j(t) \; DN_\epsilon(s(t); s(t_i)) \; dt, & 1 \le i, j \le p, \\ \int_{\partial\Omega} L_j(t) dt, & i = p+1, \;\; 1 \le j \le p, \end{cases}$$

and the right-hand side given by

$$\varphi_i = \frac{k^+(s(t_i))}{\epsilon \; a_+} \left(\int_{\Gamma_+} f \; N_\epsilon(\cdot \; ; s(t_i)) \; ds \; - \; \int_{\Gamma_-} g \; N_\epsilon(\cdot \; ; s(t_i)) \; ds \right).$$

When $0 < t_j < T$ is one of the two parameter values corresponding to the endpoints of Γ_I an extra integral has to be added to the formula for m_{ij}. This integral corresponds to the parameter interval $(0, t_1)$ in the case of the left endpoint and the parameter interval (t_p, T) in the case of the right endpoint. Almost all the (nonzero) integrals in the definition of the matrix m are calculated by means of a Simpson's composite rule. An exception is made when $j = i$ and $s(t_i)$ is a collocation point adjacent to one of the (interior) corners of Γ_I; in this case the integral is calculated using a highly accurate adaptive quadrature routine. We solve the $(p+1) \times p$ system of equations (58) by means of linear least squares. Once we obtain the coefficients c_j, we have an approximation to v_ϵ along $\partial\Omega \cup \Gamma_I$. In order to reconstruct the interface function $h(x)$, we most often also require an approximation to $\epsilon \frac{\partial v_\epsilon}{\partial x}$ along Γ_+ and Γ_-. Having computed the values

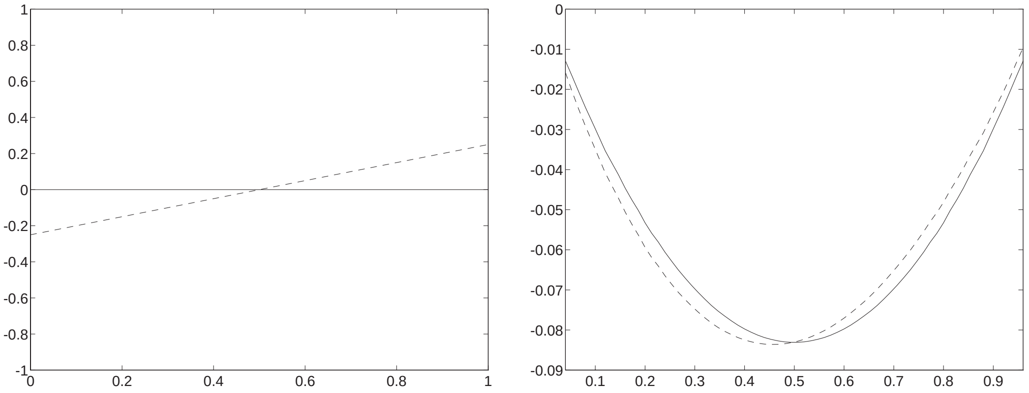
$$v_\epsilon^{(p)}(0,1), \; v_\epsilon^{(p)}\left(\frac{1}{N}, 1\right), \; v_\epsilon^{(p)}\left(\frac{2}{N}, 1\right), \ldots, v_\epsilon^{(p)}(1,1)$$

and

$$v_\epsilon^{(p)}(0,-1), \; v_\epsilon^{(p)}\left(\frac{1}{N}, -1\right), \; v_\epsilon^{(p)}\left(\frac{2}{N}, -1\right), \ldots, v_\epsilon^{(p)}(1,-1)$$

at the $N+1$ equidistant collocation points along Γ_+ and Γ_-, respectively, we approximate $\epsilon \frac{\partial v_\epsilon}{\partial x}$ by numerical differentiation, using the five-point rule

$$\begin{aligned} &\epsilon \frac{\partial v_\epsilon}{\partial x}\left(\frac{i}{N}, \pm 1\right) \\ &\quad \approx \frac{\epsilon}{12} \frac{v_\epsilon^{(p)}(\frac{i-2}{N}, \pm 1) - 8 v_\epsilon^{(p)}(\frac{i-1}{N}, \pm 1) + 8 v_\epsilon^{(p)}(\frac{i+1}{N}, \pm 1) - v_\epsilon^{(p)}(\frac{i+2}{N}, \pm 1)}{\frac{1}{N}}, \end{aligned}$$



(a) (b)

FIG. 2. *Two interfaces* (a) *and the corresponding boundary voltage derivatives* (b).

$i = 2,\dots, N$-2. Figure 2b displays two examples of this approximate, rescaled derivative from the bottom boundary Γ_-. The computations were both done with $\epsilon = 0.01$, $a_+ = 1$, $a_- = 5$, and with boundary currents $f = -2x + 1$, $g = 2x - 1$. Two different h's were used, namely $h(x) = 0$ and $h(x) = \frac{1}{4}x - \frac{1}{8}$, as displayed in Figure 2a. The solid (dashed) interface in Figure 2a corresponds to the solid (dashed) approximate derivative graphed in Figure 2b. Notice that the approximate derivatives are reasonably different except near $x = 0.5$, where the two h's agree.

In the following section approximate values such as these will be used as "synthetic" data in order to establish the effectiveness of the interface reconstruction formulas derived in section 3.

6. Computational results. Before we present our computational results let us first note any restrictions or conditions imposed on our data. In all our computations the boundary currents $f(x)$ and $g(x)$ are piecewise-linear, but not necessarily continuous, functions which satisfy

$$\int_{\Gamma^+} f(x)dx - \int_{\Gamma^-} g(x)dx = 0. \tag{59}$$

The "true" interface function $h(x)$ is a piecewise-linear continuous function, and the conductivities a_+ and a_- are positive constants with $a_+ \neq a_-$. In the examples presented in this section the "synthetic" data have been generated with N between 50 and 150, corresponding to a total number of collocation points between 349 and 1049. A relatively large number of collocation points was used for small ϵ, large conductivity ratios, and very "rough" h's. The reconstructed interface function is given in terms of its computed values at the equidistant points i/N, $2 \leq i \leq N - 2$. Since the formulas used in reconstructing the interface are only valid for "thin domains," we restrict our thickness parameter, ϵ, to the interval between 0 and 1/10. In two of our examples (corresponding to Figures 3b and 8) we select parameters that are directly relevant to Hall–Héroult cells. In our other examples the parameters are chosen simply to test various features of the reconstruction formulas.

We note that in many cases the voltage potential v_ϵ has a singularity at the corners of the domain, Ω, as well as at the points where the interface meets the vertical boundaries. Therefore the approximation we calculate may not be very accurate near these (six) points. Furthermore, the value of $\int_0^x (f-g)(t)dt$ is always very close to zero when x is near 0 or 1, and so the reconstruction formula based on the derivative data degenerates near these points. Recognizing these facts, we cannot expect to recover the interface function $h(x)$ very well for values near 0 or 1. Consequently, we have graphed all the reconstructed interfaces only for $0.05 < x < 0.95$.

In the first example we examine the effect of different conductivity ratios on the reconstructed interface. We use the derivative-based formula (41). Here and in all the following examples the cutoff parameter δ (in the formula for $DV_{\pm\epsilon}$) is chosen very small (so small that δa_- is always of magnitude less than 10^{-5}). We compare the reconstruction results for the following two sets of conductivities: (a) $a_+ = 1,\ a_- = 2$, and (b) $a_+ = 1,\ a_- = 10^4$, with all other data remaining the same. The boundary currents are $f(x) = -x + \frac{1}{2}$ and $g(x) = \ x - \frac{1}{2}$. The interface function, $h(x)$, is defined by

$$h(x) = \begin{cases} 0.50, & 0 \quad \le x \le 0.30, \\ -100x + 30.5, & 0.30 \quad < x < 0.31, \\ -0.50, & 0.31 \quad \le x \le 1, \end{cases}$$

and the thickness parameter used is $\epsilon = 0.02$. Each of the frames in Figure 3 shows two reconstructions of $h(x)$ obtained by using the bottom derivative data (the dot-dashed line) and the top derivative data (the dashed line). As in all the figures to follow the solid line represents the "true" location of the interface. Figure 3a corresponds to the first set of conductivities ($a_+ = 1,\ a_- = 2$) and Figure 3b corresponds to the second set. It is evident that if the conductivities are close in size, either set of data (top or bottom) suffices for the reconstruction. However, if the conductivities are very different in magnitude, as in case (b), the bottom data (from the boundary adjacent to the subdomain with the large conductivity) produces the only acceptable reconstruction of $h(x)$. This is consistent with our remarks at the end of section 4. In the rest of our examples we always use the bottom data to calculate the boundary voltage derivative, unless otherwise stated. We note that there are oscillations present in the (synthetic) boundary derivative data (particularly in that which corresponds to the top boundary and a conductivity ratio $a_+/a_- = 10^{-4}$). These oscillations could be made smaller by use of more collocation points and/or a more sophisticated differentiation strategy. This last comment is only of minor practical interest since "real" measured data will come with a limited accuracy, which we cannot as easily change.

In our next example we analyze the reconstruction results obtained by means of (41) for different ϵ. In each case the boundary currents are given by $f(x) = \frac{x}{2}$ and $g(x) = \frac{x}{4} + \frac{1}{8}$; the interface, $h(x)$, is defined by

$$h(x) = \begin{cases} 0, & 0 \quad \le x < 0.25, \\ 2x - 0.5, & 0.25 \quad \le x < 0.375, \\ -2x + 1, & 0.375 \quad \le x < 0.625, \\ 2x - 1.5, & 0.625 \quad \le x < 0.75, \\ 0, & 0.75 \quad \le x \le 1, \end{cases}$$

and the conductivities are $a_+ = 1$ and $a_- = 5$. We shall compare the results obtained for the following three cases: (a) $\epsilon = 0.10$, (b) $\epsilon = 0.05$, and (c) $\epsilon = 0.01$. Since the performance of the reconstruction formula is based on how well $\epsilon\frac{\partial}{\partial x}u_\epsilon(x, \pm 1)$ approximates $\frac{d}{dx}v^{(-1)}(x)$, $0 \le x \le 1$, we can expect better results for smaller ϵ. On

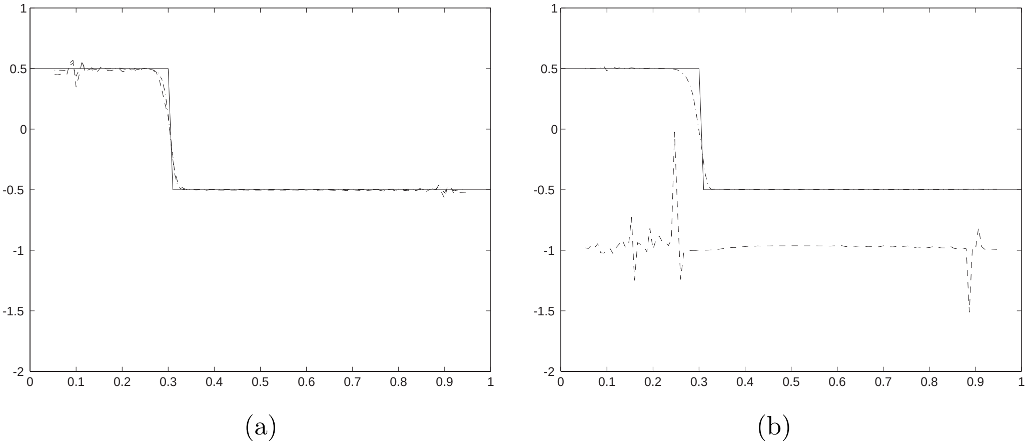


FIG. 3. *Reconstructed interfaces for two different conductivity ratios.* (a) $a_+/a_- = 0.5$, (b) $a_+/a_- = 10^{-4}$.

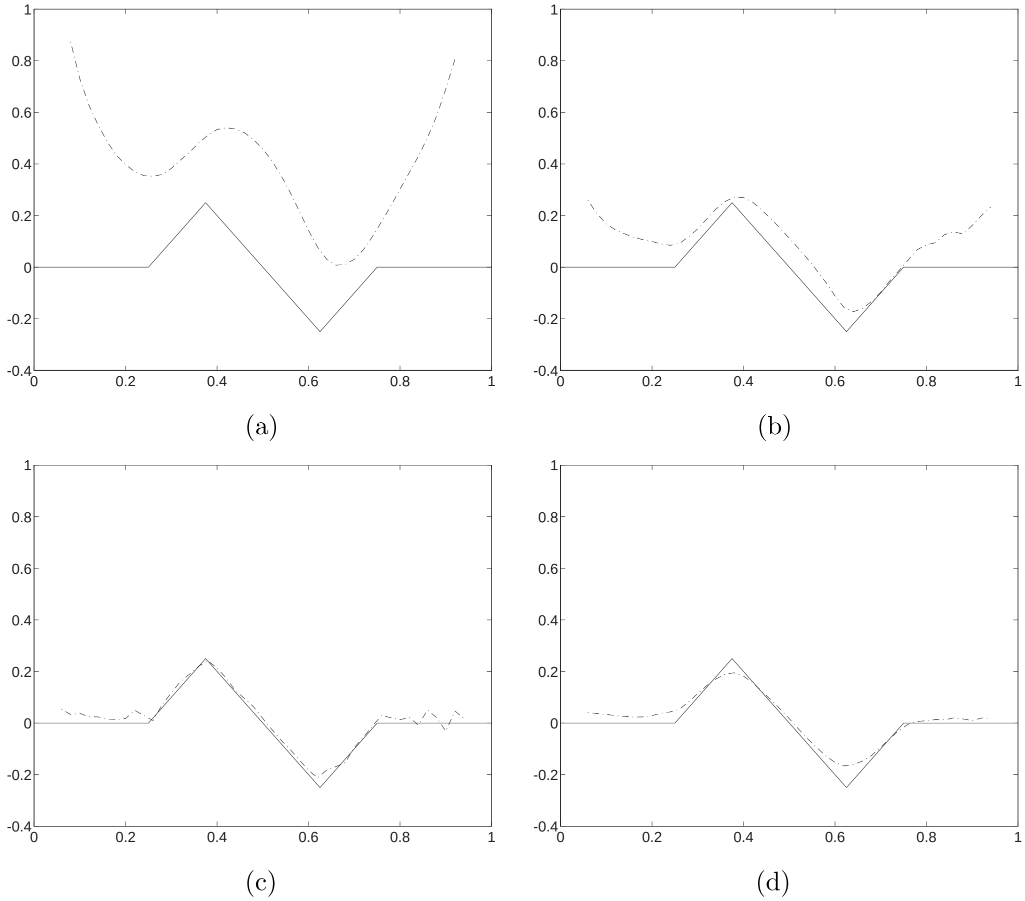
the other hand we note that as ϵ decreases the solution to (1) becomes increasingly more difficult to approximate; i.e., the quality of our synthetic data declines. Our computational results displayed in Figure 4a–c confirm both of these observations. Figure 4a shows the reconstruction of $h(x)$ for $\epsilon = 0.10$. The smoothness of the reconstruction directly corresponds to the smoothness of our approximation $u_\epsilon^{(p)}$. In this case the approximation is very smooth, but the thickness parameter ϵ is not sufficiently small, resulting in a smooth, but fairly inaccurate, reconstruction. Figure 4b shows the reconstruction with $\epsilon = 0.05$. The improved accuracy and increased irregularity is evident. Figure 4c shows the reconstruction of $h(x)$ for $\epsilon = 0.01$. The effect of decreasing ϵ is clearly displayed. The reconstruction is highly accurate but is also somewhat irregular. We may regularize the reconstructed curve in the following fashion. Let $h_r(x)$ be the originally reconstructed interface function and consider the energy

$$\begin{aligned} E_c(\tilde{h}_r) &= \|\tilde{h}_r - h_r\|^2_{L^2(0,1)} + c\|\tilde{h}'_r\|^2_{L^2(0,1)} \\ &= \quad E_1 \quad + \quad cE_2. \end{aligned}$$

The regularized interface function is an approximate minimizer of this energy. We obtain this approximation by solving a finite difference version of the corresponding Euler–Lagrange equation

$$-c\frac{d^2\tilde{h}_r}{dx^2} + \tilde{h}_r = h_r,$$

with the boundary conditions $\tilde{h}'_r = 0$ at $x = 0, 1$. Our selection of the value of c is based on an L-curve approach. Quite frequently, as c increases up to a particular value, c_{opt}, E_2 rapidly decreases while E_1 slightly increases. Then, for $c > c_{opt}$, E_2 decreases at a declining rate while E_1 rapidly increases. In some sense c_{opt} represents the optimal choice. For our example we examined the results obtained for several values of c. The reconstruction obtained after regularization of the interface in Figure



(a)

(b)

(c)

(d)

FIG. 4. *Reconstructions for various values of* ϵ. (a) $\epsilon = 0.1$, (b) $\epsilon = 0.05$, (c) $\epsilon = 0.01$, (d) $\epsilon = 0.01$, *regularized.*

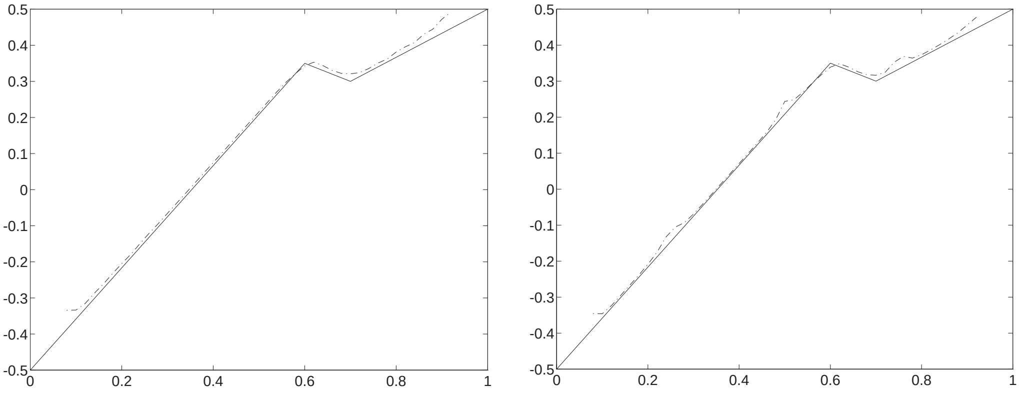
4c is graphed in Figure 4d. For a further discussion of this regularization approach and the selection of c, we refer the reader to [6] and [10].

In our next example we examine the effect of the smoothness of the boundary currents on the reconstructed interface. We again base the reconstruction on (41). For our example we take conductivity values $a_+ = 1$ and $a_- = 5$ and thickness parameter $\epsilon = 0.02$. We compare the computational results for the reconstruction of the interface:

$$h(x) = \begin{cases} \frac{17}{12}x - \frac{1}{2}, & 0 \quad\ \le x < 0.60, \\ -\frac{1}{2}x + \frac{13}{20}, & 0.60 \quad \le x < 0.70, \\ \frac{2}{3}x - \frac{1}{6}, & 0.70 \quad \le x \le 1.00 \end{cases}$$

for the following two sets of boundary currents:

$$\text{(a)} \quad f(x) = x - \frac{1}{2},$$
$$g(x) = 0,$$



(a) (b)

FIG. 5. *The effect of discontinuities in the currents on the reconstruction.*

and

$$\text{(b)} \quad f(x) = \begin{cases} -\frac{3}{8}, & 0 \quad \leq x < 0.25, \\ -\frac{1}{8}, & 0.25 \quad \leq x < 0.50, \\ \frac{1}{8}, & 0.50 \quad \leq x < 0.75, \\ \frac{3}{8}, & 0.75 \quad \leq x \leq 1.00, \end{cases}$$

$$g(x) = 0.$$

The current distribution f in (b) is a very natural piecewise-constant approximation to the continuous distribution in (a). Figure 5a shows the reconstruction corresponding to the smooth f. Away from the endpoints ($x = 0, 1$), the reconstructed interface is uniformly accurate in its approximation of $h(x)$. Figure 5b shows the reconstruction corresponding to the piecewise-constant f. Away from the points of discontinuity of f as well as the endpoints, the reconstructed interface provides an excellent approximation to $h(x)$. This reconstruction is clearly less accurate near the points of discontinuity of f. The local inaccuracy of the reconstruction at these points of discontinuity is caused by two facts: 1) the fact that the boundary voltage derivative $\epsilon\frac{\partial}{\partial x}u_\epsilon$ is a poorer approximation to $\frac{d}{dx}v^{(-1)}$, and 2) the fact that our synthetic data (the approximation of u_ϵ) are less accurate.

In our next example we use model data corresponding to a cell in which both (41) and (43) must be employed to reconstruct $h(x)$. The conductivities are $a_+ = 1$ and $a_- = 2$ and once again we use a thickness parameter of $\epsilon = 0.02$. The "true" interface is given by

$$(60) \qquad h(x) = \begin{cases} -\frac{3}{4}x + \frac{2}{5}, & 0 \quad \leq x < 0.20, \\ x + \frac{1}{20}, & 0.20 \quad \leq x < 0.25, \\ -\frac{3}{4}x + \frac{39}{80}, & 0.25 \quad \leq x < 0.65, \\ x - \frac{13}{20}, & 0.65 \quad \leq x < 0.70, \\ -\frac{3}{4}x + \frac{23}{40}, & 0.70 \quad \leq x \leq 1.00, \end{cases}$$

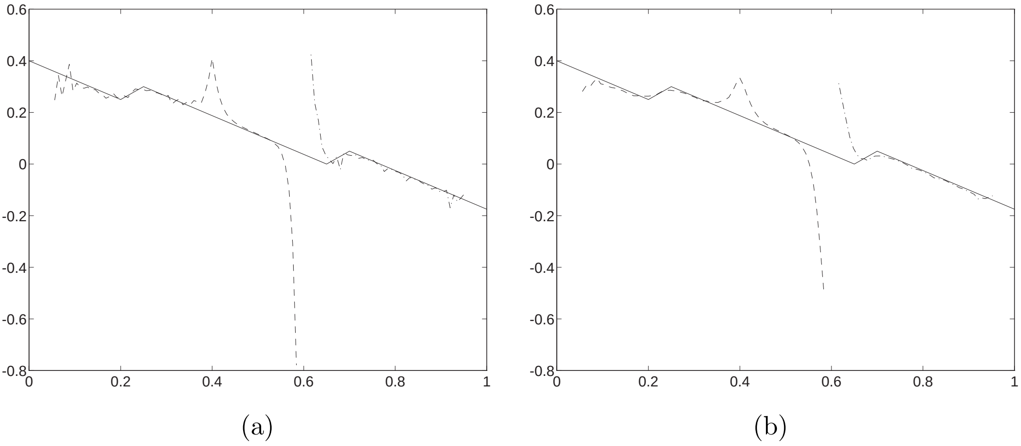


FIG. 6. *Reconstruction using a combination of* (41) *and* (43).

and the boundary currents are defined by

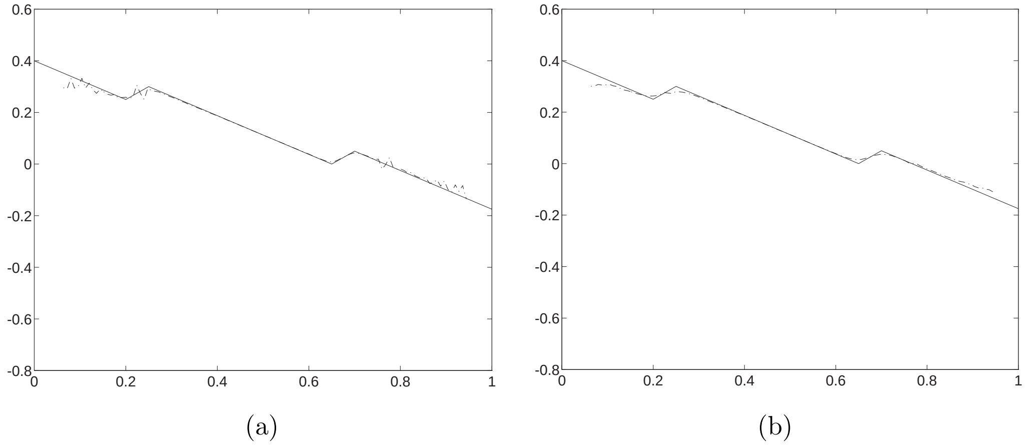
$$f(x) = \begin{cases} 2, & 0 \leq x < 0.40, \\ -1, & 0.40 \leq x < 0.60, \\ 2, & 0.60 \leq x \leq 1.00, \end{cases}$$

$$g(x) = \begin{cases} 1, & 0 \leq x < 0.60, \\ \\ 2, & 0.60 \leq x \leq 1.00. \end{cases}$$

Since $\int_0^x (f(t) - g(t))\, dt \neq 0, \quad 0 < x < 0.60$, (41) is used to reconstruct $h(x)$ for $x < 0.60$. The result is the dashed curve to the left in Figure 6a. Since $\int_0^x (f(t) - g(t))\, dt \equiv 0, \quad 0.60 < x < 1.00$, (43) is used to reconstruct $h(x)$ for $x > 0.60$. The result is the dot-dashed curve to the right in Figure 6a. Away from the endpoints the reconstruction is a quite accurate approximation to $h(x)$ except near the points $x = 0.40$ and $x = 0.60$. Note that $x = 0.40$ is a point of discontinuity for $f(x)$ while $x = 0.60$ is a point of discontinuity for both $f(x)$ and $g(x)$ as well as the point at which the formula used to reconstruct $h(x)$ changes. Even with perfect boundary voltage data we could not expect to reconstruct $h(x)$ with extremely good accuracy near $x = 0.40$ and $x = 0.60$. Since the reconstruction is rather unsmooth we regularize the curves using the process described above. The regularized result is graphed in Figure 6b.

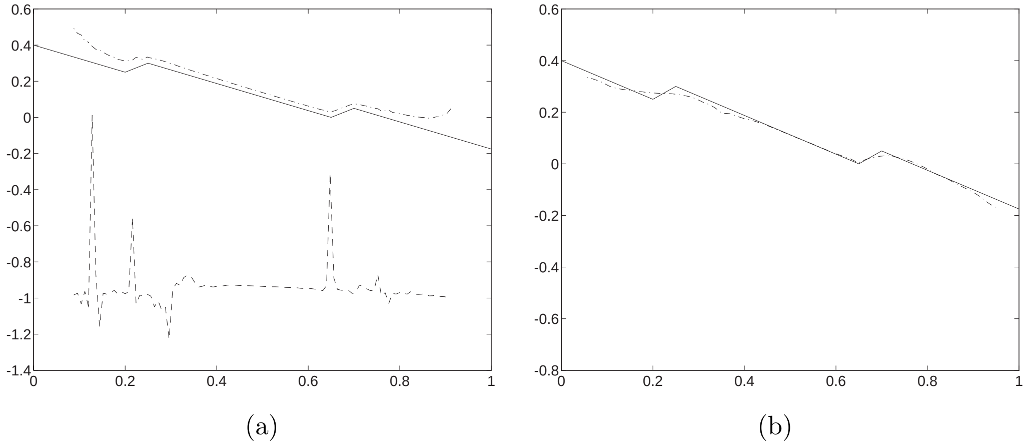
It is not that the interface (60) is particularly difficult to reconstruct. It may be quite accurately reconstructed using smoother (and less degenerate) boundary currents. Figure 7 shows the reconstruction of this interface for conductivities $a_+ = 1$, $a_- = 2$ and thickness parameter $\epsilon = 0.02$ (as before). However, this time we use the boundary currents $f(x) = -2x + 1$ and $g(x) = 2x - 1$. Figure 7a is the reconstruction obtained by use of the derivative-based formula (41); Figure 7b is the same reconstruction, only this time regularized. This example confirms that the inaccuracies in the previous reconstruction were largely a result of the discontinuities and the degeneracy of the applied boundary currents.

In our final example we demonstrate that an interface such as (60) may also be quite accurately reconstructed in the case $a_+ = 1$, $a_- = 10^4$, and $\epsilon = 0.02$. Figure



(a) (b)

FIG. 7. *The same interface and conductivities as in the previous figure, but with smoother, less degenerate boundary currents.*



(a) (b)

FIG. 8. *Same interface as in Figure* 7*, but a conductivity ratio* $a_+/a_- = 10^{-4}$*. Reconstructions based on* (41) *to the left, reconstruction based on* (43) *to the right. The boundary currents for* (a) *and* (b) *differ.*

8a corresponds to applied boundary currents $f(x) = -2x + 1$ and $g(x) = 2x - 1$ and use of (41). The dot-dashed (fairly good) reconstruction corresponds to the use of the bottom derivative data. The dashed (useless) reconstruction is obtained by use of the top derivative data. No regularization was performed on the reconstructions shown in Figure 8a. Figure 8b corresponds to the applied boundary currents $f = g = 1$ and use of (43). This reconstruction has been regularized using the technique described earlier.

REFERENCES

[1] G. ALESSANDRINI AND A. DIAZ VALENZUELA, *Unique determination of multiple cracks by two measurements*, SIAM J. Control Optim., 34 (1996), pp. 913–921.

[2] S. ANDRIEUX, A. BEN ABDA, AND M. JAOUA, *Identifiabilité de frontière inaccessible par des*

mesures de surface, C. R. Acad. Sci. Paris Sér. I Math., 316 (1993), pp. 429–434.

[3] V. Bojarevics and M.V. Romerio, *Long waves instability of liquid metal-electrolyte interface in aluminium electrolytic cells: a generalization of Sele's criterion*, European J. Mech. B Fluids, 13 (1994), pp. 33–56.

[4] K. Bryan, V. Liepa, and M. Vogelius, *Reconstruction of multiple cracks from experimental, electrostatic boundary data*, in Inverse Problems and Optimal Design in Industry, H.W. Engl and J. McLaughlin, eds., Teubner, Stuttgart, 1994, pp. 147–167.

[5] P. G. Ciarlet and P. Destuynder, *A justification of the two-dimensional linear plate model*, J. Mécanique, 18 (1979), pp. 315–344.

[6] D. J. Cedio-Fengya, Ph.D. thesis, Rutgers University, New Brunswick, NJ, 1997.

[7] A. Friedman and V. Isakov, *On the uniqueness in the inverse conductivity problem with one measurement*, Indiana Univ. Math. J., 38 (1989), pp. 563–579.

[8] A. Friedman and M. Vogelius, *Determining cracks by boundary measurements*, Indiana Univ. Math. J., 38 (1989), pp. 527–556.

[9] P. Kaup and F. Santosa, *Nondestructive evaluation of corrosion damage using electrostatic boundary measurements*, J. Nondestructive Evaluation, 14 (1995), pp. 127–136.

[10] P. Kaup, F. Santosa, and M. Vogelius, *A method for imaging corrosion damage in thin plates from electrostatic data*, Inverse Problems, 12 (1996), pp. 279–293.

[11] R. Kress, *Linear Integral Equations*, Springer-Verlag, New York, 1989.

[12] D. Morgenstern and I. Szabo, *Vorlesungen über Theoretische Mechanik*, Springer-Verlag, New York, 1961.

[13] C. Schwab, *Boundary layer resolution in hierarchical models of laminated composites*, RAIRO Modél. Math. Anal. Numér., 28 (1994), pp. 517–537.

[14] M. Vogelius and I. Babuska, *On a dimensional reduction method. The optimal selection of basis functions*, Math. Comp., 37 (1981), pp. 31–46.

[15] M. Vogelius, *Crack detection by electrical impedance tomography. A survey of results*, to appear.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 1, pp. 278–302

CONVEX OPTIMIZATION VIA FEEDBACKS[*]

ARKADII V. KRYAZHIMSKII[†]

Abstract. Three dynamical systems are associated with a problem of convex optimization in a finite-dimensional space. For system trajectories $x(t)$, the ratios $x(t)/t$ are, respectively, (i) solution tracking (staying within the solution set X^0), (ii) solution abandoning (reaching X^0 as time t goes back to the initial instant), and (iii) solution approaching (approaching X^0 as time t goes to infinity). The systems represent a closed control system with appropriate feedbacks. In typical cases, the structure of the trajectories is simple enough. For instance, for a problem of quadratic programming with linear and box constraints, solution-approaching dynamics are described by a piecewise-linear ODE with a finite number of polyhedral domains of linearity. Finding the order of visiting these domains yields an analytic resolution of the original problem; a detailed analysis is given for a particular example. A discrete-time approach is outlined.

Key words. feedback control, differential inclusions, convex programming, linear equality constraints

AMS subject classifications. 90C25, 93B52, 49M30, 34A60

PII. S036301299528030X

1. Introduction.

Our goal in this paper is to expose new connections between dynamical systems and static convex optimization problems. The idea of approaching a static optimizer along a trajectory of a continuous-time dynamical system has been exploited in various aspects. Gradient processes shifting the primal and dual variables against the gradients of the Lagrange function so as to attain its saddle point have perhaps the longest history [1]. Gradient-type processes in feasible sets [2], continuous-time gradient projection [3], and continuous-time barrier projection [4] illustrate, among other things, further developments in the gradient approach. The homotopy approach rests on the idea of tracking solutions of time-parametrized problems by smoothly transforming an easily solvable problem into the original one. In this context we mention path-following processes [5] and interior-point homotopy methods [6]. Finally, we refer to continuous-time counterparts of proximal and Newton-type optimization methods (see, e.g., [3]).

In this paper we hold a viewpoint of control theory. We describe three dynamical systems (ODEs) whose state-over-time ratios converge to static optimizers. The ODEs represent a closed control system with appropriate feedbacks. The structure of the feedbacks originates from the method of shifting control [7] and its regularized modifications [8].

The paper has two objectives.

The first objective is qualitative. We show that the dynamical systems corresponding to the *Filippov shifting*, *regularized shifting*, and *penalized shifting* feedbacks converge to an optimizer differently. The Filippov shifting system is fixed on the optimizer. The regularized shifting system reaches the optimizer as time goes back to the initial instant. The penalized shifting system finds the optimizer as time goes to

[*] Received by the editors January 23, 1995; accepted for publication (in revised form) February 9, 1998; published electronically November 5, 1998. This work was supported in part by the Russian Foundation for Basic Research under grant 97-01-01060 and the International Institute for Applied Systems Analysis.

http://www.siam.org/journals/sicon/37-1/28030.html

[†] Mathematical Steklov Institute, Gubkina 8, GSP-1, 117966 Moscow, Russia (kryazhim@genesis.mi.ras.ru).

infinity. Thus, we have solution-tracking dynamics, solution-abandoning dynamics, and solution-approaching dynamics. Up to now, dynamical systems reaching the optimizer at infinity (a prototype of solution approaching) have been considered. Solution tracking and solution abandoning extend the variety of qualitative links between the dynamical systems and the solvers to convex optimization problems.

Our second objective is to outline an analytic optimization approach associated with explicit solutions of the designed ODEs. Generally, these ODEs have multivalued right-hand sides, and nonstandard solution techniques are required. However, if the minimized function is strictly convex, the right-hand sides of the ODEs are single valued (the Filippov shifting system is a single exception). For instance, the penalized shifting system associated with a problem of quadratic programming with linear and box constraints is described by a piecewise-linear ODE with a finite number of polyhedral domains of linearity. System trajectory is represented explicitly in each domain. A problem is to identify the order of visiting the domains. In section 7 we give an example of an analytic resolution of this problem.

We restrict this study to convex optimization problems whose feasible sets are described by inclusions and linear equalities. The linear equality constraints allow us to implement the idea of extremal shifting with minimum modifications (in this paper we deal with a minimum of technical details).

In section 2 we discuss the underlying idea of our method. In sections 3, 4, and 6 we introduce the Filippov shifting system, the regularized shifting system, and the penalized shifting system, and prove their convergence properties. Sections 5 and 7 are devoted to the analytic design of the regularized shifting trajectories and the penalized shifting trajectories for two particular examples. In section 8 we consider discrete-time analogues of the shifting and penalized shifting dynamics.

The study was initiated in [9]. In this paper the method of regularized shifting was used for the justification of a finite approximate optimization algorithm. A preliminary text of the present paper was published in [10]. A family of numerical algorithms for convex optimization adjoining the suggested approach is described in [11].

2. Outline of the method: Definitions. We are concerned with the optimization problem

$$\begin{aligned} &\text{minimize } J(x),\\ &x \in M,\\ &Fx = b. \end{aligned} \tag{2.1}$$

Here J is a convex function on R^n, M is a closed, convex, and bounded set in R^n, F is an $r \times n$ matrix, and $b \in R^r$. As usual, R^k is a k-dimensional Euclidean space of column vectors, x_i is the ith coordinate of $x \in R^n$, x^T stands for x transposed, and $|\cdot|$ denotes the Euclidean norm. A point $x \in R^n$ such that $x \in M$, $Fx = b$ is called feasible in problem (2.1). The collection of all points feasible in (2.1) forms the feasible set of problem (2.1). We assume that the feasible set of problem (2.1) is nonempty. J^0 and X^0 denote the optimal value and the solution set in problem (2.1), respectively. Note that X^0 is nonempty. In what follows, $\operatorname{argmin}\{f(x) : x \in E\}$ stands for the collection of all minimizers of function f in E.

Let us outline informally a method of building a dynamical system tracking the solution set X^0. We begin with an ODE with a fixed (zero) initial state and a variable (controllable) right-hand side $u(t)$:

$$\dot{x}(t) = u(t), \quad x(0) = 0, \tag{2.2}$$

where time t varies between zero and infinity. Impose the constraint $u(t) \in M$. Then, no matter how $u(t)$ is formed, $x(t)/t \in M$ for all $t > 0$; this follows from the convexity of M. In other words, the ratio $x(t)/t$ automatically satisfies the inclusion constraint in problem (2.1). We treat $x(t)/t$ as a candidate for tracking X^0. Let us find a control law ensuring $Fx(t)/t = b$, or, equivalently, $\epsilon(t) = |Fx(t) - bt|^2 = 0$ for all $t \geq 0$. The zero initial condition implies $\epsilon(0) = 0$. So, it is sufficient to have $\dot{\epsilon}(t) \leq 0$. The differentiation yields $\dot{\epsilon}(t) = 2(Fx(t) - bt)^T(Fu(t) - b)$. We get $\dot{\epsilon}(t) \leq 0$ by letting $u(t) \in L^-(t, x(t))$, where

$$L^-(t,x) = \{u \in M : (Fx - bt)^T(Fu - b) \leq 0\}. \tag{2.3}$$

Any $u(t) \in L^-(t, x(t))$ keeps $x(t)/t$ within the feasible set of problem (2.1) for all $t > 0$. We associate this argument with the method of shifting control [7]. Control $u(t) \in L^-(t, x(t))$ shifts discrepancy $\epsilon(t)$ toward zero. Now let $u(t)$ minimize J in $L^-(t, x(t))$, i.e., $u(t) \in U^s(t, x(t))$, where

$$U^s(t,x) = \operatorname{argmin}\{J(u) : u \in L^-(t,x)\}. \tag{2.4}$$

We call $U^s(t, x)$ the *shifting feedback*. We have $Fx^0 - b = 0$, where x^0 is a solution of problem (2.1). Hence $(Fx(t) - bt)^T(Fx^0 - b) = 0$, and $x^0 \in L^-(t, x(t))$. Therefore $J(u(t)) \leq J(x^0) = J^0$ for all $t > 0$. Due to the convexity of J,

$$J\left(\frac{x(t)}{t}\right) \leq \frac{1}{t}\int_0^t J(u(s))ds \leq \frac{1}{t}\int_0^t J^0 = J^0.$$

Now take into account that $x(t)/t$ is feasible in problem (2.1). We immediately get that $x(t)/t$ is a solution of (2.1), or, equivalently, $x(t)/t \in X^0$ for all $t > 0$. We find that $x(t)$, a trajectory of the control system (2.2) under the shifting feedback $U^s(t, x(t))$, tracks the solution set X^0.

Unfortunately, the above argument does not work, for we cannot guarantee the existence of a trajectory under the shifting feedback $U^s(t, x)$. Let us give a simple nonexistence example. We minimize the scalar variable x under the trivial constraints $x \in [-1, 1]$ and $x = 0$. We have $L^-(t, x) = [-1, 0]$ for $x \geq 0$, and $L^-(t, x) = [0, 1]$ for $x < 0$. Hence $U^s(t, x) = \{u^s(t, x)\}$, where

$$u^s(t,x) = \begin{cases} -1, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

The differential equation for system (2.2) closed with feedback $U^s(t, x)$ is

$$\dot{x}(t) = u^s(t, x(t)), \quad x(0) = 0. \tag{2.5}$$

Its right-hand side is discontinuous. Obviously, this equation has neither classical nor Carathéodory solutions.

The paradox disappears if we assume Filippov's definition of solutions. Following [12], we identify a solution of (2.5) with a solution of the differential inclusion

$$\dot{x}(t) \in U^{fs}(t, x(t)), \quad x(0) = 0.$$

Here

$$U^{fs}(t,x) = \begin{cases} \{-1\}, & x > 0, \\ \{0\}, & x < 0, \\ [-1, 0], & x = 0. \end{cases}$$

It is easily seen that $x(t) = 0$ is a single Filippov solution of (2.5) and $x(t)/t \in X^0$ for all $t > 0$ ($X^0 = \{0\}$).

In section 3 the features of the above example will be extended to the general case.

In what follows, we deal with multivalued feedbacks for system (2.2). A (multivalued) *feedback* U is defined to be a map from $R_+ \times R^n$ into the collection of all nonempty sets in M; here $R_+ = [0, \infty]$. A feedback U such that $U(t,x)$ is one-element for all $(t,x) \in E$ is called *single valued on* E; if $E = R_+ \times R^n$, we call U *single valued.* A *trajectory* (of system (2.2)) *under feedback* U is a solution of

$$\dot{x}(t) \in U(t, x(t)), \quad x(0) = 0. \tag{2.6}$$

More accurately, a function $x(\cdot) : R_+ \mapsto R^n$ is called a trajectory under U if $x(\cdot)$ is absolutely continuous on every bounded interval, $x(0) = 0$, and the inclusion in (2.6) holds for almost all $t \geq 0$.

The convexity of M and the fact that the values of feedbacks are contained in M yield the following.

LEMMA 2.1. *Let* $x(\cdot)$ *be a trajectory under an arbitrary feedback. Then* $x(t)/t \in M$ *for all* $t > 0$.

We conclude this section with some known definitions and properties of multivalued maps; they will be referred to in what follows. A multivalued (set-valued) map $\mathcal{F}$ on $E \subset R_+ \times R^n$ associates with every $(t,x) \in E$ a nonempty set $\mathcal{F}(t,x) \subset R^n$; if $E = R_+ \times R^n$, we call $\mathcal{F}$ a multivalued map (without mentioning its set of definition). The continuity, upper semicontinuity, and lower semicontinuity of a multivalued map $\mathcal{F}$ on E at a point are understood in a standard way (see [13, pp. 41, 43]). A multivalued map $\mathcal{F}$ whose restriction to a set E is continuous at every point $(t,x) \in E$ is said to be continuous on E; if $E = R_+ \times R^n$, then $\mathcal{F}$ is called continuous. The same applies for the upper and lower semicontinuity. The next two lemmas follow from standard results of set-valued analysis (see [14, section 1.4]).

LEMMA 2.2. *Let* $\mathcal{F}$ *be a closed and convex-valued multivalued map continuous on* E, *and* f *be a convex function on* R^n. *Then the multivalued map* $(t,x) \mapsto \operatorname{argmin}\{f(u) : u \in \mathcal{F}(t,x)\}$ *is closed and convex valued and upper semicontinuous on* E.

LEMMA 2.3. *Let* $E_1, \ldots, E_m$ *be closed sets in* $R_+ \times R^n$, *and a multivalued map* $\mathcal{F}$ *be upper semicontinuous on each of them. Then* $\mathcal{F}$ *is upper semicontinuous on* $\cup\{E_j : j = 1, \ldots, m\}$.

3. Solution-tracking dynamics. In this section we extend the argument of section 2 to the general case. We state that the trajectories under the shifting feedback are solution tracking but usually do not exist. Next, we introduce the Filippov shifting feedback. We prove that there always exists a trajectory under this feedback, and every such trajectory is solution tracking.

We call a function $x(\cdot) : R_+ \mapsto R^n$ *solution tracking* if $x(t)/t \in X^0$ for all $t > 0$. Let us fix a continuous feedback Q such that $Q(t,x) \cap X^0$ is nonempty for all $(t,x) \in R_+ \times R^n$. The *shifting* feedback U^s is defined by (2.4), where

$$L^-(t,x) = \{u \in Q(t,x) : (Fx - bt)^T(Fu - b) \leq 0\} \tag{3.1}$$

(in (2.3) we have $Q(t,x) = M$). Note that U^s is closed and convex valued.

Let us give a criterion for a function $x(\cdot)$ to be a trajectory under the shifting feedback. We set

$$Q^0(t,x) = \operatorname{argmin}\{J(u) : u \in Q(t,x)\}. \tag{3.2}$$

Lemma 2.3 and the continuity of Q yield the following.

LEMMA 3.1. *Feedback Q^0 (3.2) is upper semicontinuous.*

THEOREM 3.1. *Let $x(\cdot) : R_+ \mapsto R^n$ be absolutely continuous on every bounded interval, and $x(0) = 0$. Then the next statements are equivalent:*

(i) *$x(\cdot)$ is a trajectory under the shifting feedback;*

(ii) *$\dot{x}(t) \in Q^0(t, x(t)) \cap X^0$ for almost all $t \geq 0$.*

Proof. Let (i) hold. By the definition of the shifting feedback (see (2.4)), $\dot{x}(t) \in L^-(t, x(t))$ for almost all t. Hence (see (3.1)), the function $\epsilon(\cdot) : t \mapsto \mid Fx(t) - tb \mid^2$ is nonincreasing. Since $\epsilon(0) = 0$, we have $Fx(t) = tb$ for all $t \geq 0$. Therefore $L^-(t, x(t)) = Q(t, x(t))$ for $t \geq 0$ (see (3.1)). By (3.2) $U^s(t, x(t)) = Q^0(t, x(t))$. Consequently $\dot{x}(t) \in Q^0(t, x(t))$ for almost all $t \geq 0$. In particular $J(\dot{x}(t)) \leq J^0$ for almost all $t \geq 0$. It follows from $Fx(t) = tb$ that $\dot{x}(t)$ is feasible in problem (2.1) for almost all $t \geq 0$. Hence $J(\dot{x}(t)) = J^0$ for almost all $t \geq 0$. Therefore $\dot{x}(t) \in X^0$ for almost all $t \geq 0$.

Let (ii) hold. For almost all t, $\dot{x}(t)$ is simultaneously a minimizer of J in $Q(t, x(t))$ and a minimizer of J in the feasible set of problem (2.1). Since $L^-(t, x(t))$ contains the feasible set of problem (2.1) and is contained in $Q(t, x(t))$, $\dot{x}(t)$ minimizes J in $L^-(t, x(t))$, or, equivalently, $\dot{x}(t) \in U^s(t, x(t))$ for almost all t. □

COROLLARY 3.1. *Every trajectory under the shifting feedback is solution tracking.*

Proof. Let $x(\cdot)$ be a trajectory under the shifting feedback. By Theorem 3.1(ii), $\dot{x}(t) \in X^0$ for almost all $t \geq 0$. Since X^0 is closed and convex, $x(t)/t \in X^0$ for all $t > 0$. □

COROLLARY 3.2. *Let there exist a trajectory under the shifting feedback. Then $\cup\{Q_0(t, x) \cap X^0 : x \in tX^0\}$ is nonempty for all $t \geq 0$.*

Proof. Let $x(\cdot)$ be a trajectory under the shifting feedback. By statement (ii) of Theorem 3.1, $x(t) \in tX^0$ for all $t \geq 0$ and $Q^0(t, x(t)) \cap X^0$ is nonempty for almost all $t \geq 0$. Due to the upper semicontinuity of feedback Q^0 (Lemma 3.1), $Q^0(t, x(t)) \cap X^0$ is nonempty for all $t \geq 0$. □

COROLLARY 3.3. *Let X^0 be one-element. The next statements are equivalent:*

(i) *there exists a trajectory under the shifting feedback;*

(ii) *for all $t \geq 0$ it holds that $x^0 \in Q^0(t, tx^0)$, where x^0 is a single element in X^0.*

Proof. If (i) holds, then (ii) is true by Corollary 3.2. Let (ii) hold. By Theorem 3.1, $x(\cdot) : t \mapsto tx^0$ is a trajectory under the shifting feedback. □

The necessary condition for the existence of a trajectory under the shifting feedback given in Corollary 3.2 is a severe constraint on the problem's data and feedback Q. For instance, if $Q(t, x) = M$ (see section 2), this condition requires that a minimizer in the original problem (2.1) minimizes J in M. The latter situation is degenerate.

To avoid the nonexistence phenomenon, we introduce a Filippov modification of the shifting feedback. In what follows, conv E designates the closure of the convex hull of a set $E \subset R^n$. The *Filippov shifting feedback* U^{fs} is defined by

$$U^{fs}(t, x) = \cap_{\delta > 0} \operatorname{conv} \cup \{U^s(t, y) : |y - x| \leq \delta\}. \tag{3.3}$$

THEOREM 3.2. *There exists a trajectory under the Filippov shifting feedback, and every such trajectory is solution tracking.*

Proof. We can easily verify that the Filippov shifting feedback U^{fs} is upper semicontinuous. Obviously, it is bounded. These properties imply the existence of a trajectory under U^{fs} (see, e.g., [13, Theorem 4, p. 101]). Now let $x(\cdot)$ be a trajectory under U^{fs}. We must prove that $x(\cdot)$ is solution tracking. Take a small $\delta > 0$ and arbitrary $t \geq 0$ such that $\dot{x}(t) \in U^{fs}(t, x)$. By (3.3) there are points $y_1, \ldots, y_k$ in the closed δ-neighborhood of $x(t)$, elements $u_i \in U^s(t, y_i)$ $(i = 1, \ldots, k)$, and coefficients $\beta_i \geq 0$ $(i = 1, \ldots, k)$, $\sum_{i=1}^k \beta_i = 1$, such that

$$\left| \dot{x}(t) - \sum_{i=1}^{k} \beta_i u_i \right| < \delta. \tag{3.4}$$

By the definition of $U^s(t, x)$ (see (2.4)), $u_i \in L^-(t, y_i)$. So, $(Fy_i - tb)^T(Fu_i - b) \leq 0$. Fix an arbitrary $\mu > 0$, and, with no loss of generality, take δ so small that $(Fx(t) - tb)^T(Fu_i - b) \leq \mu$ for all i. Then $(Fx(t) - tb)^T$ $(F \sum_{i=1}^k \beta_i u_i - b) \leq \mu$. Due to (3.4) we can choose δ so small that

$$(Fx(t) - tb)^T(F\dot{x}(t) - b) \leq 2\mu. \tag{3.5}$$

Note that $Q(t, y_i) \cap X^0$ is nonempty for every i. Let $x_i^0 \in Q(t, y_i) \cap X^0$. Obviously, $(Fy_i - tb)^T(Fx_i^0 - b) = 0$. Hence $x_i^0 \in X^0 \cap L^-(t, y_i)$. Consequently $J(u_i) \leq J(x^0) = J^0$ (see (2.4)). By the convexity of J we have $J(\sum_{i=1}^k \beta_i u_i) \leq J^0$. Taking into account (3.4) and (if needed) decreasing δ, we get

$$J(\dot{x}(t)) \leq J^0 + \mu. \tag{3.6}$$

In (3.5) and (3.6) $\mu > 0$ is arbitrary. Omitting μ yields

$$(Fx(t) - tb)^T(F\dot{x}(t) - b) \leq 0, \tag{3.7}$$

$$J(\dot{x}(t)) \leq J^0. \tag{3.8}$$

Recall that t is arbitrary and satisfies $\dot{x}(t) \in U^{fs}(t, x)$. The latter holds for almost all $t \geq 0$. Therefore for almost all $t \geq 0$ estimates (3.7) and (3.8) are valid. Set $\epsilon(t) = |Fx(t) - tb|^2$. Obviously $\epsilon(0) = 0$. By (3.7) $\dot{\epsilon}(t) \leq 0$ for almost all $t \geq 0$. Thus for all $t \geq 0$ we have $\epsilon(t) = 0$, or $Fx(t)/t = b$. So far as $x(t)/t \in M$ (Lemma 2.1), $x(t)/t$ is feasible in problem (2.1) for all $t > 0$. Now integrate (3.8) and use the convexity of J. For arbitrary $t > 0$ we obtain

$$J\left(\frac{x(t)}{t}\right) \leq \frac{1}{t}\int_0^t J(\dot{x}(s))ds \leq \frac{1}{t}\int_0^t J^0 = J^0. \tag{3.9}$$

Since $x(t)/t$ is feasible in problem (2.1), $x(t)/t$ is a minimizer in (2.1). Equivalently, $x(t)/t \in X^0$. Thus trajectory $x(\cdot)$ is solution tracking. □

Note that if J is strictly convex, the shifting feedback U^s is single valued (see (2.4)). This is not true for the Filippov shifting feedback U^{fs}.

THEOREM 3.3. *Let J be strictly convex, and the Filippov shifting feedback be single valued. Then there exists a trajectory under the shifting feedback.*

Proof. Since J is strictly convex, each of the sets X^0 and Q^0 is one-element. We denote by x^0 a single element in X^0 and by $q^0(t, x)$ a single element in $Q^0(t, x)$. Suppose there does not exist a trajectory under the shifting feedback. Then statement (ii) of Corollary 3.3 is untrue; i.e., there is a $\tau \geq 0$ such that $x^0 \neq q$, where $q = q^0(\tau, \tau x^0)$. Then q (lying in M) does not satisfy the equality constraint in problem (2.1), $Fq \neq b$. Take $\epsilon > 0$ and set $y_\epsilon^- = \tau x^0 - (\epsilon q - x^0)$. We have

$$\begin{aligned}(Fy_\epsilon^- - \tau b)^T(Fq - b) &= (F\tau x^0 - \tau b - \epsilon F(q - x^0))^T(Fq - b)\\ &= -\epsilon(Fq - b)^T(Fq - b) < 0.\end{aligned}$$

By Lemma 3.1 the function $(t, x) \mapsto q^0(t, x)$ is continuous. Consequently for sufficiently small $\epsilon > 0$, $(Fy_\epsilon^- - \tau b)^T\ (Fq^0(\tau, y_\epsilon^-) - b) < 0$, or $q^0(\tau, \tau y_\epsilon^-) \in L^-(\tau, y_\epsilon^-)$. Hence for small $\epsilon > 0$, $q^0(\tau, \tau y_\epsilon^-)$ minimizes J in $L^-(\tau, y_\epsilon^-)$, i.e., $q^0(\tau, y_\epsilon^-) \in U^s(\tau, y_\epsilon^-)$. By the definition of the Filippov shifting feedback, $q = \lim_{\epsilon\to+0} q^0(\tau, y_\epsilon^-)$ is contained in $U^{fs}(\tau, \tau x^0)$. Now let

$$y_\epsilon^+ = \tau x^0 + \epsilon(q - x^0),$$

$u_\epsilon^+ \in U^s(\tau, y_\epsilon^+)$, and u^+ be a limit point of the sequence $(u_{1/i}^+)$. By (3.3) $u^+ \in U^{fs}(\tau, \tau x^0)$. Since $u_{1/i}^+ \in L^-(\tau, y_{1/i}^+)$, we have $(Fy_{1/i}^+ - \tau b)^T\ (Fu_{1/i}^+ - b) \le 0$. Dividing by ϵ and using $Fx^0 = b$, we get $(Fq - b)^T\ (Fu_{1/i}^+ - b) \le 0$. Hence $(Fq - b)^T$ $(Fu^+ - b) \le 0$. If $u^+ = q$, then $Fq - b = 0$. The latter does not hold by assumption. Therefore $u^+ \ne q$. Thus $U^{fs}(\tau, \tau x^0)$ contains the two different elements q and u^+. We find that the Filippov shifting feedback U^{fs} is not single valued, which contradicts the assumption. □

4. Solution-abandoning dynamics. The single-validity is a useful property of a feedback. It allows us to design trajectories using standard ODE techniques. Theorem 3.3 says that the Filippov shifting feedback U^{fs} is rarely single valued.

In this section we consider a feedback that is single valued on $(0, \infty) \times R^n$ (provided J is strictly convex) whose trajectories possess a property slightly weaker than solution tracking (we call it solution abandoning). The single-validity on $(0, \infty) \times R^n$, i.e., everywhere except the hyperplane $t = 0$, is less convenient than the single-validity everywhere but still gives a chance for using ODE techniques. The trajectories under a feedback U that is single valued on $(0, \infty) \times R^n$ are governed by an ODE whose right-hand side is well defined everywhere except $t = 0$. If the right-hand side is continuous on $(0, \infty) \times R^n$ (which is the case in this section), the solutions of the ODE are continuable to zero and bounded in its neighborhood. A solution $x(\cdot)$ is a trajectory under U whenever $\lim_{t\to+0} x(t) = 0$.

In what follows, $\operatorname{dist}(x, X)$ stands for the distance from a point $x \in R^n$ to a set $X \subset R^n$. We call a function $x(\cdot) : R_+ \mapsto R^n$ *solution abandoning* if $\operatorname{dist}(x(t)/t, X^0) \to 0$ as $t \to +0$. Define the *regularized shifting* feedback U^{rs} by

$$U^{rs}(t, x) = \begin{cases} (1 - \omega(t, x))U^s(t, x) + \omega(t, x)Q^0(t, x), & t > 0, \\ Q(0, x), & t = 0. \end{cases} \tag{4.1}$$

Here $\omega(\cdot)$ is a continuous *pasting function* defined on $(0, \infty) \times R^n$. It satisfies $\omega(t, x) = 0$ if $|Fx - tb| \ge \rho(t)$, and $\omega(t, x) = 1$ if $|Fx - tb| = 0$. A continuous *barrier function* $\rho(\cdot)$ is defined on R_+, takes positive values everywhere except the origin, vanishes at the origin, and satisfies

$$\lim_{t\to+0} \frac{\rho(t)}{t} = 0. \tag{4.2}$$

If J is strictly convex, then feedbacks U^s (2.4) and Q^0 (3.2) are single valued, hence U^{rs} is single valued on $(0, \infty) \times R^n$. To state the existence of a trajectory under the regularized shifting feedback we need to assume that feedback L^- (3.1) is continuous on

$$E^+ = \{(t, x) : t > 0,\ x \in R^n,\ Fx - tb \ne 0\}. \tag{4.3}$$

Note that the continuity of L^- on E^+ does not imply its continuity on $R_+ \times R^n$. For instance, for problem (2.1) with the trivial constraints $x \in [-1,1]$, $x = 0$, and $Q(t,x) = [-1,1]$, we have $L^-(t,x) = [-1,0]$ if $x > 0$, $L^-(t,x) = [0,1]$ if $x > 0$, and $L^-(t,x) = [-1,1]$ if $x = 0$. We see that L^- is continuous on $E^+ = \{(t,x) : t > 0,\ x \in R^1,\ x \neq 0\}$ and discontinuous at points $(t,0)$.

Let us give a condition sufficient for L^- to be continuous on E^+. We write int E for the interiority of a set $E \subset R^n$.

Lemma 4.1. *Let* int $Q(t,x)$ *intersect the feasible set of problem* (2.1) *for every* $(t,x) \in E^+$. *Then feedback* L^- *is continuous on* E^+.

Proof. Since Q is continuous, L^- is upper semicontinuous. We complete the proof by showing that L^- is lower semicontinuous on E^+. Let $L_0^-(t,x) = \{u \in R^n : (Fx - bt)^T(Fu - b) \leq 0\}$. By (3.1) $L^-(t,x) = \{u \in L_0^-(t,x)\ : u \in Q(t,x)\}$. Then by Proposition 1.5.2 in [15], the next conditions are sufficient for L^- to be lower semicontinuous on E^+:

(i) L_0^- is lower semicontinuous on E^+;

(ii) $Q(t,x)$ is convex;

(iii) the graph of the multivalued map $(t,x) \mapsto \text{int } Q(t,x)$ defined on E^+ is open; and

(iv) $L_0^-(t,x) \cap \text{int } Q(t,x)$ is nonempty for all $(t,x) \in E^+$.

Condition (i) is obviously satisfied. Conditions (ii) and (iv) follow from the assumptions. It remains to prove (iii). Take arbitrary $(t_*,x_*) \in E^+$ and $u_* \in \text{int } Q(t,x)$. Let $\epsilon > 0$ be the radius of a neighborhood of u_* contained in $Q(t_*,x_*)$. Then

$$\Delta(\psi \mid Q(t_*,x_*)) - \epsilon|\psi| \geq \psi^T u_*$$

for all $\psi \in R^n$. Here and in what follows $\Delta(\cdot \mid D)$ is the support function of $D \subset R^n$, i.e., $\Delta(\psi \mid D) = \sup\{\psi^T u : u \in D\}$. By the continuity of Q, if (τ,ξ) is sufficiently close to (t_*,x_*), then

$$\Delta(\psi \mid Q(\tau,\xi)) \geq \Delta(\psi \mid Q(t_*,x_*)) - \frac{\epsilon}{2}|\psi|$$

for all $\psi \in R^n$. Hence for (τ,ξ) close to (t_*,x_*) it holds that

$$\Delta(\psi \mid Q(\tau,\xi)) - \frac{\epsilon}{2}|\psi| \geq \psi^T u_*$$

for all $\psi \in R^n$; equivalently, the $(\epsilon/2)$-neighborhood of u_* is contained in $Q(\tau,\xi)$. Thus, (t_*,x_*,u_*) lies in the interior of the graph of $(t,x) \mapsto \text{int } Q(t,x)$. This graph is therefore open. Condition (iii) is proved. □

Lemma 4.2. *Let feedback* L^- *be continuous on* E^+. *Then the regularized shifting feedback is closed and convex valued and upper semicontinuous.*

Proof. Observing (4.1), we easily see that the regularized shifting feedback U^{rs} is closed and convex valued. We must prove that U^{rs} is upper semicontinuous. Let

$$D^+ = \{(t,x) \in (0,\infty) \times R^n : |Fx - tb| \geq \rho(t)\},$$
$$D^- = \{(t,x) \in (0,\infty) \times R^n : |Fx - tb| \leq \rho(t)\},$$
$$D_0^+ = D^+ \cup (\{0\} \times R^n),$$
$$D_0^- = D^- \cup (\{0\} \times R^n).$$

Obviously D_0^+ and D_0^- are closed, and $D_0^+ \cup D_0^- = R_+ \times R^n$. In view of Lemma 2.3 it is sufficient to prove that U^{rs} is upper semicontinuous on each of these sets. Note that $U^{rs}(t,x) \subset Q(t,x)$ and $U^{rs}(0,x) = Q(0,x)$ (see (4.1)). Since Q is continuous, U^{rs} is upper semicontinuous on $\{0\} \times R^n$. It remains to show that U^{rs} is upper semicontinuous on D^+ and D^-. By assumption L^- is continuous on D^+. Then by Lemma 2.2 the shifting feedback U^s is upper semicontinuous on D^+. On the other hand, U^{rs} and U^s coincide on D^+ (see (4.1)). Thus U^{rs} is upper semicontinuous on D^+. Take arbitrary $(t_*, x_*) \in D^-$ and show that U^{rs} is upper semicontinuous at (t_*, x_*). Let $Fx_* \neq bt_*$. Feedback L^- is continuous at (t_*, x_*). Hence by Lemma 2.2 U^s is upper semicontinuous at (t_*, x_*). By Lemma 3.1 Q^0 is upper semicontinuous at (t_*, x_*). Observe formula (4.1) for $U^{rs}(t,x)$. So far as $\omega(\cdot)$ is continuous, and U^s and Q^0 are upper semicontinuous at (t_*, x_*), U^{rs} is upper semicontinuous at (t_*, x_*). Let $Fx_* = bt_*$. Now formula (4.1) for $U^{rs}(t,x)$ and the facts that Q^0 is upper semicontinuous at (t_*, x_*) and $\omega(t,x) \to 1$ as $(t,x) \to (t_*, x_*)$ yield that $U^{rs}(t,x)$ lies in an arbitrary neighborhood of $U^{rs}(t_*, x_*) = Q^0(t_*, x_*)$ provided (t,x) is sufficiently close to (t_*, x_*). In other words, U^{rs} is upper semicontinuous at (t_*, x_*). We have proved that U^{rs} is upper semicontinuous at every point in D^-. □

A key property of the trajectories under the regularized shifting feedback is as follows.

Lemma 4.3. *Let $x(\cdot)$ be a trajectory under the regularized shifting feedback. Then for all $t > 0$*

$$J\left(\frac{x(t)}{t}\right) \leq J^0, \tag{4.4}$$

$$\left| F\frac{x(t)}{t} - b \right| \leq \frac{\rho(t)}{t}. \tag{4.5}$$

Proof. Take a $t \geq 0$ such that $\dot{x}(t) \in U^{rs}(t, x(t))$. By (4.1) $\dot{x}(t) = (1 - \omega(t,x(t)))u_1(t) + \omega(t,x(t))u_2(t)$, where $u_1(t) \in U^s(t,x(t))$ and $u_2(t) \in Q^0(t,x(t))$. By the definition of $Q^0(t,x(t))$ (see (3.2)), $J(u_2(t)) \leq J^0$. The definition of $U^s(t,x(t))$ (see (2.4)) and the fact that $Q(t,x(t))$ intersects X^0 imply that $J(u_1(t)) \leq J^0$. Then the convexity of J and the inclusion $\omega(t,x(t)) \in [0,1]$ yield

$$J(\dot{x}(t)) \leq (1 - \omega(t,x(t)))J(u_1(t)) + \omega(t,x(t))J(u_1(t)) \leq J^0.$$

We have proved that $J(\dot{x}(t)) \leq J^0$ for almost all $t \geq 0$. Now using the convexity of J, we get (4.4) for all $t > 0$ (a detailed argument is given in (3.9)). Let us prove that (4.5) holds for all $t > 0$. Suppose the contrary, i.e., $\epsilon(t^*) > \rho(t^*)$ for some $t^* \geq 0$, where $\epsilon(t) = |Fx(t) - bt|$. Note that $\epsilon(0) = 0 = \rho(0)$. Hence $t^* > 0$, and there is a $t_* \in (0, t^*)$ such that $\epsilon(t_*) < \epsilon(t^*)$ and $\epsilon(t) > \rho(t)$ for all $t \in [t_*, t^*]$. By the definition of the pasting function $\omega(\cdot)$ we have $\omega(t,x(t)) = 0$ for $t \in [t_*, t^*]$. Therefore $U^{rs}(t,x(t)) = U^s(t,x(t))$ for $t \in [t_*, t^*]$. Hence $\dot{x}(t) \in U^s(t,x(t))$ for almost all $t \in [t_*, t^*]$. Since $U^s(t,x(t)) \subset L^-(t,x(t))$ (see (2.4)), $d\epsilon^2(t)/dt = 2(Fx(t) - bt)^T(F\dot{x}(t) - b) \leq 0$ for almost all $t \in [t_*, t^*]$. Thus $\epsilon(\cdot)$ is nonincreasing on $[t_*, t^*]$, which is not possible, since by supposition $\epsilon(t_*) < \epsilon(t^*)$. The obtained contradiction proves that (4.5) holds for all $t > 0$. □

Theorem 4.1. *Let feedback L^- be continuous on E^+. Then there exists a trajectory under the regularized shifting feedback, and every such trajectory is solution abandoning.*

Proof. By Lemma 4.2 the regularized shifting feedback U^{rs} is upper semicontinuous. Obviously, it is bounded. These properties imply the existence of a trajectory under U^{rs} (see [14, Theorem 4, p. 101]). Let $x(\cdot)$ be a trajectory under the regularized shifting feedback. By Lemma 4.3 the estimates (4.4) and (4.5) hold for all $t > 0$. Estimate (4.5) and the fact that $\rho(\cdot)$ satisfies (4.2) imply that $|Fx(t)/t - b| \to 0$ as $t \to +0$. The latter convergence and the inclusion $x(t)/t \in M$ holding for all $t > 0$ (Lemma 2.1) yield that the distance from $x(t)/t$ to the feasible set in problem (2.1) goes to zero as $t \to +0$. Then in view of (4.4), $\lim_{t\to+0} \operatorname{dist}(x(t)/t, X^0) = 0$, i.e., trajectory $x(\cdot)$ is solution abandoning. □

Lemma 4.3 provides additional information on a solution-abandoning trajectory $x(\cdot)$ under the regularized shifting feedback. In particular, we see that as $t \to +0$, values $J(x(t)/t)$ converge to the optimal value J^0 from below.

5. Solution-abandoning dynamics: Example. Here we give an example of a solution-abandoning trajectory under the regularized shifting feedback.

Consider the problem of finding the projection of a vector $z \in R^n$ onto the hyperplane H orthogonal to a vector $q \in R^n$. This problem is represented in the form (2.1), where $J(x) = |x - z|^2$, $F = q^T$, and M is a closed ball centered at zero and containing z. We assume that $q^T z < 0$ and denote by x^0 the projection of z onto H, i.e., the solution of problem (2.1). Set $Q(t, x) = M$. Note that the zero element is feasible in problem (2.1) and lies in the interior of $Q(t, x)$. Therefore by Lemma 4.1 feedback L^- is continuous on E^+. Hence by Theorem 4.1 there exists a trajectory under the regularized shifting feedback U^{rs}, and every such trajectory is solution abandoning. Let us construct a trajectory under U^{rs}. The shifting feedback U^s (2.4) has the form

$$U^s(t,x) = \begin{cases} \{z\}, & q^T x > 0, \\ \{x^0\}, & q^T x \le 0. \end{cases}$$

We have $Q^0(t, x) = \{x^0\}$ (see (3.2)). Hence the regularized shifting feedback U^{rs} (4.1) is given by

$$U^{rs}(t,x) = \begin{cases} \{u^{rs}(t,x)\}, & t > 0, \\ M, & t = 0, \end{cases}$$

where

$$u^{rs}(t,x) = \begin{cases} z, & q^T x > 0, \\ (1 - \omega(t,x))x^0 + \omega(t,x)z, & -\rho(t) < q^T x \le 0, \\ x^0, & q^T x \le -\rho(t). \end{cases}$$

We take $\rho(t) = t^2$ for the barrier function, and $\omega(t, x) = \max\{0, (1 - |q^T x|)/t^2\}$ for the pasting function. Then a trajectory under U^{rs} is described by the ODE

$$\dot{x}(t) = x^0 + \zeta(t, q^T x(t))(z - x^0), \quad x(0) = 0, \tag{5.1}$$

where

$$\zeta(t,y) = \begin{cases} 1, & y \ge 0, \\ 1 + \frac{y}{t^2}, & -t^2 < y < 0, \\ 0, & y \le -t^2. \end{cases}$$

Let $x(\cdot)$ be a solution of (5.1). Scalar multiplying (5.1) by q, we find that $y(t) = q^T x(t)$ solves

$$\dot{y}(t) = q^T x^0 + \zeta(t, y(t)) q^T (z - x^0), \quad y(0) = 0,$$

or, as long as $q^T x^0 = 0$,

$$\dot{y}(t) = -a\zeta(t, y(t)), \quad y(0) = 0, \tag{5.2}$$

where $a = -q^T z > 0$. The right-hand side in (5.2) is decreasing in the state variable. Hence $y(\cdot)$ is a unique solution of (5.2). We can easily verify that $y(\cdot)$ solves the linear ODE

$$\dot{y}(t) = -a\left(1 + \frac{y(t)}{t^2}\right), \quad y(0) = 0,$$

and is given by

$$y(t) = -a\left(\int_{1/t}^{\infty} \frac{e^{-a\tau}}{\tau^2} d\tau\right) e^{-a/t}$$

for $t > 0$. Obviously $y(t) \leq 0$. Also, we have

$$-y(t) \leq \frac{1}{(1/t)^2} a \left(\int_{1/t}^{\infty} e^{-a\tau} d\tau\right) e^{-a/t} = at^2 \int_0^{\infty} e^{-a\tau} d\tau = t^2.$$

Thus

$$-t^2 \leq y(t) \leq 0. \tag{5.3}$$

Coming back to (5.1) and taking into account (5.2), we find that $x(\cdot)$ solves the equation

$$\dot{x}(t) = x^0 - \frac{\dot{y}(t)}{a}(z - x^0), \quad x(0) = 0.$$

Hence trajectory $x(\cdot)$ is unique and given by

$$x(t) = x^0 t - \frac{y(t)}{a}(z - x^0).$$

Owing to (5.3) we have

$$\left|\frac{x(t)}{t} - x^0\right| \leq \frac{t}{a}|z - x^0|. \tag{5.4}$$

Thus $x(t)/t$ converges to x^0, the solution of problem (2.1), as $t \to +0$. We found a formula for $x(\cdot)$, a (single) trajectory under the regularized shifting feedback, and showed that $x(\cdot)$ is solution abandoning.

6. Solution-approaching dynamics. In this section we introduce a feedback whose trajectories approach the solution set X^0 as time goes to infinity.

We call a function $x(\cdot) : R_+ \mapsto R^n$ *solution approaching* if $\text{dist}(x(t)/t, X^0) \to 0$ as $t \to \infty$. Note that a solution-abandoning trajectory (see section 4) identifies a solution within an arbitrarily short initial time interval. A solution-approaching trajectory finds a solution in infinite time. Thus, solution-approaching trajectories are much slower "solution identifiers" than solution-abandoning trajectories. On the other hand, the structure of the ODEs for solution-approaching trajectories is usually simpler.

We define the *penalized shifting* feedback U^{ps} by

$$U^{ps}(t,x) = \text{argmin}\{2(Fx - tb)^T Fu + \alpha J(u) : u \in Q(t,x)\}, \tag{6.1}$$

where α is a positive *penalty parameter.* A key property of the trajectories under the penalized shifting feedback is as follows.

LEMMA 6.1. *Let $x(\cdot)$ be a trajectory under the penalized shifting feedback. Then for all $t \geq 0$*

$$\lambda(t) \leq 0, \tag{6.2}$$

where

$$\lambda(t) = |Fx(t) - tb|^2 + \alpha \int_0^t J(\dot{x}(\tau))d\tau - \alpha t J^0. \tag{6.3}$$

Proof. For almost all t we have

$$\dot{\lambda}(t) = 2(Fx(t) - tb)^T (F(\dot{x}(t) - b) + \alpha J(\dot{x}(t))) - \alpha J^0.$$

Letting $x^0(t) \in X^0 \cap Q(t, x(t))$ (note that by assumption $Q(t, x(t))$ intersects X^0) and observing that $Fx^0(t) - b = 0$ and $J^0 = J(x^0(t))$, we continue as follows:

$$\begin{aligned}\dot{\lambda}(t) = {} & [2(Fx(t) - tb)^T (F(\dot{x}(t) - b) + \alpha J(\dot{x}(t)))] \\ & -[2(Fx(t) - tb)^T (F(x^0(t) - b) + \alpha J(x^0(t)))].\end{aligned}$$

Now take into account that for almost all $t \geq 0$ we have $\dot{x}(t) \in U^{ps}(t, x(t))$. Owing to (6.1) and the fact that $x^0 \in Q(t, x(t))$ we get $\dot{\lambda}(t) \leq 0$ for almost all $t \geq 0$. Since $\lambda(\cdot)$ is absolutely continuous and $\lambda(0) = 0$, (6.2) holds for all $t \geq 0$. □

The next lemma is a simple corollary of Lemma 6.1.

LEMMA 6.2. *Let $x(\cdot)$ be a trajectory under the penalized shifting feedback. Then for all $t > 0$ we have* (4.4) *and*

$$\left| F\frac{x(t)}{t} - b \right|^2 \leq \frac{\alpha K}{t}, \tag{6.4}$$

where $K = 2\max\{|J(u)| : u \in M\}$.

Proof. Take arbitrary $t > 0$. By Lemma 6.1, $\lambda(t)$ defined by (6.3) satisfies (6.2). Dividing (6.2) by t^2, we get

$$\left| F\frac{x(t)}{t} - b \right|^2 + \frac{1}{t^2}\alpha \int_0^t J(\dot{x}(\tau))d\tau - \frac{1}{t^2}\alpha t J^0 \leq 0. \tag{6.5}$$

Hence

$$\frac{1}{t}\int_0^t J(\dot{x}(\tau))d\tau \le J^0,$$

and we obtain (4.4) due to the convexity of J. Noticing that the second and third terms in (6.5) do not exceed $\alpha K/2t$, we arrive at (6.4). □

Theorem 6.1. *There exists a trajectory under the penalized shifting feedback, and every such trajectory is solution approaching.*

Proof. Observing (6.1), we see that the penalized shifting feedback U^{ps} is closed and convex valued. By Lemma 2.2 U^{ps} is upper semicontinuous. Obviously, it is bounded. These properties imply the existence of a trajectory under U^{rs} (see [14, Theorem 4, p. 101]). Let $x(\cdot)$ be a trajectory under the regularized shifting feedback. By Lemma 6.2, for all $t > 0$ we have (4.4) and (6.4). The latter estimate implies that $|Fx(t)/t - b| \to 0$ as $t \to \infty$. Recall that by Lemma 2.1, $x(t)/t \in M$. Hence we conclude that the distance from $x(t)/t$ to the feasible set of problem (2.1) tends to zero as $t \to \infty$. Then in view of (4.4), $\lim_{t\to\infty} \operatorname{dist}(x(t)/t, X^0) = 0$, i.e., trajectory $x(\cdot)$ is solution approaching. □

Lemma 6.2 provides additional information on trajectories $x(\cdot)$ under the penalized shifting feedback. In particular, we see that, as $t \to \infty$, values $J(x(t)/t)$ converge to the optimal value J^0 from below.

It is easily seen that the penalized shifting feedback U^{ps} is single valued if J is strictly convex. For some typical problems, including problems of linear and quadratic programming, the feedback U^{ps} is specified explicitly. Consider, for instance, a problem of quadratic programming under linear and box constraints, i.e., let $J(x) = |x|^2$ and $M = \{x \in R^n : x_i^- \le x \le x_i^+,\ \ i = 1, \dots, n\}$. Set $Q(t,x) = M$. Then $u^{ps}(t,x)$, a single element in $U^{ps}(t,x)$, is represented as follows. Let $v(t,x)$ be a minimizer of $u \mapsto 2(Fx - tb)^T Fu + \alpha|u|^2$ in R^r, i.e.,

$$v(t,x) = -\frac{F^T(Fx - tb)}{\alpha}.$$

We have

$$u^{ps}(t,x)_i = \begin{cases} v(t,x)_i, & x_i^- \le v(t,x)_i \le x_i^+, \\ x_i^-, & v(t,x)_i < x_i^-, \\ x_i^+, & v(t,x)_i > x_i^+ \end{cases}$$

$$(i = 1, \dots, n).$$

A trajectory under feedback U^{ps} is described by the ODE

$$\dot{x}(t) = u^{ps}(t, x(t)), \quad x(0) = 0.$$

Its right-hand side is continuous and piecewise linear, with a finite number of linearity domains. Each linearity domain is characterized by a finite number of linear inequalities. A trajectory is unique and represented explicitly in each linearity domain. In order to build the trajectory, we must identify the order of visiting the linearity domains. In the next section we give an example of an analysis of this kind.

7. Solution-approaching dynamics: Example. Here we consider a particular problem of quadratic programming with linear and box constraints. Our goal is to find an explicit solution through the analytic design of a trajectory under the penalized shifting feedback. The trajectories are described by a piecewise-linear ODE. First, we show a trajectory locked in a single linearity domain. Then we change the parameters and build a trajectory visiting two linearity domains.

The problem under consideration is a discrete counterpart of a one-dimensional linear quadratic optimal control problem with state constraints:

$$\text{minimize } \sum_{i=1}^{N+1} y_i^2 + \sum_{i=1}^{N} s_i^2,$$

$$\begin{gathered} y_{i+1} = y_i + s_i \quad (i = 1, \dots, N), \\ y_i \in [y_i^-, y_i^+] \quad (i = 1, \dots, N+1), \\ s_i \in [s_i^-, s_i^+] \quad (i = 1, \dots, N). \end{gathered}$$

We set

$$x = (y_1, \dots, y_{N+1}, s_1, \dots, s_N)^T \tag{7.1}$$

and represent the problem in the form (2.1), where

$$J(x) = |x|^2, \quad M = \prod_{i=1}^{N+1} [y_i^-, y_i^+] \times \prod_{i=1}^{N} [s_i^-, s_i^+],$$

$$F = (F_1 \ F_2), \quad b = 0.$$

The $N \times (N+1)$ matrix F_1 is given by

$$F_1 = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 & 0 \\ & & & \dots & & & \\ 0 & 0 & 0 & \dots & -1 & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & -1 & 1 \end{pmatrix},$$

and $F_2 = -I$, where I is an $N \times N$ identity matrix. We assume that the feasible set of problem (2.1) is nonempty. In our argument we use the matrix $F^T F$. We have

$$F^T F = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix},$$

where the $(N+1) \times (N+1)$ matrix G_{11} is given by

$$G_{11} = \begin{pmatrix} 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ & & & \dots & & & \\ 0 & 0 & 0 & \dots & 2 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{pmatrix},$$

the $(N+1) \times N$ matrix G_{12} is given by

$$G_{12} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ -1 & 1 & \dots & 0 & 0 \\ 0 & -1 & \dots & 0 & 0 \\ & & \dots & & \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & -1 & 1 \\ 0 & 0 & \dots & 0 & -1 \end{pmatrix},$$

and

$$G_{21} = G_{12}^T, \quad G_{22} = I.$$

We set $Q(t,x) = M$ and fix a penalty parameter $\alpha > 0$. The penalized shifting feedback U^{ps} is single valued. Let $u^{ps}(t,x)$ denote a single element in $U^{ps}(t,x)$. By (6.1), $u^{ps}(t,x)$ is a minimizer of $2(Fx)^T Fu + \alpha|u|^2$ in M. Note that $u^{ps}(t,x)$ does not depend on t; therefore we use the simpler notation $u^{ps}(x)$. Setting

$$u = (v_1, \dots, v_{N+1}, w_1, \dots, w_N)^T,$$

we get

$$2(Fx)^T Fu + \alpha|u|^2 = 2(F^T Fx)u + \alpha|u|^2$$

$$= 2\left(\sum_{i=1}^{N+1} f_i(x)v_i + \sum_{i=1}^{N} g_i(x)w_i\right) + \alpha \sum_{i=1}^{N+1} v_i^2 + \alpha \sum_{i=1}^{N} w_i^2, \tag{7.2}$$

where

$$(f_1(x), \dots, f_{N+1}(x), g_1(x), \dots, g_N(x)) = (F^T Fx)^T.$$

Using the form of $F^T F$ and (7.1), we obtain

$$\begin{gathered} f_1(x) = 2y_1 - y_2 + s_1, \\ f_i(x) = -y_{i-1} + 2y_i - y_{i+1} - s_{i-1} + s_i \ (i = 2, \dots, N), \\ f_{N+1}(x) = -y_N + 2y_{N+1} - s_N, \\ g_i(x) = y_i - y_{i+1} + s_i \ (i = 1, \dots, N). \end{gathered} \tag{7.3}$$

For

$$u^{ps}(x) = (v_1^{ps}(x), \dots, v_{N+1}^{ps}(x), w_1^{ps}(x), \dots, w_N^{ps}(x))^T,$$

a minimizer of (7.2) in M, we have

$$v_i^{ps}(x) = \begin{cases} -\frac{f_i(x)}{\alpha}, & y_i^- \le -\frac{f_i(x)}{\alpha} \le y_i^+, \\ y_i^-, & -\frac{f_i(x)}{\alpha} \le y_i^-, \\ y_i^+, & -\frac{f_i(x)}{\alpha} \ge y_i^+, \end{cases} \tag{7.4}$$

$$w_i^{ps}(x) = \begin{cases} -\frac{g_i(x)}{\alpha}, & s_i^- \le -\frac{g_i(x)}{\alpha}, \le s_i^+, \\ s_i^-, & -\frac{g_i(x)}{\alpha} \le s_i^-, \\ s_i^+, & -\frac{g_i(x)}{\alpha} \ge s_i^+. \end{cases} \tag{7.5}$$

A trajectory under the penalized shifting feedback U^{ps} is described by the ODE

$$\dot{x}(t) = u^{ps}(x(t)) \tag{7.6}$$

and the initial condition $x(0) = 0$. The right-hand side in (7.6) is linear in closed polyhedral domains. Each polyhedral linearity domain is defined by a combination of the linear inequalities given in (7.4) and (7.5). Within each of the linearity domains (7.6) is solved explicitly. We only need to identify the linearity domains visited by the trajectory. Here we restrict ourselves to the cases where the trajectory stays within a single linearity domain and crosses two linearity domains.

In what follows,

$$y_i^- > 0 \quad (i = 1, \dots, N+1), \quad s_i^- < 0 < s_i^+ \quad (i = 1, \dots, N). \tag{7.7}$$

Let $x(\cdot)$ be a trajectory under U^{ps}:

$$x(t) = (y_1(t), \dots, y_{N+1}(t), s_1(t), \dots, s_N(t))^T.$$

The initial conditions $y_i(0) = 0$, $s_i(0) = 0$, and assumptions (7.7) imply that for t close to the origin

$$-\frac{f_i(x(t))}{\alpha} < y_i^- \quad (i = 1, \dots, N+1), \tag{7.8}$$

$$s_i^- < -\frac{g_i(x(t))}{\alpha} < s_i^+ \quad (i = 1, \dots, N) \tag{7.9}$$

(see also (7.3)). These inequalities correspond to the linearity domain L_1, where

$$v_i^{ps}(x) = y_i^- \quad (i = 1, \dots, N+1),$$
$$w_i^{ps}(x) = -\frac{g_i(x)}{\alpha} \quad (i = 1, \dots, N)$$

(see (7.4) and (7.5)). Referring to the formula for $g_i(x)$ in (7.3), we find that

$$\dot{y}_i(t) = y_i^- \quad (i = 1, \dots, N+1), \tag{7.10}$$

$$\dot{s}_i(t) = -\frac{s_i(t)}{\alpha} + \frac{y_{i+1}(t) - y_i(t)}{\alpha} \quad (i = 1, \dots, N) \tag{7.11}$$

in a neighborhood of the origin. Explicitly,

$$y_i(t) = y_i^- t \quad (i = 1, \dots, N+1), \tag{7.12}$$

$$s_i(t) = (y_{i+1}^- - y_i^-)t - \alpha(y_{i+1}^- - y_i^-)(1 - e^{t/\alpha}) \quad (i = 1, \dots, N). \tag{7.13}$$

Case 1. Assume that

$$s_i^- \le y_{i+1}^- - y_i^- \le s_i^+ \quad (i = 1, \dots, N), \tag{7.14}$$

$$y_2^- - y_1^- \le y_1^-, \tag{7.15}$$

$$y_N^- - y_{N+1}^- \le 0, \tag{7.16}$$

$$y_{i-1}^- - 2y_i^- + y_{i+1}^- \le y_i^- \quad (i = 2, \dots, N). \tag{7.17}$$

Let us show that in this case $x(t)$ never escapes the linearity domain L_1, i.e., inequalities (7.8), (7.9) hold for all $t \ge 0$. Assume the contrary. Then $x(t)$ hits the boundary of L_1. Let ξ be the first hitting time, or, more accurately, the supremum of all $\tau \ge 0$ such that (7.8), (7.9) hold for all $t \in [0, \tau]$. We have either

$$-\frac{f_i(x(\xi))}{\alpha} = y_i^- \tag{7.18}$$

for some $i \in \{1, \dots, N+1\}$, or

$$-\frac{g_i(x(\xi))}{\alpha} \in \{s_i^-, s_i^+\} \tag{7.19}$$

for some $i \in \{1, \dots, N+1\}$. For all $t \in [0, \xi]$ the trajectory representation (7.12), (7.13) is valid. Substitute (7.12), (7.13) into the formulas for $f_i(x(t))$, $g_i(x(t))$ (see (7.3)). We get

$$-\frac{f_1(x(t))}{\alpha} = -\frac{y_1^- t}{\alpha} + (y_2^- - y_1^-)(1 - e^{-t/\alpha}), \tag{7.20}$$

$$-\frac{f_i(x(t))}{\alpha} = (y_{i-1}^- - 2y_i^- + y_{i+1}^-)(1 - e^{-t/\alpha}) \quad (i = 2, \dots, N), \tag{7.21}$$

$$-\frac{f_{N+1}(x(t))}{\alpha} = -\frac{y_{N+1}^- t}{\alpha} - (y_{N+1}^- - y_N^-)(1 - e^{-t/\alpha}), \tag{7.22}$$

$$-\frac{g_i(x(t))}{\alpha} = (y_{i+1}^- - y_i^-)(1 - e^{-t/\alpha}) \quad (i = 1, \dots, N) \tag{7.23}$$

for all $t \in [0, \xi]$. Estimate the right-hand sides using (7.15), (7.17), (7.7), (7.16), and (7.14). We obtain

$$-\frac{f_i(x(t))}{\alpha} < y_1^- \quad (i = 1, \dots, N+1),$$

$$-\frac{g_i(x(t))}{\alpha} \in (s_i^-, s_i^-) \quad (i = 1, \dots, N)$$

for all $t \in [0, \xi]$. Therefore (7.18) and (7.19) are violated. A contradiction proves that (7.8) and (7.9) hold true for all $t \ge 0$ and the trajectory representation (7.12), (7.13) is valid for all $t \ge 0$. Hence

$$\frac{y_i(t)}{t} = y_i^- \quad (i = 1, \dots, N+1),$$

$$\lim_{t \to \infty} \frac{s_i(t)}{t} = y_{i+1}^- - y_i^- \quad (i = 1, \dots, N).$$

Consequently

$$\lim_{t \to \infty} \frac{x(t)}{t} = x^0,$$

where

$$\begin{gathered} x^0 = (y_1^0, \dots , y_{N+1}^0, s_1^0, \dots , s_N^0)^T, \\ y_i^0 = y_i^- \quad (i = 1, \dots , N + 1), \\ s_i^0 = y_{i+1}^- - y_i^- \quad (i = 1, \dots , N). \end{gathered}$$

By Theorem 6.1 x^0 is a (unique) solution of problem (2.1).

Using the trajectory representation (7.12), (7.13), we can easily compute the discrepancies in constraints and objective values for the approximate solution $x(t)/t$. We have

$$\left| F\frac{x(t)}{t} \right| = \frac{\alpha}{t}(1 - e^{-t/\alpha}) \left(\sum_{i=1}^{N+1} (y_{i+1}^- - y_i^-)^2 \right)^{1/2},$$

$$J^0 - J\left(\frac{x(t)}{t}\right) = \frac{\alpha}{t}(1 - e^{-t/\alpha})[2 - (1 - e^{-t/\alpha})] \sum_{i=1}^{N+1} (y_{i+1}^- - y_i^-)^2.$$

The discrepancies vanish as $t \to \infty$. A simple analysis shows that discrepancy $J^0 - J(x(t)/t)$ is positive. This provides an explicit justification for estimate (4.4) given in Lemma 6.2.

Case 2. Let conditions (7.7), (7.14), (7.15), and (7.16) be again satisfied, and one of the inequalities (7.17) be violated. More accurately, instead of (7.17) we have

$$y_{i-1}^- - 2y_i^- + y_{i+1}^- \le y_i^- \quad (i = 2, \dots , N, \ i \ne k), \tag{7.24}$$

$$y_{k-1}^- - 2y_k^- + y_{k+1}^- > y_k^-, \tag{7.25}$$

where $2 \le k \le N$. Rough conditions sufficient for the fact that trajectory $x(\cdot)$ visits precisely two linearity domains also involve inequalities (7.49), (7.50), and (7.51) introduced later. (These inequalities require that y_k^+, the upper bound for variable y_k, is sufficiently large, and intervals $[s_{k-1}^-, s_{k-1}^+]$ and $[s_k^-, s_k^+]$ admissible for variables s_{k-1} and s_k are wide enough.)

As in Case 1, for t in a neighborhood of the origin we have inequalities (7.8) and (7.9) implying the explicit trajectory representation (7.12), (7.13) and formulas (7.20) through (7.23). Due to (7.25) the right-hand side in (7.21), where $i = k$, exceeds $y_k^-(1 - e^{-t/\alpha})$. Hence $-f_k(x(t))\alpha$ reaches y_k^-, i.e., $x(t)$ hits the boundary of the linearity domain L_1 at a finite time. Let ξ be the first hitting time defined as in Case 1. Arguing as in Case 1, we find that neither of the hitting conditions (7.18) with $i \ne k$ and (7.19) can hold. Therefore

$$-\frac{f_k(x(\xi))}{\alpha} = y_k^-, \tag{7.26}$$

$$-\frac{f_i(x(t))}{\alpha} < y_k^- \quad (t \in [0, \xi), \ i = 1, \dots , N + 1), \tag{7.27}$$

$$s_i^- < -\frac{g_i(x(t))}{\alpha} < s_i^+ \quad (t \in [0, \xi), \ i = 1, \dots , N). \tag{7.28}$$

Let us identify a linearity domain containing $x(t)$ after the hitting time ξ. According to (7.21), the left derivative of $f_k^*(\cdot) : t \mapsto f_k(x(t))$ at point ξ is given by $(y_{k-1}^- - 2y_k^- + y_{k-1}^-)e^{-\xi/\alpha}$. Assumption (7.25) shows that it is positive. Since $f_k(\cdot)$ (see (7.3)) and $x(\cdot)$ are continuously differentiable, $f_k^*(\cdot)$ is continuously differentiable as well, hence $\dot{f}_k^*(t)$ is positive for t in a right neighborhood of ξ. Due to (7.26), for these t we have

$$\frac{f_k(x(t))}{\alpha} > y_k^- . \tag{7.29}$$

On the other hand, (7.27) and (7.28) imply that for t in a right neighborhood of ξ,

$$-\frac{f_k(x(t))}{\alpha} < y_k^- \quad (i = 1, \dots , N + 1,\ i \neq k), \tag{7.30}$$

$$s_i^- < -\frac{g_i(x(t))}{\alpha} < s_i^+ \quad (i = 1, \dots , N). \tag{7.31}$$

Now (7.4) and (7.5) show that for t varying in a right neighborhood of ξ, point $x(t)$ moves within the linearity domain L_2, where the right-hand side of the ODE (7.6) is given by

$$v_i^{ps}(x) = y_i^- \quad (i = 1, \dots , N + 1,\ i \neq k), \tag{7.32}$$

$$v_k^{ps}(x) = \frac{f_k(x)}{\alpha}, \tag{7.33}$$

$$w_i^{ps}(x) = \frac{g_i(x)}{\alpha} \quad (i = 1, \dots , N). \tag{7.34}$$

Using formulas (7.3), we specify (7.6) as

$$\dot{y}_i(t) = y_i^- \quad (i = 1, \dots , N + 1,\ i \neq k), \tag{7.35}$$

$$\dot{y}_k(t) = \frac{y_{k-1}(t) - 2y_k(t) + y_{k+1}(t)}{\alpha} + \frac{s_{k-1}(t) + s_k(t)}{\alpha}, \tag{7.36}$$

$$\dot{s}_i(t) = -\frac{s_i(t)}{\alpha} + \frac{y_{i+1}(t) - y_i(t)}{\alpha} \quad (i = 1, \dots , N). \tag{7.37}$$

We see that the ODEs for $y_i(\cdot)$, where $i \neq k$, and $s_i(\cdot)$, where $i \neq k - 1, k$, do not differ from those in the linearity domain L_1 (see (7.10), (7.11)). Therefore for t close to ξ we have as in Case 1

$$y_i(t) = y_i^- t \quad (i = 1, \dots , N + 1,\ i \neq k), \tag{7.38}$$

$$\dot{s}_i(t) = -\frac{s_i(t)}{\alpha} + \frac{y_{i+1}(t) - y_i(t)}{\alpha} \quad (i = 1, \dots , N + 1,\ i \neq k - 1, k) \tag{7.39}$$

(see (7.12), (7.13)). Substitute (7.38) into (7.36), (7.37) for $i = k - 1, k$. We obtain a three-dimensional ODE for $y_k(\cdot)$, $s_{k-1}(\cdot)$, $s_k(\cdot)$. We can solve this ODE explicitly and find that the next representations hold in a right neighborhood of ξ:

$$\begin{aligned} y_k(t) = y_k^- &+ \frac{y_{k-1}^- + y_{k+1}^-}{3}(t - \xi) \\ &+ \frac{\alpha}{3}\left(y_k^- - \frac{y_{k-1}^- + y_{k+1}^-}{3}\right)(1 - e^{-3(t-\xi)/\alpha}), \end{aligned} \tag{7.40}$$

$$\begin{aligned} s_{k-1}(t) = {} & \left(\frac{y_{k-1}^- + y_{k+1}^-}{3} - y_{k-1}^-\right) t \\ & - \frac{\alpha}{6}\left(\frac{y_{k-1}^- + y_{k+1}^-}{3} - y_k^-\right)(1 - e^{-3(t-\xi)/\alpha}) \\ & + \alpha c_{k-1}\left(1 - e^{-(t-\xi)/\alpha}\right), \end{aligned} \tag{7.41}$$

$$\begin{aligned} s_k(t) = {} & \left(y_{k+1}^- - \frac{y_{k-1}^- + y_{k+1}^-}{3}\right) t \\ & + \frac{\alpha}{6}\left(\frac{y_{k-1}^- + y_{k+1}^-}{3} - y_k^-\right)(1 - e^{-3(t-\xi)/\alpha}) \\ & + \alpha c_k(1 - e^{-(t-\xi)/\alpha}). \end{aligned} \tag{7.42}$$

Here

$$c_{k-1} = (y_k^- - y_{k-1}^-)(1 - e^{-\xi/\alpha}) + \frac{y_k^- + y_{k-1}^- - y_{k+1}^-}{2},$$

$$c_k = (y_{k+1}^- - y_k^-)(1 - e^{-\xi/\alpha}) - \frac{y_k^- + y_{k-1}^- - y_{k+1}^-}{2}.$$

The differentiation yields that in a right neighborhood of ξ,

$$\dot y_k(t) = \phi_k(t), \quad \dot s_{k-1}(t) = \psi_{k-1}(t), \quad \dot s_k(t) = \psi_k(t),$$

where

$$\begin{aligned} \phi_k(t) = {} & \frac{y_{k-1}^- + y_{k+1}^-}{3} \\ & + \left(y_k^- - \frac{y_{k-1}^- + y_{k+1}^-}{3}\right) e^{-3(t-\xi)/\alpha}, \end{aligned} \tag{7.43}$$

$$\begin{aligned} \psi_{k-1}(t) = {} & \frac{y_{k-1}^- + y_{k+1}^-}{3} - y_{k-1}^- \\ & - \frac{1}{2}\left(y_k^- - \frac{y_{k-1}^- + y_{k+1}^-}{3}\right) e^{-3(t-\xi)/\alpha} \\ & + c_{k-1}e^{-(t-\xi)/\alpha}, \end{aligned} \tag{7.44}$$

$$\begin{aligned} \psi_k(t) = {} & y_{k+1}^- - \frac{y_{k-1}^- + y_{k+1}^-}{3} \\ & + \frac{1}{2}\left(y_k^- - \frac{y_{k-1}^- + y_{k+1}^-}{3}\right) e^{-3(t-\xi)/\alpha} \\ & + c_k e^{-(t-\xi)/\alpha}. \end{aligned} \tag{7.45}$$

If

$$y_k^- < \phi_k(t) < y_k^+, \tag{7.46}$$

$$s_{k-1}^- < \psi_{k-1}(t) < s_{k-1}^+, \tag{7.47}$$

$$s_k^- < \psi_k(t) < s_k^+ \tag{7.48}$$

for all $t > \xi$, then $x(t)$ lies in the linearity domain L_2 for all $t > \xi$, and consequently (7.40), (7.41), and (7.42) hold for all $t > \xi$. We can prove this by contradiction arguing as in Case 1. Let us specify the problem's parameters so as to guarantee (7.46), (7.47), (7.48) for all $t > \xi$. We start with (7.46). The second term in the right-hand side of (7.43) is negative for $t > \xi$ due to assumption (7.25). Hence the infimum and supremum of $\phi_k(t)$ in $[\xi, \infty)$ are achieved at ξ and infinity and are equal to y_k^- and $(y_{k-1}^- + y_{k+1}^-)/3$, respectively. Therefore

$$y_k^- < \phi_k(t) < \frac{y_{k-1}^- + y_{k+1}^-}{3}$$

for all $t > \xi$. We require

$$\frac{y_{k-1}^- + y_{k+1}^-}{3} \leq y_k^+ \tag{7.49}$$

and thus ensure (7.46) for all $t > \xi$.

The sum of the second and third terms on the right-hand side in (7.44) is estimated in absolute value by

$$\psi_{k-1}^* = \frac{1}{2}\left| y_k^- - \frac{y_{k-1}^- + y_{k+1}^-}{3} \right| + |y_k^- - y_{k-1}^-| + \frac{1}{2}|y_k^- + y_{k-1}^- - y_{k+1}^-|.$$

We guarantee (7.47) by assuming

$$s_{k-1}^- + \psi_{k-1}^* < \frac{y_{k-1}^- + y_{k+1}^-}{3} - y_{k-1}^- < s_{k-1}^+ - \psi_{k-1}^*. \tag{7.50}$$

Using (7.45), we similarly guarantee (7.48) by

$$s_k^- + \psi_k^* < y_{k+1}^- - \frac{y_{k-1}^- + y_{k+1}^-}{3} < s_k^+ - \psi_k^*; \tag{7.51}$$

here

$$\psi_k^* = \frac{1}{2}\left| y_k^- - \frac{y_{k-1}^- + y_{k+1}^-}{3} \right| + |y_{k+1}^- - y_k^-| + \frac{1}{2}|y_k^- + y_{k-1}^- - y_{k+1}^-|.$$

Thus assuming (7.49), (7.50), and (7.51), we have (7.40), (7.41), and (7.42) for all $t > \xi$. Hence

$$\lim_{t\to\infty} \frac{y_k(t)}{t} = \frac{y_{k-1}^- + y_{k+1}^-}{3},$$

$$\lim_{t\to\infty} \frac{s_{k-1}(t)}{t} = \frac{y_{k-1}^- + y_{k+1}^-}{3} - y_k^-,$$

$$\lim_{t\to\infty} \frac{s_k(t)}{t} = y_{k+1}^- - \frac{y_{k-1}^- + y_{k+1}^-}{3}.$$

The relations (7.35) and (7.37) show that $y_i(t)$ for $i \neq k$ and $s_i(t)$ for $i \neq k-1, k$ behave as in Case 1, i.e.,

$$\frac{y_i(t)}{t} = y_i^- \quad (i = 1, \ldots, N+1,\ i \neq k),$$

$$\lim_{t\to\infty} \frac{s_i(t)}{t} = y_{i+1}^- - y_i^- \quad (i = 1, \ldots, N,\ i \neq k-1, k).$$

Therefore

$$\lim_{t\to\infty} \frac{x(t)}{t} = x^0,$$

where

$$\begin{aligned} x^0 &= (y_1^0, \ldots, y_{N+1}^0, s_1^0, \ldots, s_N^0)^T, \\ y_i^0 &= y_i^- \quad (i = 1, \ldots, N+1,\ i \neq k), \\ y_k^0 &= \frac{y_{k-1}^- + y_{k+1}^-}{3}, \\ s_i^0 &= y_{i+1}^- - y_i^- \quad (i = 1, \ldots, N+1,\ i \neq k-1, k), \\ s_{k-1}^0 &= \frac{y_{k-1}^- + y_{k+1}^-}{3} - y_{k-1}^-, \\ s_k^0 &= y_{k+1}^- - \frac{y_{k-1}^- + y_{k+1}^-}{3}. \end{aligned}$$

By Theorem 6.1 x^0 is a (unique) solution of problem (2.1).

8. Discrete dynamics. In this section we consider discrete-time analogues of the shifting and penalized shifting trajectories. Here the feedback Q (see section 2) is not necessarily continuous.

Let (t_k) $(k = 1, 2, \ldots)$ be a sequence of positive instants increasing and convergent to infinity. We set

$$t_0 = 0, \quad \delta_k = t_{k+1} - t_k, \quad \tau_k = \sum_{i=0}^{k-1} \delta_i^2.$$

A *discrete-time trajectory under a feedback* U is defined to be a sequence $(x(t_k))$ in R^n satisfying $x(t_0) = 0$ and $x(x_{k+1}) \in x(t_k) + U(t_k, x(t_k))\delta_k$. Obviously, a discrete-time trajectory exists for every feedback. We can hardly expect a discrete-time trajectory to be solution tracking (see section 3) or solution abandoning (see section 4); each of these properties would require $x(t_1)/t_1$ to lie in the solution set or close to it, which is hardly possible.

We call a sequence $(x(t_k))$ in R^n *solution approaching* if $\mathrm{dist}(x(t_k), X^0) \to 0$ as $k \to \infty$. In what follows we assume that either

$$\lim_{k\to\infty} \frac{\tau_k}{t_k^2} = 0, \tag{8.1}$$

or

$$\lim_{k\to\infty} \frac{\tau_k}{t_k} = 0. \tag{8.2}$$

Since $t_k \to \infty$, (8.2) implies (8.1). Obviously (8.2) holds if the sequence (τ_k) is bounded, or, equivalently, the sum of the series $\delta_0^2 + \delta_1^2 + \cdots$ is finite, for instance, $\delta_k = 1/k$. If δ_k does not depend on k, $\delta_k = \delta$, then $\tau_k = \delta t_k$, and we have (8.1), while (8.2) is not valid.

THEOREM 8.1. *Let* (8.1) *be satisfied. Then every discrete-time trajectory under the shifting feedback is solution approaching.*

Proof. Let $(x(t_k))$ be a discrete-time trajectory under the shifting feedback U^s. We have $x(x_{k+1}) = x(t_k) + u(t_k)\delta$, where $u(t_k) \in U^s(t_k, x(t_k))$. By (2.4) $u(t_k) \in L^-(t, y_i)$, i.e., $(Fx(t_k) - t_k b)^T(Fu(t_k) - b) \le 0$. Hence

$$\begin{aligned}|Fx(t_{k+1}) - t_{k+1}b|^2 = {}&|Fx(t_k) - t_k b|^2\\ &+ 2(Fx(t_k) - t_k b)^T(Fu(t_k) - b)\delta_k + |Fu(t_k) - b|^2\delta_k^2\\ \le{}& |Fx(t_k) - t_k b|^2 + K\delta_k^2,\end{aligned}$$

where $K = \sup\{|Fu - b|^2 : u \in M\}$. So far as $Fx(t_0) - t_0 b = 0$, we obtain $|Fx(t_k) - t_k b|^2 \le K\tau_k$. Hence $|Fx(t_k/t_k) - b|^2 \le K\tau_k/t_k^2$. In view of (8.1), $\lim_{k\to\infty} |Fx(t_k/t_k) - b|^2 = 0$. Note that $x(t_k)/t_k \in M$. Therefore the sequence $(x(t_k)/t_k)$ converges to the feasible set of problem (2.1) (more accurately, the distance from $x(t_k)/t_k$ to the feasible set of problem (2.1) goes to zero as $k \to \infty$). By assumption $Q(t, x(t_k)) \cap X^0$ is nonempty. For $x_k^0 \in Q(t, x(t_k)) \cap X^0$ we have $(Fy_i - tb)^T(Fx_k^0 - b) = 0$. Hence $x_k^0 \in X^0 \cap L^-(t, x(t_k))$. Since $u(t_k) \in U^s(t_k, x(t_k))$, it holds that $J(u(t_k)) \le J(x_k^0) = J^0$. Using the convexity of J, we get

$$J\left(\frac{x(t_k)}{t_k}\right) \le \frac{1}{t_k}\sum_{i=0}^{k-1} J(u(t_k))\delta_k \le \frac{1}{t_k}\left(\sum_{i=0}^{k-1}\delta_k\right) J^0 = J^0.$$

Hence the sequence $(x(t_k)/t_k)$ (convergent to the feasible set of problem (2.1)) converges to the solution set X^0, and $\mathrm{dist}(x(t_k), X^0) \to 0$ as $k \to \infty$. □

THEOREM 8.2. *Let* (8.2) *be satisfied. Then every discrete-time trajectory under the penalized shifting feedback is solution approaching.*

Proof. Let $(x(t_k))$ be a discrete-time trajectory under the penalized shifting feedback U^{ps}, i.e., $x(x_{k+1}) = x(t_k) + u(t_k)\delta$, where $u(t_k) \in U^{ps}(t_k, x(t_k))$. Let

$$\lambda_k = |Fx(t_k) - t_k b|^2 + \alpha\sum_{i=0}^{k-1} J(u(t_k))\delta_k - \alpha t_k J^0.$$

Obviously,

$$\lambda_{k+1} - \lambda_k \le 2(Fx(t_k) - t_k b)^T(Fu(t_k) - b)\delta_k + K\delta_k^2 + \alpha J(u(t_k))\delta_k - \alpha J^0\delta_k,$$

where $K = \sup\{|Fu - b|^2 : u \in M\}$. For $x_k^0 \in Q(t, x(t_k)) \cap X^0$ we have $Fx_k^0 - b = 0$ and $J^0 = J(x^0(t))$. Therefore

$$\begin{aligned}\lambda_{k+1} - \lambda_k \le{} & [2(Fx(t_k) - t_k b)^T (Fu(t_k) - b) + \alpha J(u(t_k))]\delta_k \\ & - [2(Fx(t_k) - t_k b)^T (Fx_k^0 - b) + \alpha J^0]\delta_k + K\delta_k^2.\end{aligned}$$

The inclusion $u(t_k) \in U^{ps}(t_k, x(t_k))$ implies that the sum of the first two terms on the right is nonpositive. Thus $\lambda_{k+1} - \lambda_k \le K\delta_k^2$. Since $\lambda_0 = 0$, we have $\lambda_k \le K\tau_k$. Dividing by t_k^2, we get

$$\left| F\frac{x(t_k)}{t_k} - b \right|^2 \le -\alpha \frac{1}{t_k^2} \left(\sum_{i=0}^{k-1} J(u(t_k))\delta_k - t_k J^0 \right) + K\frac{\tau_k}{t_k^2}.$$

The right-hand side goes to zero as $k \to \infty$. That follows from (8.2) and the fact that J is bounded on M. Therefore the sequence $(x(t_k)/t_k)$ converges to the feasible set of problem (2.1). Now divide the inequality $\lambda_k \le K\tau_k$ by t_k. We obtain

$$\alpha \left(\frac{1}{t_k} \sum_{i=0}^{k-1} J(u(t_k))\delta_k - J^0 \right) \le K\frac{\tau_k}{t_k} - \frac{1}{t_k}|Fx(t_k) - b|^2 \le K\frac{\tau_k}{t_k}.$$

By (8.2) the right-hand side goes to zero as $k \to \infty$, and we have

$$\limsup_{k\to\infty} \frac{1}{t_k} \sum_{i=0}^{k-1} J(u(t_k))\delta_k \le J^0.$$

By the convexity of J

$$J\left(\frac{x(t_k)}{t_k} \right) \le \frac{1}{t_k} \sum_{i=0}^{k-1} J(u(t_k))\delta_k.$$

Therefore $\limsup_{k\to\infty} J(x(t_k)/t_k) \le J^0$. Since $(x(t_k)/t_k)$ converges to the feasible set of problem (2.1), it converges to the solution set X^0, and $\operatorname{dist}(x(t_k), X^0) \to 0$ as $k \to \infty$. □

REFERENCES

[1] K. J. Arrow, L. Hurwitz, and H. Uzava, *Studies in Linear and Nonlinear Programming*, Stanford University Press, Stanford, CA, 1958.

[2] K. A. Tanabe, *A geometric method in nonlinear programming*, J. Optim. Theory Appl., 30 (1980), pp. 181–210.

[3] A. S. Antipin, *Continuous and iterative processes with projection operators*, in Voprosy Kibernetiki. Vychislitilnye Voprosy Analisa Bolshikh System (Questions of Cibernetics. Computational Questions of the Analysis of Large Systems), Nauka, Moscow, 1989, pp. 1–43 (in Russian).

[4] Yu. G. Evtushenko and V. G. Zhadan, *Barrier-projection methods for solving nonlinear programming problems*, Zh. Vychisl. Mat. Mat. Fiz., 34 (1994), pp. 669–684 (in Russian). English translation: Comput. Math. Math. Phys., 34 (1994), pp. 579–590.

[5] W. I. Zangwill and C. B. Garcia, *Pathways to Solutions, Fixed Points and Equilibria*, Prentice–Hall, Englewood Cliffs, NJ, 1981.

[6] G. Sonnevend, *An "analytic center" for polyhedrons and new classes of global algorithms for linear (smooth convex) programming*, in Proc. 12th Conference on System Modelling and Optimization, Budapest, 1985, Lecture Notes in Control and Inform. Sci. 84, Springer-Verlag, Berlin, 1986, pp. 866–876.

[7] N. N. Krasovskii and A. I. Subbotin, *Game-Theoretical Control Problems*, Springer-Verlag, Berlin, 1988.

[8] Yu. S. Osipov and A. V. Kryazhimskii, *Inverse Problems for Ordinary Differential Equations: Dynamical Solutions*, Gordon and Breach, London, 1995.

[9] A. V. Kryazhimskii and Yu. S. Osipov, *To a regularization of a convex extremal problem with inaccurately given constraints. An application to an optimal control problem with state constraints*, in Some Methods of Positional and Program Control, Urals. Sci. Center, Sverdlovsk, 1987, pp. 34–54 (in Russian).

[10] A. V. Kryazhimskii, *Convex Optimization via Feedbacks*, Working Paper WP-94-109, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1994.

[11] Yu. M. Ermoliev, A. V. Kryahimskii, and A. Ruszczyński, *Constraint aggregation principle in convex optimization*, Math. Programming Ser B, 34 (1997), pp. 353–372.

[12] A. F. Filippov, *Differential equations with discontinuous right hand side*, Mat. Sb., 51 (1960), pp. 99–128 (in Russian).

[13] J.-P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

[14] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.

SIAM J. Control Optim.
Vol. 37, No. 1, pp. 303–318

SECOND-ORDER CONDITIONS FOR OPTIMIZATION PROBLEMS WITH CONSTRAINTS[*]

JEAN-PAUL PENOT[†]

Abstract. Using a projective approach, new necessary conditions and new sufficient conditions for optimization problems with explicit or implicit constraints are examined. They are compared to previous ones. A particular emphasis is given to mathematical programming problems with nonpolyhedral constraints. This case occurs in particular when the constraints are defined in functional spaces.

Key words. Lagrangian, mathematical programming, multiplier, optimality conditions, projective tangent set, second-order conditions

AMS subject classifications. 49K27, 90C30, 46A20, 46N10, 52A05, 52A40

PII. S0363012996311095

1. Introduction.

Devising efficient optimality conditions is an important objective when dealing with optimization problems. Therefore the literature on the subject is rich. (See [11], [15], [25], [29], [30], [41] for recent contributions.) Structured problems, such as mathematical programming problems, optimal control problems, continuous time problems, and semi-infinite programming problems, require a particular attention because the constraints are not necessarily defined by a finite number of scalar functions. This lack of polyhedrality causes a gap between necessary conditions and sufficient conditions, (see, for instance, [24], [28]). Moreover, the conditions cannot be given the simple and aesthetic form of the cases in which the constraints are polyhedral, as in [3], [4], [7], [17]–[19], and [23], for instance.

In [37] we reduced this gap to an acceptable extent: when the decision space is finite dimensional, the sufficient condition differs from the necessary condition by the replacement of an inequality by a strict inequality. As the unconstrained case shows, this difference is unavoidable. However, the second-order conditions of [37] are complex, and so are the conditions of [20], [27], and [31]. It is the purpose of the present work to present more handy conditions inspired by [11] and to compare them with recent proposals. It appears that the new conditions are not as selective as the previous ones: the sufficient (resp., the necessary) condition is a consequence of the sufficient (resp., necessary) condition of [37]. However, the new necessary condition is close to the sufficient condition, and such a fact is rather satisfactory.

For simplicity, we limit our study to the second-order case and we do not insist on the projective aspect of the tangent sets we deal with, which is just pointed out in section 2, although it is probably the main novelty here.

The optimality conditions are presented in section 3 along with a comparison with the results of [37]. Mathematical programming problems are considered in section 4. We devote section 5 to comparisons with recent works which came to our attention after the original version of the present paper was submitted. We are especially indebted to the referees for references [12] and [26]. We hope the clarifications we give

[*] Received by the editors October 21, 1996; accepted for publication (in revised form) December 29, 1997; published electronically November 17, 1998.
http://www.siam.org/journals/sicon/37-1/31109.html

[†] Université de Pau et des Pays de l'Adour, Laboratoire de Mathématiques appliquées, CNRS UPRES A 5033 64000 Pau, France (jean-paul.penot@univ-pau.fr).

will provide hints for obtaining concrete and convenient conditions in the specially structured cases mentioned above.

2. Projective tangent sets. In what follows, we denote by $\mathbb{P}$ (resp., $\mathbb{R}_+$) the set of positive (resp., nonnegative) real numbers. The closed ball with center x and radius r in a normed vector space (n.v.s.) X is denoted by $B(x,r)$. The closure of a subset F of X is denoted by cl$F$. Recall that the projective space $P\left(X\right)$ associated with a vector space X is the set of equivalence classes of pairs $(v,r)\in X\times\mathbb{R}_+$ for the relation

$$(v,r)\sim(v',r')\ \text{ if }\ (v',r')=(tv,tr)\ \text{ for some } t>0.$$

Obviously, $P\left(X\right)$ can be identified with the union

$$P\left(X\right)=X_1\cup X_0,$$

where X_1 (resp., X_0) is the image of $X\times\{1\}$ (resp., $X\times\{0\}$) under the canonical mapping $p:X\times\mathbb{R}_+\to P(X)$. We write $[v,r]$ to denote $p\left(v,r\right)$ and we call p the projective projection. If Y is another vector space and if $h:X\to Y$ is a positively homogeneous mapping, then h induces a mapping $h^P:P\left(X\right)\to P\left(Y\right)$ satisfying $h^P\left(p\left(x,r\right)\right)=p\left(h\left(x\right),r\right)$ for each $(x,r)\in X\times\mathbb{R}_+$, hence $h^P\left(X_1\right)\subset Y_1$, $h^P\left(X_0\right)\subset Y_0$, and if $h^P([x,1])=[y,1]$, then $h^P([x,0])=[y,0]$. Conversely, any mapping $\widehat{h}:P\left(X\right)\to P\left(Y\right)$ satisfying these conditions is the mapping h^P associated with some positively homogeneous map $h:X\to Y$.

Definition 1. *Given an integer $k\geq 2$, a subset F of an n.v.s. X, $x\in$ cl F, $v_1,\ldots,v_{k-1}\in X$, the projective tangent set of order k to F at $(x,v_1,\ldots v_{k-1})$ is the image $PT^k\left(F,x,v_1,\ldots,v_{k-1}\right)$ by the projective projection p of the set $\widehat{T}^k\left(F,x,v_1,\ldots,v_{k-1}\right)$ of pairs $(w,r)\in X\times\mathbb{R}_+$ such that there exist sequences $\left(t_n\right)$, $\left(r_n\right)$ in $\mathbb{P}$ with limits 0 and r, resp., $\left(w_n\right)\overset{\sigma}{\to}w$ (weak convergence) such that $\left(r_n^{-1}t_n\right)\to 0$ and*

$$x_n:=x+t_nv_1+\frac{t_n^2}{2}v_2+\cdots+\frac{t_n^{k-1}}{(k-1)!}v_{k-1}+\frac{t_n^k}{k!}\frac{w_n}{r_n}\in F$$

for each n.

The preceding definition has been inspired by a notion presented in [11]; it is closely related to two notions given in [26]. A precise comparison will be given in the last section of the paper. Several variants are possible. For instance, one can take strong convergence instead of weak convergence in what precedes, or weak* convergence if X is a dual space. One could also use nets (or, rather, bounded nets). Also for some purposes, it would be possible to replace the condition $\left(r_n^{-1}t_n\right)\to 0$ by the weaker condition $\left(r_n^{-1}t_nw_n\right)\to 0$. Clearly, by its very definition, the weak tangent set of order k to F at $(x,v_1,\ldots,v_{k-1})$ (also denoted by $F^k\left(x,v_1,\ldots,v_{k-1}\right)$),

$$T^k\left(F,x,v_1,\ldots,v_{k-1}\right)=\limsup_{t\downarrow 0}k!t^{-k}\left(F-x-tv_1\cdots-\frac{t^{k-1}}{(k-1)!}v_{k-1}\right)$$

coincides with the set $F_1^k\left(x,v_1,\ldots,v_{k-1}\right)$, where

$$F_r^k\left(x,v_1,\ldots,v_{k-1}\right):=\left\{w\in X:(w,r)\in\widehat{T}^k\left(F,x,v_1,\ldots,v_{k-1}\right)\right\}.$$

Here the limit sup is the sequential limit sup with respect to the weak topology.

It may be useful to split the set $PT^2(F,x,v)$ into two parts.. We observe that this set is the union of $p\left(F^2(x,v)\times\{1\}\right)$ and $p\left(F_0^2(x,v)\times\{0\}\right)$, where

$$F^2(x,v)=\left\{w\in X:\exists(t_n)\searrow 0,\exists(w_n)\xrightarrow{\sigma}w,\ x+t_nv+\frac{1}{2}t_n^2w_n\in F\ \ \forall n\in\mathbb{N}\right\}$$

is the familiar (weak upper) *second-order tangent set* to F at (x,v) and

$$F_0^2(x,v)=\Big\{w\in X:\exists(t_n)\downarrow 0,\exists(r_n)\downarrow 0,\exists(w_n)\xrightarrow{\sigma}w,\left(r_n^{-1}t_n\right)\to 0,$$
$$x+t_nv+\frac{1}{2}r_n^{-1}t_n^2w_n\in F\ \ \forall n\Big\}$$

is what will be called the *asymptotic second-order tangent cone* to F at (x,v).

Similar decompositions hold for higher-order projective tangent sets. For the sake of simplicity, in what follows we focus our attention on the second-order case only.

Although the second-order tangent set to a smooth subset may be empty, as the example of

$$F:=\left\{(r,s)\in\mathbb{R}^2:r^2=s^3\right\},\ x=(0,0),\ v:=(1,0)$$

shows, the following result asserts that the projective tangent set of order two in the reflexive case is always nonempty.

PROPOSITION 2.1. *Let $v\in F'(x):=T(F,x)$, where F is an arbitrary subset of the reflexive Banach space X and $x\in\operatorname{cl}F$. Then either $F^2(x,v)$ or $F_0^2(x,v)$ is nonempty.*

Proof. By assumption, there exists a sequence $(t_n)\searrow 0$ such that the sequence (s_n) given by $s_n:=t_n^{-1}d(x+t_nv,F)$ converges to 0. Since $0\in F^2(x,v)$ if $s_n=0$ for infinitely many n, we may assume $s_n>0$ and set $r_n=\frac{1}{2}s_n^{-1}t_n$, $w_n=s_n^{-1}t_n^{-1}(z_n-x-t_nv)$, where $z_n\in F$ is such that $\|x+t_nv-z_n\|\leq 2s_nt_n$. Then $(r_n^{-1}t_n)=(2s_n)\to 0$ and

$$x+t_nv+\frac{1}{2}t_n^2r_n^{-1}w_n=z_n\in F.$$

Taking a subsequence if necessary, we may suppose $(r_n)\to r$ for some $r\in[0,\infty]$ and (w_n) has a weak limit w in $2B_X$. If $r=\infty$, setting $w_n'=r_n^{-1}w_n$, we get $(w_n')\to 0$ and $0\in F^2(x,v)$ (strong). If $r\in\mathbb{P}$ the same choice of (w_n') shows that $r^{-1}w\in F^2(x,v)$ (weak). Finally, if $r=0$ we have $w\in F_0^2(x,v)$. □

EXAMPLE 2.1. *Suppose F is the graph of a twice differentiable mapping $g:U\to V$ in $X=U\times V$, where U and V are n.v.s. Then, for $(u_0,v_0)\in F$, $(u,v)\in F'(u_0,v_0)$, $(w,z)\in X$, $r>0$, one has*

$$((w,z),r)\in\widehat{T}^2F((u_0,v_0),(u,v))$$

iff $z=g'(u_0)w+rg''(u_0)uu$, as an easy calculation shows. Since any submanifold of a normed vector space can be represented locally as a graph, this example applies in a variety of situations.

The following proposition shows the concept of projective tangent set is invariant under C^k-diffeomorphisms and thus can be extended to subsets of C^k-manifolds. We take $k=2$ for simplicity.

PROPOSITION 2.2. *Let $g:X\to Y$ be a mapping of class C^2 on an open subset X_0 of X, let B be a subset of X, $x\in X_0\cap\operatorname{cl}B$, and let C be a subset of Y with $g(B)\subset C$. Then for each $v\in X,(w,r)\in\widehat{T}^2(B,x,v)$ one has*

$$(g'(x)w+rg''(x)vv,r)\in\widehat{T}^2(C,g(x),g'(x)v).$$

Proof. By definition there exist sequences $(t_n) \searrow 0$, $(w_n, r_n) \to (w, r)$ such that $\{r_n\} \subset \mathbb{P}$, $\left(r_n^{-1} t_n\right) \to 0$, and $x_n := x + t_n v + 2^{-1} r_n^{-1} t_n^2 w_n \in B$ for each n. Then $g\left(x_n\right) \in C$, and since $(t_n v_n) \to 0$ strongly if $v_n := v + 2^{-1} r_n^{-1} t_n w_n$, for some $(y_n) \to 0$, one has

$$g\left(x_n\right) = g\left(x\right) + t_n g'\left(x\right) v + \frac{1}{2} r_n^{-1} t_n^2 \left(g'\left(x\right) w_n + r_n g''\left(x\right) vv\right) + t_n^2 y_n$$

in view of Taylor's expansion. Setting $z_n := g'\left(x\right) w_n + r_n g''\left(x\right) vv + 2 r_n y_n$ and observing that $(z_n) \xrightarrow{\sigma} z := g'\left(x\right) w + r g''\left(x\right) vv$ and that $\left(r_n^{-1} t_n\right) \to 0$, the result follows. □

The following property will be useful. For $r = 1$ it corresponds to a property similar to the one observed in [34], [13, Proposition 3.1].

Proposition 2.3. *Let C be a convex subset of X, and let $x \in C$, $v \in T\left(C, x\right)$. Then, for any $z \in T(T(C, x), v)$, $(w, r) \in \widehat{T}^2\left(C, x, v\right)$, one has*

$$(w + z, r) \in \widehat{T}^2\left(C, x, v\right).$$

Proof. By definition, $w \in F_r^2(x, v) := \left\{w \in X : (w, r) \in \widehat{T}^2\left(C, x, v\right)\right\}$, so that there exist sequences $(r_n) \to r$, $(t_n) \to 0_+$, $(w_n) \to w$ such that $\left(r_n^{-1} t_n\right) \to 0$, $r_n > 0$, and

$$x_n = x + t_n v + \frac{1}{2} r_n^{-1} t_n^2 w_n \in C$$

for each n. For any $y \in C$ and any $p, q \geq 0$ we have

$$\left(1 - \frac{1}{2} p r_n^{-1} t_n\right) x_n + \frac{1}{2} p r_n^{-1} t_n (x + q t_n (y - x)) \in C;$$

as $\left(r_n^{-1} t_n\right) \to 0$ we obtain $w + p(q(y - x) - v) \in F_r(x, v)$. As this set is closed, and as $\mathbb{R}_+(C - x)$ is dense in $T(C, x)$, we also have $w + p(T(C, x) - v) \subset F_r(x, v)$. Since $T(C, x)$ is convex, $\mathbb{R}_+(T(C, x) - v)$ is dense in $T(T(C, x), v)$ and we get $w + T(T(C, x), v) \subset F_r(x, v)$. □

Among the variants of Definition 1, the following one seems to be noticeable. We will see in section 5 that this variant is closely related to Definition 2.2 of [26].

Definition 2. *The second-order projective incident set to a subset F of X at (x, v) with $x \in F, v \in T(F, x)$ is the image by the projective projection of the set $\widehat{T}^{ii}(F, x, v)$ of $(w, r) \in X \times \mathbb{R}_+$ such that for any sequences $(t_n) \to 0_+$, $(r_n) \to r$ with $r_n > 0$, $(r_n^{-1} t_n) \to 0$, there exists a sequence $(w_n) \to w$ such that $x + t_n v + 2^{-1} r_n^{-1} t_n^2 w_n \in F$ for each n.*

Given $r \in \mathbb{R}_+$ we denote by $\widehat{T}^{ii}(F, x, v, r)$ the set of $w \in X$ such that $(w, r) \in \widehat{T}^{ii}(F, x, v)$, and we use the similar notation $\widehat{T}^2(F, x, v, r)$ when $\widehat{T}^{ii}(F, x, v)$ is replaced with $\widehat{T}^2(F, x, v)$.

Part of the interest of this notion stems from the following property, the proof of which follows easily from the definition.

Proposition 2.4. (a) *If F is convex, then $\widehat{T}^{ii}(F, x, v, r)$ is convex for each (x, v, r);*

(b) *if F is convex, then $(1 - \lambda)\widehat{T}^{ii}(F, x, v, r) + \lambda \widehat{T}^2(F, x, v, r) \subset \widehat{T}^2(F, x, v, r)$ for any $\lambda \in [0, 1]$;*

(c) $\widehat{T}^{ii}(F \times G, (x, y), (u, v), r) = \widehat{T}^{ii}(F, x, u, r) \times \widehat{T}^{ii}(G, y, v, r)$;

(d) $\widehat{T}^{ii}(F, x, u, r) \times \widehat{T}^2(G, y, v, r) \subset \widehat{T}^2(F \times G, (x, y), (u, v), r) \subset \widehat{T}^2(F, x, u, r) \times \widehat{T}^2(G, y, v, r)$.

Sequential concepts as in [21], [22], [37], and [38] can be devised for similar aims.

3. Optimality conditions. The following necessary optimality condition justifies the introduction of the projective tangent set of order two. The proof we present has been devised independently of the one in [12, Theorem 2]; however, results of this kind had been announced earlier by A. Cambini at a lecture in Marseille (see [11] for a partial account and section 5 for a comparison).

PROPOSITION 3.1. *Suppose $f : X \to \mathbb{R}$ is twice differentiable at $x \in F$ and attains a (local) minimum on $F \subset X$ at x. Then*

$$f'(x)\, v \geq 0 \text{ for each } v \in F'(x) = T(F,x),$$

and whenever $v \in F'(x) \cap \ker f'(x)$, one has

$$f'(x)\, w + r f''(x)\, vv \geq 0 \text{ for each } (w,r) \in \widehat{T}^2(F,x,v).$$

Clearly this last condition can be formulated on the second projective tangent space $PT^2(F,x,v)$.

Proof. Since for $g(\cdot) := f(\cdot) - f(x)$ we have $g(F) \subset \mathbb{R}_+$, the result follows from Proposition 2.2 and from the fact that for $y = 0$, $v = 0$ one has

$$\widehat{T}^2(\mathbb{R}_+, y, v) = \{(z,r) \in \mathbb{R} \times \mathbb{R}_+ : z \geq 0\}. \qquad \square$$

EXAMPLE 3.1. *Let $F = \mathbb{R} \times \{0\} \cup \{0\} \times \mathbb{R}$ in $X = \mathbb{R}^2$. Then $F'(0) = F$, and, as for $x = 0, v \in F$, the set $\widehat{T}^2(F,x,v)$ contains $(w,1)$, with $w = 0$, a necessary optimality condition for f on F at 0 is $f'(0) = 0$, $f''(0)\, vv \geq 0$ for each $v \in F$.* $\square$

It may be useful to split the condition of Proposition 3.1 into two parts, using the decomposition of $PT^2(F,x,v)$ we described above.

COROLLARY 3.2. *If $f : X \to \mathbb{R}$ is twice differentiable at $x \in F$ and attains a local minimum on F at x, then $f'(x)\, v \geq 0$ for each $v \in F'(x)$, and when $v \in F'(x) \cap \ker f'(x)$, one has*

$$f'(x)\, w + f''(x)\, vv \geq 0 \text{ for each } w \in F^2(x,v),$$
$$f'(x)\, w \geq 0 \text{ for each } w \in F_0^2(x,v).$$

The first condition is well known but the second one is new.

EXAMPLE 3.2. *Let $F = \left\{(r,s) \in \mathbb{R}^2 : s = \mid r \mid^{\alpha}\right\}$, where $\alpha \in]1,2[$. Then for $x = (0,0)$, $v = (1,0)$, the set $F^2(x,v)$ is empty but $F_0^2(x,v)$ contains $w = (0,1)$. Thus a necessary condition for $(0,0)$ to be a minimizer of f on F is $f'(0,0) = 0$. Such a condition can also be obtained from* [37, Theorem 1.2] *via a computation similar to the one in* [37, Example 1.4].

EXAMPLE 3.3. *Given a subset F of the space X, $x \in F$, $v \in X \backslash \{0\}$, given $0 < p < q$, let us denote by $T^{q/p}(F,x,v)$ the set of vectors w, such that for some sequences $(s_n) \to 0_+$, $(w_n) \to w$ one has $x + s_n^p v + s_n^q w_n \in F$ for each n. Then if $q > 2p$, one has $0 \in F^2(x,v)$ whenever $T^{q/p}(F,x,v)$ is nonempty, while for $q = 2p$ one has $F^2(x,v) = T^{q/p}(F,x,v)$; for $q < 2p$ and for $w \in T^{q/p}(F,x,v)$ one has $(w,0) \in \widehat{T}^2(F,x,v)$, as one can see by taking $(t_n) := (s_n^p)$, $(r_n) := (s_n^{2p-q})$. In the last case, a necessary condition for f to attain a local minimum on F at x is $f'(x)w \geq 0$ whenever $f'(x)v = 0$ and $w \in T^{q/p}(F,x,v)$. The relationships with the higher-order optimality conditions of* [14] *and* [15] *will be considered elsewhere.* $\square$

The preceding examples prompt us to clarify the relationships between Corollary 3.2 (which is equivalent to Proposition 3.1) and [37, Theorem 1.2].

PROPOSITION 3.3. *The necessary optimality condition of* [37]*:*

$$\frac{1}{2} f''(x)\, vv + \liminf_{(t,u) \to (0,v),\ t>0,\ x+tu \in F} f'(x) t^{-1}(u - v) \geq 0 \quad \forall v \in F'(x) \cap \ker f'(x)$$

implies the necessary condition of Corollary 3.2.

Proof. Let $v \in F'(x) \cap \ker f'(x)$. The condition $f''(x)vv + f'(x)w \geq 0$ for each $w \in F^2(x,v)$ is a consequence of [37, Theorem 1.2] by [37, Corollary 1.3]. Let us derive the condition $f'(x)w \geq 0$ for $w \in F_0^2(x,v)$. Suppose on the contrary that $f'(x)w < 0$ for some $w \in F_0^2(x,u)$. Then there exist sequences $(t_n) \downarrow 0$, $(r_n) \downarrow 0$, $(w_n) \xrightarrow{\sigma} w$ such that $\left(r_n^{-1}t_n\right) \downarrow 0$, $x_n := x + t_n v + (2r_n)^{-1} t_n^2 w_n \in F$ for each n. Setting $v_n := v + (2r_n)^{-1} t_n w_n$ we see that $(v_n) \to v$, $x + t_n v_n = x_n \in F$ and $f'(x)t_n^{-1}(v_n - v) = (2r_n)^{-1} f'(x)(w_n) \to -\infty$, which is a contradiction with our assumption. □

Although the necessary condition of Proposition 3.1 is not as strong as [37, Theorem 1.2], one can still associate to it a sufficient condition of the same type (see also [11] and [12, Theorem 2] for a closely related result).

Proposition 3.4. *If X is finite dimensional, if f is twice differentiable at $x \in F$, and if the following conditions hold, then x is a local strict minimizer of f on F:*

(a) $f'(x)v \geq 0$ *for each* $v \in F'(x)$;

(b) *if* $v \in F'(x) \cap \ker f'(x)$, $v \neq 0$, *then* $f'(x)w + rf''(x)vv > 0$ *for each* $(w,r) \in \widehat{T}^2(F,x,v) \setminus \{(0,0)\}$.

Proof. Suppose on the contrary there exists a sequence (x_n) of $F \setminus \{x\}$ with limit x such that $f(x_n) \leq f(x)$ for each $n \in \mathbb{N}$. Let $t_n := \|x_n - x\|$, $v_n := t_n^{-1}(x_n - x)$. Taking a subsequence if necessary, we may suppose (v_n) has a limit v with norm 1. Let $s_n := \|v_n - v\|$. When $s_n = 0$ for infinitely many n we get $0 \in F^2(x,v)$ and $f'(x)v = 0$ (by (a) and the inequality $t_n^{-1}(f(x+t_nv_n) - f(x)) \leq 0$), and $f''(x)vv \leq 0$, a contradiction, as we can take $(w,r) = (0,1)$ in (b). Thus we may suppose $s_n > 0$ for each n and assume that the sequence (r_n) given by $r_n := (2s_n)^{-1} t_n$ has a limit r in $\mathbb{R}_+ \cup \{\infty\}$, and the sequence $(w_n) := \left(s_n^{-1}(v_n - v)\right)$ has a limit w with norm 1. Then $\left(r_n^{-1}t_n\right) = (2s_n) \to 0$ and

$$x + t_n v + \frac{1}{2} r_n^{-1} t_n^2 w_n = x_n \in F \qquad \forall n.$$

Thus, when r is finite, we have $(w,r) \in \widehat{T}^2(F,x,v)$, and, since $f(A) \subset f(x) - \mathbb{R}_+$ for $A = \{x_n : n \in \mathbb{N}\}$, we obtain $f'(x)v \leq 0$, hence $f'(x)v = 0$, and

$$f'(x)w + rf''(x)vv \leq 0$$

by a computation similar to the one in Proposition 2.2. This is a contradiction, as $w \neq 0$.

When $r = \infty$, setting $r' = 1$, $r'_n = 1$, $w'_n = r_n^{-1} w_n$, we observe that $(w'_n) \to 0$ and $x + t_n v + \frac{1}{2} t_n^2 w'_n = x_n \in F$ for each n, so that $0 \in F^2(x,v)$ or $(w',r') \in \widehat{T}^2(F,x,v)$, and we get, as above, $f''(x)vv \leq 0$, which is a contradiction, since $(0,r') \in \widehat{T}^2(F,x,v)$, $v \neq 0$. □

The preceding sufficient condition is in fact a consequence of the sufficient condition of [37, Theorem 1.7], as the following result shows.

Proposition 3.5. *Suppose X is finite dimensional. If f is twice differentiable at $x \in F$ and $v \in F'(x) \cap \ker f'(x)$, $v \neq 0$, the condition*

(1) $$f'(x)w + rf''(x)vv > 0 \textit{ for each } (w,r) \in \widehat{T}^2(F,x,v) \setminus \{(0,0)\}$$

implies the condition

(2) $$\frac{1}{2} f''(x)vv + \liminf_{(t,u) \to (0,v),\ t>0,\ x+tu \in F} f'(x)t^{-1}(u-v) > 0 \ \ \forall v \in F'(0) \cap \ker f'(x).$$

Proof. Suppose on the contrary that the first condition holds and there exist sequences $(t_n) \to 0_+$, $(v_n) \to v$ such that $x + t_n v_n \in F$ for each n and

$$\frac{1}{2} f''(x)\, vv + f'(x)\, t_n^{-1}(v_n - v) \to c \leq 0.$$

Let $s_n = \|v_n - v\|$. If $s_n = 0$ for infinitely many n, we have $w := 0 \in F^2(x,v)$ and $\frac{1}{2} f''(x)\, vv = c \leq 0$, so that for $r = \frac{1}{2}$, $w = 0$, we get a contradiction with (1).

Thus, we may assume $s_n > 0$ for each n and that $(r_n) := \left(2^{-1} s_n^{-1} t_n\right)$ has a limit r in $\mathbb{R}_+ \cup \{\infty\}$ and $(w_n) := \left(s_n^{-1}(v_n - v)\right)$ has a limit $w \neq 0$. If $r = \infty$, setting $w'_n = 2 s_n t_n^{-1} w_n$, we see that $(w'_n) \to 0$, $x_n := x + t_n v + \frac{1}{2} t_n^2 w'_n = x + t_n v_n \in F$ for each n, hence $w' := 0 \in F^2(x,v)$, and, as $t_n^{-1}(v_n - v) = \frac{1}{2} w'_n$, we get $f'(x)\, 0 + \frac{1}{2} f''(x)\, vv = c \leq 0$, a contradiction with $(0,1) \in \widehat{T}^2(F,x,v)$. If $r < \infty$ we have

$$x_n := x + t_n v_n = x + t_n v + \frac{1}{2} r_n^{-1} t_n^2 w_n \in F$$

and $\left(r_n^{-1} t_n\right) = (2 s_n) \to 0$, so that $(w,r) \in \widehat{T}^2(F,x,v)$. Since

$$f'(x)\, w = \lim s_n^{-1} t_n f'(x)\, t_n^{-1}(v_n - v) = 2r\left(c - \frac{1}{2} f''(x)\, vv\right)$$

we get

$$f'(x)\, w + r f''(x)\, vv = 2rc \leq 0,$$

a contradiction. □

However, the implication shown in the preceding condition can be partly reversed.

PROPOSITION 3.6. *Suppose that f is twice differentiable at $x \in F$. Then, for each $v \in (F'(x) \setminus \{0\}) \cap \ker f'(x)$, the condition*

$$\liminf_{(t,u) \to (0,v),\ t>0,\ x+tu \in F} f'(x) t^{-1}(u - v) + \frac{1}{2} f''(x)\, vv > 0 \tag{3}$$

implies the condition

(4) $f'(x)\, w + r f''(x)\, vv > 0$ *for each* $(w,r) \in \widehat{T}^2(F,x,v)$ *with* $r \neq 0$.

Proof. Let $(w,r) \in \widehat{T}^2(F,x,v)$ with $r > 0$: there exist positive sequences $(t_n) \to 0$, $(r_n) \to r$, and a sequence $(w_n) \xrightarrow{\sigma} w$ such that $\left(r_n^{-1} t_n\right) \to 0$ and

$$x_n := x + t_n v + \frac{1}{2} r_n^{-1} t_n^2 w_n \in F$$

for each n. Let $v_n := t_n^{-1}(x_n - x) = v + \frac{1}{2} r_n^{-1} t_n w_n$, so that $(v_n) \to v$, $x + t_n v_n = x_n \in F$. By assumption, there exists some $c > 0$ such that, for n large enough, one has

$$\frac{1}{2} f''(x)\, vv + f'(x)\, t_n^{-1}(v_n - v) \geq c,$$

hence

$$\frac{1}{2} r_n f''(x)\, vv + \frac{1}{2} f'(x)\, w_n > c r_n > \frac{1}{2} c r > 0$$

as $r > 0$, and the result follows by taking limits. □

4. Application to mathematical programming. Let us consider in this section the mathematical programming problem

$$(\mathcal{M}) \quad \text{minimize } f\left(x\right) : x \in F := g^{-1}\left(C\right),$$

where $f : X \to \mathbb{R}$, $g : X \to Z$ are twice differentiable mappings, C is a closed convex subset of Z, and X and Z are Banach spaces. Such a formulation encompasses problems in which equality and inequality constraints are present.

We will need a series of preliminary results of some independent interest. The first one gives a characterization of the projective tangent set of order two to the feasible set F. It uses a condition of metric regularity introduced in [36]. Here, for $z \in Z$ we set $d(z, C) = \inf_{c \in C} \|z - c\|$ to denote the distance function to C, and we adopt a similar notation for subsets of X.

Proposition 4.1. *Suppose the following directional metric regularity condition is satisfied for $x \in X$, $v \in X$:*

(DMR) *there exists $\mu > 0$, $\rho > 0$ such that for $t \in (0, \rho)$, $u \in B\left(v, \rho\right)$ one has*

$$d\left(x + tu,\, g^{-1}\left(C\right)\right) \leq \mu\, d\left(g\left(x + tu\right), C\right).$$

Then, for $F = g^{-1}\left(C\right)$, one has

$$(w, r) \in \widehat{T}^2\left(F, x, v\right) \Leftrightarrow \left(g'\left(x\right)w + rg''\left(x\right)vv, r\right) \in \widehat{T}^2\left(C, g\left(x\right), g'\left(x\right)v\right).$$

Proof. In view of Proposition 2.2 it suffices to prove that $(w, r) \in \widehat{T}^2\left(F, x, v\right)$ whenever $\left(g'\left(x\right)w + rg''\left(x\right)vv, r\right) \in \widehat{T}^2\left(C, g\left(x\right), g'\left(x\right)v\right)$. Let $(r_n) \to r$, $(t_n) \to 0_+$, $(z_n) \to z := g'\left(x\right)w + rg''\left(x\right)vv$ be such that $\left(r_n^{-1}t_n\right) \to 0$, $r_n > 0$ and

$$g\left(x\right) + t_n g'\left(x\right)v + \frac{1}{2}r_n^{-1}t_n^2 z_n \in C$$

for each n. For n large enough we have $t_n \in (0, \rho)$, $u_n := v + 2^{-1}r_n^{-1}t_n w \in B\left(v, \rho\right)$, so that

$$\begin{aligned} d\left(x + t_n u_n, F\right) &\leq \mu\ d\left(g\left(x + t_n u_n\right), C\right) \\ &\leq \mu\ \left\|g\left(x + t_n u_n\right) - g\left(x\right) - t_n g'\left(x\right)v - 2^{-1}r_n^{-1}t_n^2 z_n\right\| \\ &\leq \frac{1}{2}r_n^{-1}t_n^2 \mu \left\|g'\left(x\right)w + r_n g''\left(x\right)vv - z_n\right\| + o\left(t_n^2\right). \end{aligned}$$

Since $(z_n) \to z$ we can find $x_n \in F$ such that $\left(r_n t_n^{-2}\left\|x + t_n u_n - x_n\right\|\right) \to 0$. Defining w_n by $x_n := x + t_n v + 2^{-1}r_n^{-1}t_n^2 w_n$ we get $(w_n) \to w$, so that $w \in \widehat{T}^2\left(F, x, v\right)$. □

Let us observe that condition (DMR) is a consequence of the following metric regularity condition:

(MR) there exist $\mu > 0$, $\delta > 0$ such that for each $x' \in B\left(x, \delta\right)$ one has

$$d\left(x', g^{-1}\left(C\right)\right) \leq \mu\, d\left(g\left(x'\right), C\right).$$

This condition is of more common use than the directional metric regularity condition (DMR). In turn, condition (MR) has been shown to be a consequence of the classical Mangasarian–Fromovitz qualification [28], [18] and of its extension to the infinite dimensional case in [40], [32], [6], [16], and [42], which can be written

$$(\mathrm{R^r}) \qquad g'\left(x\right)\left(X\right) - \mathbb{R}_+\left(C - g\left(x\right)\right) = Z.$$

When the interior int C of C is nonempty, it has been shown in [32] that the radial tangent cone $T^r(C,x) := \mathbb{R}_+(C - g(x))$ in the preceding condition can be replaced by the usual tangent cone $T(C, g(x)) = \mathrm{cl}(T^r(C, g(x)))$:

$$\text{(R)} \qquad g'(x)(X) - T(C, g(x)) = Z.$$

However, in general, condition (R) is weaker than condition (R^r) and does not imply (MR). We will use a second-order qualification condition which generalizes the Ben-Tal qualification condition [2]:

$$\text{(TR)} \qquad g'(x)(X) - T(T(C, g(x)), g'(x)v) = Z,$$

in which v is a given vector of X; it is still weaker than (R).

We will also need the following duality result.

LEMMA 4.2. *Let P and Q be closed convex cones of the Banach spaces X and Z, resp., and let $A : X \to Z$, $c : X \to \mathbb{R}$ be linear and continuous and such that for some $m \in \mathbb{R}, b \in Z$*

$$c(x) \geq m \text{ for each } x \in P \cap A^{-1}(b + Q).$$

Then, if $A(P) - Q = Z$, there exists $y \in Q^0$ such that for each $x \in P$

$$c(x) + \langle y, Ax - b\rangle \geq m.$$

Since P is a cone, the conclusion can be written $0 \in c+y\circ A+P^0$ and $\langle y, -b\rangle \geq m$. When $P = X$, we have $c + y \circ A = 0$. Taking $m = 0$, $b = 0$ we get a Farkas lemma:

$$-c \in (A^{-1}(Q) \cap P)^0 \Rightarrow \exists y \in Q^0 : -(c + y \circ A) \in P^0.$$

In what follows we say that $v \in X$ is a *critical vector* at x if $f'(x)v = 0$, $g'(x)v \in T(C, g(x))$, and we write $v \in K(x)$.

THEOREM 4.3. *Let x be a (local) solution to problem $(\mathcal{P})$. Suppose conditions* (DMR) *and* (TR) *are satisfied at x. Then, for each critical vector $v \in K(x)$, $v \neq 0$ and each $(z, r) \in \widehat{T}^2(C, g(x), g'(x)v)$ there exists some $y \in N(T(C, g(x)), g'(x)v)$ such that*

$$f'(x) + y \circ g'(x) = 0,$$
$$r(f''(x)vv + \langle y, g''(x)vv\rangle) \geq \langle y, z\rangle.$$

Proof. Given $v \in K(x)\backslash\{0\}$, $(z, r) \in \widehat{T}^2(C, g(x), g'(x)v)$, for each $w \in X$ such that

$$g'(x)w + rg''(x)vv - z \in T(T(C, g(x)), g'(x)v),$$

Proposition 2.3 ensures that

$$(g'(x)w + rg''(x)vv, r) \in \widehat{T}^2(C, g(x), g'(x)v).$$

It follows from Proposition 4.1 that

$$(w, r) \in \widehat{T}^2(F, x, v).$$

Then, by Proposition 3.1, we have

$$f'(x)w \geq -rf''(x)vv.$$

Taking in Lemma 4.2, $A = g'(x)$, $b = z - rg''(x)vv$, $c = f'(x)$, $P = X$, $Q = T(T(C, g(x)), g'(x)v)$, $m = -rf''(x)vv$, and observing that $A(P) - Q = Z$ by condition (TR), we get some $y \in Q^0 = N(T(C, g(x)), g'(x)v)$ such that

$$f'(x) + y \circ g'(x) = 0,$$

$$\langle y, -z + rg''(x)vv\rangle \geq -rf''(x)vv.$$

Thus the result is established. □

Let us present a variant of the preceding necessary condition.

THEOREM 4.4. *Let x be a (local) solution to problem $(\mathcal{P})$. Suppose conditions* (DMR) *and* (TR) *are satisfied at x. Then for each non-null critical vector $v \in K(x)$ and each nonempty closed convex subcone $\widehat{Q}$ of $\widehat{T}^2(C, g(x), g'(x)v)$ not contained in $Z \times \{0\}$, there exists some $y \in N(T(C, g(x)), g'(x)v)$ such that*

$$f'(x) + y \circ g'(x) = 0,$$
$$\inf_{(z,r)\in\widehat{Q}} [r(f''(x)vv + \langle y, g''(x)vv\rangle) - \langle y, z\rangle] \geq 0.$$

Proof. Given $v \in K(x)\setminus\{0\}$, and a cone $\widehat{Q}$ as above, in view of Proposition 2.3, for each $(w, r) \in X \times \mathbb{R}_+$ such that

$$(g'(x)(w) + rg''(x)vv, r) \in \mathrm{cl}(\widehat{Q} + T \times \{0\})$$

with $T := T(T(C, g(x)), g'(x)v)$, we have

$$(g'(x)(w) + rg''(x)vv, r) \in \widehat{T}^2(C, g(x), g'(x)v)$$

since $\widehat{T}^2(C, g(x), g'(x)v)$ is closed. It follows from Proposition 4.1 that

$$(w, r) \in \widehat{T}^2(F, x, v).$$

And then, by Proposition 3.1,

$$f'(x)w + rf''(x)vv \geq 0.$$

Setting $P = X \times \mathbb{R}_{+,}$ $Q = \mathrm{cl}(\widehat{Q} + T \times \{0\})$, and defining A by

$$A(w, r) := (g'(x)w + rg''(x)vv, r)$$

so that $A(P) - Q = Z \times \mathbb{R}$, as is easily seen, it follows from the Farkas lemma recalled above that there exists $(y, -s) \in Q^0 = (\widehat{Q} + T \times \{0\})^0$ such that

$$f'(x)w + rf''(x)vv - rs + \langle y, g'(x)w + rg''(x)vv\rangle \geq 0$$

for each $(w, r) \in X \times \mathbb{R}_+$. It follows that $y \in T^0 := N(T(C, g(x)), g'(x)v)$ and that

$$f'(x) + y \circ g'(x) = 0,$$
$$r(f''(x)vv + \langle y, g''(x)vv\rangle) \geq rs.$$

Since $rs \geq \langle y, z\rangle$ for each $(z, r) \in Q$, the result follows. □

Since the preceding optimality condition has been derived from Proposition 3.1, and since that criterion is a consequence of the results of [37], one may guess that it is a consequence of the necessary condition of [37] for mathematical programming problems. This is the case. Given $v \in K(x)$ and a nonempty closed convex subcone $\widehat{Q}$

of $\widehat{T}^2\left(C, g\left(x\right), g'\left(x\right)v\right)$, let us consider two cases. When $\widehat{Q}$ is contained in $Z \times \{0\}$ the condition $-\langle y, z\rangle \geq 0$ for each $z \in \widehat{Q}$ is satisfied by any y in the set $M(x)$ of multipliers, as is easily seen. When $\widehat{Q} \cap Z \times \mathbb{P}$ is nonempty, taking $T = X \times (\widehat{Q} \cap Z \times \{1\})$ in [37, Corollary 3.6] we get some $y \in M(x)$ such that

$$f''\left(x\right)vv + \langle y, g''\left(x\right)vv\rangle \geq \langle y, z\rangle$$

for each z such that $(z, 1) \in \widehat{Q}$. Taking into account the remarks above and a homogeneity argument, the conclusion follows.

Now, let us turn to sufficient conditions.

THEOREM 4.5. *The following conditions ensure that an element x of F is a strict local minimizer:*

(a) *X is finite dimensional;*

(b) *the set $M\left(x\right) = \{y \in N\left(C, g\left(x\right)\right) : f'\left(x\right) + y \circ g'\left(x\right) = 0\}$ of multipliers at x is nonempty;*

(c) *for each $v \in F'\left(x\right) \setminus \{0\}$ with $f'\left(x\right)v = 0$ and each $(w, r) \in X \times \mathbb{R}_+ \setminus \{(0,0)\}$ such that $(z, r) := (g'\left(x\right)w + rg''\left(x\right)vv, r) \in \widehat{T}^2\left(C, g\left(x\right), g'\left(x\right)v\right)$ there exists $y \in M\left(x\right)$ such that*

$$r\left(f''\left(x\right)vv + \langle y, g''\left(x\right)vv\rangle\right) > \langle y, z\rangle.$$

Proof. The existence of a multiplier y ensures condition (a) of Proposition 3.4 since for any $v \in F'\left(x\right)$ we have $g'\left(x\right)v \in T\left(C, g\left(x\right)\right)$ and $y \in N\left(C, g\left(x\right)\right)$, hence $\langle y, g'\left(x\right)v\rangle \leq 0$ and $f'\left(x\right)v \geq 0$.

In order to check condition (b) of Proposition 3.4, let us consider $v \in F'\left(x\right) \cap \ker f'\left(x\right)$ with $v \neq 0$ and $(w, r) \in \widehat{T}^2\left(F, x, v\right)$ with $(w, r) \neq (0, 0)$. Then Proposition 2.2 ensures that $(z, r) \in \widehat{T}^2\left(C, g\left(x\right), g'\left(x\right)v\right)$ for $z = g'\left(x\right)w + rg''\left(x\right)vv$. Then, taking $y \in M(x)$ as in assumption (c) we get

$$f'\left(x\right)w + rf''\left(x\right)vv > -\langle y, g'\left(x\right)w\rangle + \langle y, z\rangle - r\langle y, g''\left(x\right)vv\rangle = 0,$$

and condition (b) is satisfied. □

5. Comparisons with other works. As mentioned above, the definition we gave for the second-order projective incident cone $\widehat{T}^{ii}(F, x, v)$ to a subset F of X at (x, v) seems to be closely related to Definition 2.2 of [26]: $(w, r) \in TC^{(2)}(F, x, v)$ iff there exist $\varepsilon > 0$ and $\alpha : [0, \varepsilon] \to X$ such that $\alpha(s) \to 0$ as $s \to 0$,

$$x + s\sqrt{r}v + s^2 w + s^2\alpha(s) \in F \qquad \forall s \in [0, \varepsilon].$$

In fact, supposing X is finite dimensional, so that the weak topology coincides with the strong topology, setting $t = s\sqrt{r}$ we see that for $r > 0$ $(w, r) \in TC^{(2)}(F, x, v)$ iff $(w, r) \in \widehat{T}^{ii}(F, x, v)$ iff $r^{-1}w \in T^{ii}(F, x, v) := \liminf_{t \to 0_+} 2t^{-2}(F - x - tv)$. However, $(w, 0) \in TC^{(2)}(F, x, v)$ iff $w \in T^i(F, x) := \liminf_{t \to 0_+} t^{-1}(F - x)$, the first-order incident tangent cone, and there is no relationship with the case $(w, 0) \in \widehat{T}^{ii}(F, x, v)$. Another definition is given in [26], in the style of the Dubovitskii–Milyutin work: $(w, r) \in FC^{(2)}(F, x, v)$ iff there exists $\varepsilon > 0$ such that

$$x + s\sqrt{r}v + s^2 B(w, \varepsilon) \subset F \qquad \forall s \in [0, \varepsilon].$$

Setting $G := X \backslash F$, we see that for $r > 0$ we have $(w, r) \in FC^{(2)}(G, x, v)$ iff $(w, r) \notin \widehat{T}^2(F, x, v)$. However, for $r = 0$ we have $(w, r) \in FC^{(2)}(G, x, v)$ iff $w \notin T(F, x)$ and there is no connection with $\widehat{T}^2(F, x, v)$.

As mentioned in the introduction, the definition of the projective tangent set we introduced above has been inspired by a notion given in a work of Cambini, Martein, and Komlosi [11] (or, rather, a talk around that paper). With a slight change of notation, their definition is as follows:

$$w \in TC''(F,x,v) \Leftrightarrow (\exists k \in \mathbb{R}_+ \cup \{\infty\}\ \exists (x_n) \in F^{\mathbb{N}}\ (x_n) \to x,$$
$$\exists (\alpha_n), (\beta_n) \to \infty :\ (\alpha_n \beta_n^{-1}) \to k,\ (\beta_n [\alpha_n(x_n - x) - v]) \to w).$$

Clearly, this set is a cone, as is $\widehat{T}^2(F,x,v)$. Given $(w,r) \in \widehat{T}^2(F,x,v)$ and setting $\alpha_n = t_n^{-1}$, $\beta_n = 2r_n t_n^{-1}$ one sees that $w \in TC''(F,x,v)$ so that, denoting by p_X the canonical projection of $X \times \mathbb{R}$ onto X, one has

$$p_X(\widehat{T}^2(F,x,v)) \subset TC''(F,x,v).$$

This inclusion is strict in general as a vector w such that for some sequences $(x_n) \in F^{\mathbb{N}}$, $(x_n) \to x$, $(\alpha_n), (\beta_n) \to \infty :\ (\alpha_n \beta_n^{-1}) \to 0$, $(\beta_n [\alpha_n(x_n - x) - v]) \to w$ does not belong to the left-hand side of the preceding relation. The necessary condition of [11] is thus potentially richer than the one of our Proposition 3.1. However, for a vector w as just described, the necessary condition of [11] reads as

$$f'(x)(2kw) + f''(x)(v,v) \geq 0$$

with $k = 0$ or $f''(x)(v,v) \geq 0$. Then, since $x_n = x + \alpha_n^{-1} v + \alpha_n^{-2}(\alpha_n \beta_n^{-1}) w_n$, we have $(\alpha_n \beta_n^{-1} w_n) \to 0$, so that $0 \in T^2(F,x,v)$ (see [12, Observation 7], in this connection) and the conclusion $f''(x)(v,v) \geq 0$ is contained in Proposition 3.1. For a similar reason, the assumptions of their sufficient condition are not more restrictive than the ones of our Proposition 3.4. We refer to [12] for a precise formulation of the optimality conditions of [11] and a number of observations about the second-order tangent sets described above. Among them is the following property [12, Observations 4 and 5]:

$$\widehat{T}^2(F,x,v) + \mathbb{R}T(F,x) \times \{0\} \subset \widehat{T}^2(F,x,v),$$

which is related to the inclusion

$$T^2(F,x,v) + \mathbb{R}T(F,x) \subset T^2(F,x,v)$$

contained in [13, Proposition 3.1].

Moreover, pursuing the line of thought of several papers [8], [9], [10], Cambini, Martein, and Komlosi introduce in [11] a new notion of second-order tangent set and use it for optimality conditions. When applied to mathematical programming problems, another main feature of the approach of [11] is the fact that it takes place in the image of the decision space X by the joint mapping $h := (f,g) : X \to V := \mathbb{R} \times Z$. In such a setting, X can be an arbitrary topological space, V can be an arbitrary normed vector space, and the following tools address local minimizers rather than global minimizers. Given $x \in X$, let us denote by $T(X,h,x)$ the set of $v \in V$ such that there exist sequences $(x_n) \to x$, $(t_n) \to 0_+$, $(v_n) \to v$ in X, $\mathbb{P}$, and V, resp., such that $v_n = t_n^{-1}(h(x_n) - h(x_0))$ for each n. Now, given $v \in T(X,h,x)$, let $T^2(X,h,x,v)$ be the set of limits w of sequences $(w_n) = 2t_n^{-2}(h(x_n) - h(x) - t_n v)$, where $(x_n) \to x$, $(t_n) \to 0_+$. Clearly,

$$T(X,h,x) \subset T(h(X),h(x)),\ T^2(X,h,x,v) \subset T^2(h(X),h(x),v),$$

and if X is a normed space,

$$h'(x)(X) \subset T(X,h,x),\ h'(x)(X) + h''(x)(v,v) \subset T^2(h(X),h(x),v).$$

However, these sets do not seem to be directly related to the projective tangent sets we defined, although they also give rise to optimality conditions in the form

$$T(X,h,x)\cap((-\mathbb{P})\times \mathrm{int}C)=\emptyset,$$
$$T^2(X,h,x,v)\cap((-\mathbb{P})\times \mathrm{int}C)=\emptyset \quad \forall v\in T(X,h,x)\cap Fr((-\mathbb{P})\times \mathrm{int}C).$$

On the other hand, conditions in terms of multipliers can be deduced from such relations and from the use of the set A_2 of $v\in T(X,h,x)\backslash\{0\}$ such that there exist $t>0$ and sequences $(x_n)\to x$, $(t_n)\to 0_+$, $(v_n)\to v$ in X, $\mathbb{P}$, and V, resp., such that $v_n=t_n^{-1}(h(x_n)-h(x_0))$, $t_n=t\|h(x_n)-h(x)\|$, $\|x_n-x\|^{-2}(h(x_n)-h(x))\to 0$. Such a set seems to be more closely related to our projective tangent sets.

Now let us turn to a recent contribution of Bonnans, Cominetti, and Shapiro [5] using a notion of approximation to devise a sufficient optimality condition which we intend to compare with the one in [37]. We recall them briefly. The condition in [37] relies on the notion of *compound tangent set* to $E:=(-\mathbb{R}_+)\times C$ (we suppose $f(x)=0$ for simplicity). Given $u\in X$ one denotes by

$$S_u:=\limsup_{(t,u')\to(0_+,u)} 2t^{-2}(E-h(x)-th'(x)u')$$

the set formed with limits of sequences (w_n) such that there exist sequences $(t_n)\to 0_+$, $(u_n)\to u$ in $\mathbb{P}$ and X, resp., with $h(x)+t_nh'(x)u_n+\frac{1}{2}t_n^2w_n\in E$ for each n. Then one can give a sufficient condition in order that x be an *essential local minimizer of second order* for problem $(\mathcal{M})$ in the following sense, which differs slightly from the one in [1], [5], [35], [39], and [38]: there exists $\alpha>0,\beta>0,\gamma>0$ such that

$$f(u)\geq f(x)+\alpha\|u-x\|^2 \text{ for any } u\in B(x,\beta) \text{ such that } d(g(u),C)\leq\gamma\|u-x\|^2.$$

We make use of the set $J(x)$ of F. John's multipliers at x for problem $(\mathcal{M})$, i.e., the set of $(t,y)\in\mathbb{R}_+\times N(C,g(x))$ such that

$$tf'(x)+y\circ g'(x)=0$$

and of the set of *subcritical directions*

$$K^{\leq}(x):=\{u\in X: f'(x)u\leq 0,\ g'(x)u\in T(C,g(x))\}.$$

This set obviously coincides with the set of critical directions $K(x)$ whenever the set of multipliers $M(x)=\{y:(1,y)\in J(x)\}$ is nonempty.

PROPOSITION 5.1. *The following conditions ensure that an element x of F is an essential local minimizer of second order:*

(a) *X is finite dimensional;*

(b) *the set $J(x)$ of John's multipliers at x is nonempty;*

(c) *for each $u\in K^{\leq}(x)\setminus\{0\}$ and each $(r,z)\in S_u$ there exists a multiplier $(t,y)\in J(x)$ such that*

$$tf''(x)uu+\langle y,g''(x)uu\rangle > rt+\langle y,z\rangle. \tag{5}$$

Proof. Suppose on the contrary that there exist a sequence (x_n) of X and a sequence $(\varepsilon_n)\to 0_+$ such that $0<t_n:=\|x_n-x\|<\varepsilon_n, d(g(x_n),C)\leq\varepsilon_nt_n^2$, $f(x_n)<f(x)+t_n^2\varepsilon_n$ for each n. Without loss of generality we may assume that $(t_n^{-1}(x_n-x))$ converges to some u in X. It is easy to see that $u\in K^{\leq}(x)\setminus\{0\}$ and that $(r,z):=(f''(x)uu,g''(x)uu)\in S_u$. Thus we get a contradiction with (c). □

Now, in order to present the result of [5] let us introduce the following concepts in which the Pompeiu–Hausdorff *excess* of a subset C over another subset D of a metric space is given by

$$e(C,D) := \sup_{c\in C} d(c,D).$$

DEFINITION 3. *A subset A of a metric space is an outer (or upper) hemi-limit of a family $(A_w)_{w\in W}$ of subsets of X parametrized by a subset $W$of a topological space P as $w \to w_0 \in \mathrm{cl}W$, $w \in W$ if $e(A_w, A) \to 0$ as $w \to w_0$ in W.*

Such a limit is not unique: if A' contains A, then A' is again an outer hemi-limit of (A_w). Moreover, any closed outer hemi-limit of (A_w) contains $\limsup_{w\to w_0} A_w$, as is easily seen. The concept introduced in [5] can be reformulated as follows (in the case $d = Mu$, which is of interest to us).

DEFINITION 4. *Given n.v.s. X and Z, a subset C of Z, a continuous linear mapping $M : X \to Z$, $u \in X$, $z \in C$, a subset A of Z is said to be an upper (second-order) approximation to C at z with respect to M, z, u if it is an outer hemi-limit of the family $A_{t,u'} := 2t^{-2}(C - z - tMu')$ as $(t, u') \to (0, u)$ in $\mathbb{P} \times X$.*

A simpler notion can be introduced.

DEFINITION 5. *Given a subset C of an n.v.s. Z, $z \in C$, $v \in T(C, z)$, a subset A of Z is said to be an outer (second-order) approximation to C at z in the direction v if it is an outer hemi-limit of the family $A_t := 2t^{-2}(C - z - tv)$ as $t \to 0_+$.*

This definition is less demanding than the preceding one: if A is an upper approximation to C at z with respect to M, z, u, and if $v := Au$, then A is an outer approximation to C at z in the direction v.

EXAMPLE. *For any convex subset C of Z and any $z \in C$, $v \in T(C, z)$ the cone $T(T(C, z), v)$ is an outer approximation to C at z in the direction v. In fact, for any $t > 0$, $c \in C$, setting $w := 2t^{-2}(c - z - tv)$, $v' := v + (t/2)w = t^{-1}(c - z) \in T(C, z)$, one has $w = 2t^{-1}(v' - v) \in T(T(C, z), v)$.* □

The main result of [5] states that if x is feasible for problem $(\mathcal{M})$, if for each $u \in K^{\leq}(x)$ there exists an upper approximation A to C with respect to $M := g'(x)$, $z := g(x)$, u, and if there exists $(t, y) \in J(x)$ such that

$$tf''(x)\,uu + \langle y, g''(x)\,uu\rangle > \sigma(y, A) := \sup_{a\in A}\langle y, a\rangle, \tag{6}$$

then x is a strict locally optimal solution of $(\mathcal{M})$. In fact this result can be extended to the case when A is just an outer approximation to C at z in the direction $v := g'(x)u$, and, moreover, it is a simple consequence of the preceding proposition in view of the following lemma.

LEMMA 5.2. *If condition* (6) *holds for some outer approximation to C at $g(x)$ in the direction $v := g'(x)u$, then condition* (5) *holds.*

Proof. It suffices to prove that for any $u \in K^{\leq}(x)$, any $(r, z) \in S_u$, any $(t, y) \in J(x)$, and any outer approximation A to C at $g(x)$ in the direction $v := g'(x)u$, one has

$$\sigma(y, A) \geq \langle y, z\rangle + rt.$$

Now, since $(r, z) \in S_u$ there exist sequences $(t_n) \to 0_+$, $(u_n) \to u$, $(z_n) \to z$, $(r_n) \to r$ such that

$$c_n := g(x) + t_n g'(x)u_n + \frac{1}{2}t_n^2 z_n \in C,$$
$$f'(x)u_n + \frac{1}{2}t_n r_n \leq 0.$$

Let $w_n := 2t_n^{-1}(u_n - u)$, and let $q_n := g'(x)w_n + z_n$. Since A is an outer approximation to C at $g(x)$ in the direction $v := g'(x)u$, and since $q_n = 2t_n^{-2}(c_n - g(x) - t_n g'(x)u)$, there exists $a_n \in A$ such that $\varepsilon_n := \|q_n - a_n\| \to 0$. Then, using the definitions of $J(x)$ and $K^{\leq}(x)$, we get

$$\begin{aligned}\langle y, z_n\rangle + tr_n &= \langle y, q_n\rangle - \langle y, g'(x)w_n\rangle + tr_n\\ &= \langle y, q_n\rangle + tf'(x)w_n + tr_n\\ &= \langle y, q_n\rangle + 2tt_n^{-1}\left(f'(x)u_n + \frac{1}{2}t_n r_n\right)\\ &\leq \langle y, a_n\rangle + \varepsilon_n\|y\|.\end{aligned}$$

Therefore, taking limits, we get

$$\langle y, z\rangle + rt \leq \sigma(y, A).$$

COROLLARY 5.3. *Suppose that for an element x of F the conditions* (a) *and* (b) *of the preceding proposition hold while condition* (c) *is replaced with the following condition* (c′)*. Then x is an essential local minimizer of second order:*

(c′) *for each $u \in K^{\leq}(x) \setminus \{0\}$ there exist an outer approximation A of C at $g(x)$ in the direction $g'(x)u$ and a multiplier $(t, y) \in J(x)$ such that*

$$tf''(x)\,uu + \langle y, g''(x)\,uu\rangle > \sup_{a\in A}\langle y, a\rangle. \tag{7}$$

REFERENCES

[1] A. Auslender, *Stability in mathematical programming with nondifferentiable data*, SIAM J. Control Optim., 22 (1984), pp. 239–254.

[2] A. Ben-Tal, *Second order and related extremality conditions in nonlinear programming*, J. Optim. Theory Appl., 31 (1980), pp. 143–165.

[3] A. Ben Tal and J. Zowe, *A unified theory of first and second-order conditions for extremum problems in topological vector spaces*, Math. Programming Study, 19 (1982), pp. 39–76.

[4] A. Ben Tal and J. Zowe, *Necessary and sufficient optimality conditions for a class of nonsmooth minimization problems*, Math. Programming, 24 (1982), pp. 70–91.

[5] J.F. Bonnans, R. Cominetti, and A. Shapiro, *Second order necessary conditions and sufficient optimality conditions under abstract constraints*, preprint, 1996.

[6] J. Borwein, *Stability and regular points of inequality systems*, J. Optim. Theory Appl., 48 (1986), pp. 9–52.

[7] J. Burke, *Second order necessary and sufficient conditions for convex composite NDO*, Math. Programming, 38 (1987), pp. 287–302.

[8] A. Cambini and L. Martein, *Second order necessary optimality conditions in the image space: Preliminary results*, in Scalar and Vector Optimization in Economic and Financial Problems, E. Castagnoli and G. Giorgi, eds., 1995, pp. 27–38.

[9] A. Cambini, L. Martein, and R. Cambini, *A new approach to second order optimality conditions in vector optimization*, Technical report 103, Department of Statistics and Applied Math., Univ. of Pisa, 1996.

[10] R. Cambini, *Second order optimality conditions in the image space*, Technical report 99, Department of Statistics and Applied Math., Univ. of Pisa, 1996.

[11] A. Cambini, L. Martein, and S. Komlosi, *Recent developments in second order necessary optimality conditions*, Generalized Convexity, Generalized Monoticity, J.-P. Crouzeix et al., eds., Kluwer, Amsterdam, 1998, pp. 347–356.

[12] A. Cambini, L. Martein, and M. Vlach, *Second order tangent sets and optimality conditions*, preprint, Japan Advanced Study of Science and Technology, Hokuriku, Japan, 1997.

[13] R. Cominetti, *Metric regularity, tangent sets and second-order optimality conditions*, Applied Math. Optim., 21 (1990), pp. 265–287.

[14] J.-P. Dedieu, *Third and fourth-order optimality conditions in optimization*, Optimization, 33 (1995), pp. 97–105.

[15] J.-P. Dedieu and R. Janin, *A propos des conditions d'optimalité d'ordre trois et quatre pour une fonction de plusieurs variables*, preprint, Univ. Poitiers and Toulouse, 1995.

[16] A. Dmitruk, A. Milyutin, and N. Osmolovski, *Lyusternik's theorem and the theory of extrema*, Russian Math. Surveys, 35 (1980), pp. 11–51.

[17] B. Gollan, *Higher order necessary conditions for an abstract optimization problem*, Math. Programming Study, 14 (1981), pp. 69–76.

[18] S.P. Han and O.L. Mangasarian, *Exact penalty functions in nonlinear programming*, Math. Programming, 17 (1979), pp. 251–269.

[19] K.H. Hoffmann and H.J. Kornstaedt, *Higher order necessary conditions in abstract mathematical programming*, J. Optim. Theory Appl., 26 (1978), pp. 533–569.

[20] A.D. Ioffe, *Necessary and sufficient conditions for a local minimum* 3: *Second-order conditions and augmented duality*, SIAM J. Control Optim., 17 (1979), pp. 266–288.

[21] A.D. Ioffe, *On some recent developments in the theory of second-order optimality conditions*, in Optimization, Proc. Conf. Varetz, 1988, S. Dolecki, ed., Lecture Notes in Math. 1405, Springer-Verlag, Berlin, 1989, pp. 55–68.

[22] A.D. Ioffe, *Variational analysis of a composite function: A formula for the lower second-order epi-derivative*, J. Math. Anal. Appl., 160 (1991), pp. 379–405.

[23] R. Janin, *Conditions d'optimalité d'ordre supérieur en programmation mathématique*, Univ. of Poitiers, 1995, preprint.

[24] H. Kawasaki, *An envelop-like effect of infinitely many inequality constraints on second-order necessary conditions for minimization problems*, Math. Programming, 41 (1988), pp. 73–96.

[25] H. Kawasaki, *Second-order necessary and sufficient optimality conditions for minimizing a sup type function*, Appl. Math. Optim., 26 (1992), pp. 195–220.

[26] U. Ledzewicz and H. Schaettler, *Second-order conditions for extremum problems with nonregular equality constraints*, J. Optim. Theory Appl., 86 (1995), pp. 113–144.

[27] E.S. Levitin, A.A. Milyutin, and N.P. Osmolovskii, *Higher order conditions for a local minimum in problems with constraints*, Uspehi Math. Nauk, 33 (1978), pp. 85–148.

[28] O.L. Mangasarian and S. Fromovitz, *The Fritz-John necessary optimality condition in the presence of equality and inequality constraints*, J. Math. Anal. Appl., 7 (1967), pp. 37–47.

[29] Y. Maruyama, *Second-order necessary conditions for nonlinear optimization problems in Banach spaces and their application to an optimal control problem*, Math. Oper. Res., 15 (1990), pp. 467–482.

[30] H. Maurer, *First and second-order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Stud., 14 (1981), pp. 163–177.

[31] Z. Páles and V.M. Zeidan, *Nonsmooth optimum problems with constraints*, SIAM J. Control Optim., 32 (1994), pp. 1476–1502.

[32] J.-P. Penot, *On regularity conditions in mathematical programming*, Math. Programming Stud., 19 (1982), pp. 167–199.

[33] J.-P. Penot, *A view of second-order extremality conditions*, Third Franco-German Conference in Optimization, C. Lemaréchal, ed., INRIA, Le Chesnay, France, 1984, pp. 34–39.

[34] J.-P. Penot, *A Geometric Approach to Higher Order Necessary Conditions*, manuscript, Univ. of Pau, 1984.

[35] J.-P. Penot, *Generalized higher order derivatives and higher order optimality conditions*, preprint, Univ. of Santiago, 1984.

[36] J.-P. Penot, *Differentiability of relations and differential stability of perturbed optimization problems*, SIAM J. Control Optim., 22 (1984), pp. 529–551.

[37] J.-P. Penot, *Optimality conditions in mathematical programming and composite optimization*, Math. Programming, 67 (1994), pp. 225–245.

[38] J.-P. Penot, *Sequential derivatives and composite optimization*, Rev. Roumaine Math. Pures Appl., 40 (1995), pp. 501–519.

[39] J.-P. Penot, *Central and peripheral results in the study of marginal and performance functions*, in Mathematical Programming with Data Perturbations, A. Fiacco, ed., Marcel Dekker, New York, 1997, pp. 305–337.

[40] S.M. Robinson, *Stability theory for systems of inequalities, Part* II: *Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.

[41] M. Stuniarski, *Second-order necessary conditions for optimality in nonsmooth nonlinear programming*, J. Math. Anal. Appl., 154 (1991), pp. 303–317.

[42] J. Zowe and S. Kurcyusz, *Regularity and stability for the mathematical programming problem in Banach space*, Appl. Math. Optim., 5 (1979), pp. 49–62.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 1, pp. 319–329

UNIFORM STABILIZATION OF THE DYNAMIC ELASTICA BY BOUNDARY FEEDBACK*

KAZUHO ITO†

Abstract. It is proved that the dynamic elastica, which is a nonlinear model describing the large deflecting motion of an inextensible beam, can be uniformly exponentially stabilized by boundary velocity feedback.

Key words. elastica, uniform stabilization, boundary control, exponential decay

AMS subject classifications. 93D15, 35B37, 73K05

PII. S0363012997322352

1. Introduction.

We consider the large, planar motion of an inextensible elastic beam described by the following model, called the dynamic elastica:

$$\rho u_{tt} = \lambda_x, \tag{1.1}$$

$$\rho w_{tt} = \nu_x, \tag{1.2}$$

$$\lambda \sin\varphi - \nu\cos\varphi + I_\rho \varphi_{tt} - EI\varphi_{xx} = 0, \tag{1.3}$$

$$1 + u_x = \cos\varphi, \quad w_x = \sin\varphi. \tag{1.4}$$

The unknowns are u, w, φ, λ, and ν, which are functions of two variables x and t, $0 \le x \le L$, $t \ge 0$. In this model, the centerline (i.e., the locus of the centroid of the cross section) of the beam is assumed to lie in a fixed plane with the usual Cartesian coordinate system O-XY, and, in the reference state of the beam, to occupy the interval $0 \le X \le L$ of the X-axis. The unknowns $u(x,t)$ and $w(x,t)$ denote the displacements, at time t, in the X- and Y-direction, respectively, of the particle which occupies position $(x,0)$, $0 \le x \le L$, in the reference state; thus, the centerline at time t is described by the curve $x \mapsto (x + u(x,t), w(x,t))$. The unknown $\varphi(x,t)$ denotes the angle between the tangent $(1 + u_x(x,t), w_x(x,t))$ of the centerline and the X-axis. This, together with the inextensibility condition $\sqrt{(1+u_x)^2 + w_x^2} = 1$, implies (1.4). The unknowns $\lambda(x,t)$ and $\nu(x,t)$ denote the components in the X- and Y-direction, respectively, of the force acting on $\Gamma(x)$, where $\Gamma(x)$ is the cross section that is at $X = x$ when the beam is in the reference state. In the motion equations (1.1)–(1.3), the coefficients ρ, I_ρ, and EI are the mass per unit length, the mass moment of inertia of the cross section, and the bending stiffness of the beam. We assume that these coefficients are positive constants. For more detailed explanation of (1.1)–(1.3), see, e.g., [2], [1].

The boundary conditions we consider at $x = 0$ are

$$u(0,t) = w(0,t) = 0, \tag{1.5}$$

$$\varphi(0,t) = 0, \tag{1.6}$$

*Received by the editors June 4, 1997; accepted for publication (in revised form) February 27, 1998; published electronically November 17, 1998.
http://www.siam.org/journals/sicon/x-x/32235.html

†Department of Electrical Engineering and Computer Science, Yamanashi University, 4-3-11 Takeda, Kofu 400-8511, Japan (kazuho@pine.ese.yamanashi.ac.jp).

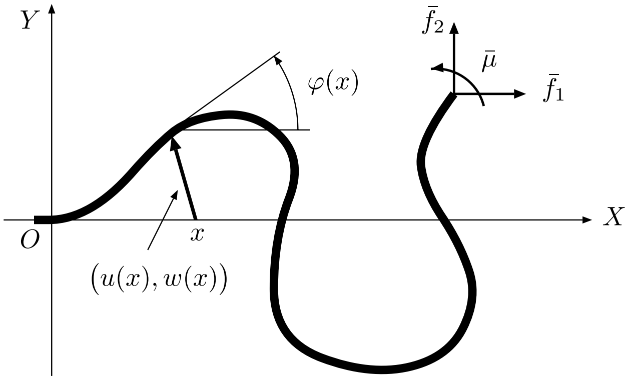


FIG. 1.1. *Elastica (the dependence of the functions upon t is suppressed).*

which represent that the beam is clamped at the end having the cross section $\Gamma(0)$, and the boundary conditions at $x = L$ are

$$\lambda(L,t) = \bar{f}_1(t), \quad \nu(L,t) = \bar{f}_2(t), \tag{1.7}$$

$$EI\varphi_x(L,t) = \bar{\mu}(t), \tag{1.8}$$

which represent that, at the cross section $\Gamma(L)$, the beam is subjected to external forces $\bar{f}_1$ and $\bar{f}_2$ in the X- and Y-direction, respectively, and to an external moment $\bar{\mu}$ (see Fig. 1.1).

For the model above, the strain energy of the beam is given by

$$U(t;\varphi) = \tfrac{1}{2}\int_0^L EI\varphi_x(x,t)^2\,dx, \tag{1.9}$$

and the kinetic energy is given by

$$K(t;\varphi) = \tfrac{1}{2}\int_0^L \left\{\rho u_t(x,t)^2 + \rho w_t(x,t)^2 + I_\rho \varphi_t(x,t)^2\right\} dx \tag{1.10}$$

(note that u and w, and thus K, are determined by φ through (1.4) and (1.5)). Indeed, the motion equations (1.1)–(1.3) and the natural boundary conditions (1.7)–(1.8) can be derived from Hamilton's principle with the Lagrangian

$$\mathcal{L}(\varphi) = \int_{t_0}^{t_1} \left\{U(t;\varphi) - K(t;\varphi) - (\bar{f}_1(t)u(L,t) + \bar{f}_2(t)w(L,t) + \bar{\mu}(t)\varphi(L,t))\right\} dt \tag{1.11}$$

and with the geometrical constraints (1.4), (1.5)–(1.6).

The purpose of this paper is to investigate uniform stabilization of the above model with the following boundary velocity feedback control (i.e., boundary damping):

$$\bar{f}_1(t) = -k_1 u_t(L,t), \tag{1.12}$$

$$\bar{f}_2(t) = -k_2 w_t(L,t), \tag{1.13}$$

$$\bar{\mu}(t) = -l\varphi_t(L,t), \tag{1.14}$$

where k_1, k_2, and l are positive constants representing feedback gains. For a linear wave equation with boundary damping, Chen [4], [5] adopted an approach using

Lyapunov-type functionals to establish exponential decay of solutions (this work was done as an extension of Quinn and Russell [20]). After that, this approach was used not only for extending Chen's result for linear wave equations (see, e.g., [14], [15], [17]), but also for obtaining exponential decay rates for some linear elastic systems with boundary damping (see, e.g., [12], [3], [7], [6]). One of the main advances in this approach was made by Komornik and Zuazua [9], [8] when they introduced a new type of Lyapunov functional. Their argument showed that the approach can be applied to nonlinear systems with boundary damping; results on uniform decay rates have been obtained for semilinear wave equations (see, e.g., [24]) and nonlinear elastic beams and plates (see, e.g., [13], [18], [10], [11], [16], [21], [19]) with linear or nonlinear boundary damping. On the other hand, there is another important approach to the problem of obtaining uniform decay rates for the similar systems. For this approach, see, for example, Tataru [22] and the bibliography therein.

We apply the Lyapunov functional approach: we construct a functional to establish that the total energy

$$E(t;\varphi) := K(t;\varphi) + U(t;\varphi) \tag{1.15}$$

of the controlled system (1.1)–(1.8), (1.12)–(1.14) has an exponential decay rate. As far as the author knows, all previous work, except [23], on uniform boundary stabilization for flexible systems (string, membrane, beam, plate) treats those models that are derived under the condition that the deflection of the body is small enough, even though the model equations are nonlinear. The present paper shows that the boundary velocity feedback control is effective for large deflecting flexible structures. Independently of the present work, Taylor has shown in an unpublished article [23] that the elastica model without the rotational inertia term (the third term of the left-hand side of (1.3)) is exponentially stabilized by the boundary feedback forces (1.12) and (1.13), assuming the existence of a solution of the model. However, for that model, the existence of a global classical solution has not been proven yet.

2. Preliminary results. In this section, we give some preliminary results, including the well-posedness of (1.1)–(1.8), (1.12)–(1.14).

First of all, we eliminate λ, ν, $\bar{f}_1$, $\bar{f}_2$, and $\bar{\mu}$ from (1.1)–(1.8), (1.12)–(1.14) to obtain

$$\begin{aligned}\int_L^x \rho u_{tt}(\xi,t)\,d\xi \sin\varphi - \int_L^x \rho w_{tt}(\xi,t)\,d\xi \cos\varphi + I_\rho \varphi_{tt} - EI\varphi_{xx} \\ = k_1 u_t(L,t)\sin\varphi - k_2 w_t(L,t)\cos\varphi,\end{aligned} \tag{2.1}$$

$$u(x,t) = \int_0^x \{\cos\varphi(\xi,t) - 1\}\,d\xi, \quad w(x,t) = \int_0^x \sin\varphi(\xi,t)\,d\xi, \tag{2.2}$$

$$\varphi(0,t) = 0, \tag{2.3}$$

$$EI\varphi_x(L,t) = -l\varphi_t(L,t). \tag{2.4}$$

In what follows, we are concerned with (2.1)–(2.4) exclusively.

Caflisch and Maddocks [2] proved the existence and uniqueness of global, classical solutions to the initial-boundary value problem corresponding to (1.1)–(1.8) with

$$\bar{f}_1(t) = \text{const}, \quad \bar{f}_2(t) = \bar{\mu}(t) = 0, \quad t \geq 0$$

(they dealt with the problem with variable coefficients). The existence result, Proposition 2.1 below, for our problem can be proved by similar argument to theirs. Thus we give only the outline of a proof.

Let C^n_{pw} ($n = 1, 2$) denote the class of those functions in C^{n-1} that have piecewise continuous derivatives of order n.

PROPOSITION 2.1. *Let $\varphi_0 \in C^2_{pw}[0, L]$, and let $\varphi_1 \in C^1_{pw}[0, L]$, where φ_0 and φ_1 satisfy*

$$\varphi_0(0) = \varphi_1(0) = 0, \quad EI\varphi_0'(L) = -l\varphi_1(L) \tag{2.5}$$

("$'$" stands for d/dx). Then there exists one and only one function $\varphi \in C^2_{pw}([0, L] \times [0, \infty))$ satisfying (2.1)–(2.4) *and*

$$\varphi(x, 0) = \varphi_0(x), \quad \varphi_t(x, 0) = \varphi_1(x). \tag{2.6}$$

Proof. Step 1: *semilinear form.* Throughout this step, the dependence of φ upon t is suppressed. Substituting (2.2) into (2.1) yields

$$\begin{aligned} (2.7)\quad & I_\rho\varphi_{tt}(x) - EI\varphi_{xx}(x) \\ &\quad + \int_0^L \hat{\rho}(x, \xi)\big\{\varphi_{tt}(\xi)\cos(\varphi(x) - \varphi(\xi)) + \varphi_t(\xi)^2\sin(\varphi(x) - \varphi(\xi))\big\}\, d\xi \\ &= -\int_0^L \varphi_t(\xi)\big\{k_1\sin\varphi(x)\sin\varphi(\xi) + k_2\cos\varphi(x)\cos\varphi(\xi)\big\}\, d\xi, \end{aligned}$$

where

$$\hat{\rho}(x, \xi) = \begin{cases} \rho(L - x) & : \xi < x, \\ \rho(L - \xi) & : x < \xi. \end{cases}$$

By integration by parts and (2.3), we have the equality

$$\begin{aligned} &\int_0^L \hat{\rho}(x, \xi)(EI/I_\rho)\varphi_{xx}(\xi)\cos(\varphi(x) - \varphi(\xi))\, d\xi \\ &= -\,\hat{\rho}(x, 0)(EI/I_\rho)\varphi_x(0)\cos\varphi(x) \\ &\quad - \int_0^L \hat{\rho}_\xi(x, \xi)(EI/I_\rho)\varphi_x(\xi)\cos(\varphi(x) - \varphi(\xi))\, d\xi \\ &\quad - \int_0^L \hat{\rho}(x, \xi)(EI/I_\rho)\varphi_x(\xi)^2\sin(\varphi(x) - \varphi(\xi))\, d\xi. \end{aligned}$$

Subtracting this equality from (2.7) yields

$$\begin{aligned} (2.8)\quad & I_\rho\varphi_{tt}(x) - EI\varphi_{xx}(x) \\ &\quad + \int_0^L \hat{\rho}(x, \xi)I_\rho^{-1}\big\{I_\rho\varphi_{tt}(\xi) - EI\varphi_{xx}(\xi)\big\}\cos(\varphi(x) - \varphi(\xi))\, d\xi = [A\varphi](x), \end{aligned}$$

where

$$\begin{aligned}[A\varphi](x) =&\hat{\rho}(x,0)(EI/I_\rho)\varphi_x(0)\cos\varphi(x)\\ &+\int_0^L \hat{\rho}(x,\xi)\big\{(EI/I_\rho)\varphi_x(\xi)^2-\varphi_t(\xi)^2\big\}\sin(\varphi(x)-\varphi(\xi))\,d\xi\\ &+\int_0^L \hat{\rho}_\xi(x,\xi)(EI/I_\rho)\varphi_x(\xi)\cos(\varphi(x)-\varphi(\xi))\,d\xi\\ &-\int_0^L \varphi_t(\xi)\big\{k_1\sin\varphi(x)\sin\varphi(\xi)+k_2\cos\varphi(x)\cos\varphi(\xi)\big\}\,d\xi.\end{aligned}$$

Define now the operator L_ψ for each $\psi\in C[0,L]$ by

$$[L_\psi f](x)=\int_0^L \hat{\rho}(x,\xi)I_\rho^{-1}\cos(\psi(x)-\psi(\xi))f(\xi)\,d\xi,\quad f\in L^2(0,L).$$

As stated in the next step, the operator $1+L_\psi$ has the inverse $(1+L_\psi)^{-1}=1+K_\psi$, with K_ψ a compact operator on $L^2(0,L)$. Thus (2.8) can be rewritten as

$$I_\rho\varphi_{tt}-EI\varphi_{xx}=(1+K_\varphi)A\varphi. \tag{2.9}$$

In the following, it is shown that this semilinear equation with the boundary conditions (2.3)–(2.4) and the initial conditions (2.6) has a unique C^2_{pw} solution.

Step 2: *properties of* L_φ, K_φ *and* $A\varphi$. Let $Q_T=[0,L]\times[0,T]$ for $T>0$, and let $Q_\infty=[0,L]\times[0,\infty)$. We denote by $||\cdot||_\infty$ and $||\cdot||_{L^2}$ the usual norms of $L^\infty(0,L)$, and $L^2(0,L)$, respectively, and define $||\cdot||_1$ by

$$||f(\cdot,t)||_1^2=||f(\cdot,t)||_{L^2}^2+||f_x(\cdot,t)||_{L^2}^2+||f_t(\cdot,t)||_{L^2}^2\quad\text{for } f\in C^1_{pw}(Q_\infty).$$

Let φ, $\psi\in C[0,L]$. The operator L_φ is a compact, positive operator on $L^2(0,L)$ into $C[0,L]$ and satisfies

$$\begin{gathered}||L_\varphi f||_\infty\le c_1||f||_{L^2},\\ ||(L_\varphi-L_\psi)f||_\infty\le c_1||\varphi-\psi||_\infty||f||_{L^2},\end{gathered}$$

where c_1 is a positive constant depending only on ρ, I_ρ, and L. If $\varphi\in C^1[0,L]$, then $L_\varphi f\in C^1[0,L]$.

Since L_φ is compact and positive definite, the operator $(1+L_\varphi)^{-1}$ on $L^2(0,L)$ is well defined, which is positive and satisfies

$$||(1+L_\varphi)^{-1}f||_{L^2}\le||f||_{L^2},\quad f\in L^2(0,L).$$

We define K_φ by

$$K_\varphi=-L_\varphi(1+L_\varphi)^{-1}=-(1+L_\varphi)^{-1}L_\varphi.$$

Thus, $(1+K_\varphi)=(1+L_\varphi)^{-1}$. By the definition, it follows that K_φ is a compact, self-adjoint operator on $L^2(0,L)$ into $C[0,L]$ and satisfies

$$\begin{gathered}||K_\varphi f||_\infty\le c_2||f||_{L^2},\\ ||(K_\varphi-K_\psi)f||_\infty\le c_2||\varphi-\psi||_\infty||f||_{L^2},\end{gathered}$$

where c_2 is a positive constant depending only on ρ, I_ρ, and L. Moreover, if $\varphi \in C^1[0,L]$, then $K_\varphi f \in C^1[0,L]$.

In the rest of this step, we suppose that $\varphi(x,t)$ and $\psi(x,t)$ belong to $C^1_{pw}(Q_\infty)$ (the dependence of φ and ψ upon t is suppressed) and $\varphi(0,t) = \psi(0,t) = 0$. By the definition of $A\varphi$ and by using the inequalities above, we obtain the following estimations, in which c_3 is a positive constant depending only on ρ, I_ρ, EI, L, k_1, and k_2:

$$||A\varphi||_\infty \leq c_3\big(||\varphi_x||_\infty + ||\varphi||_1 + ||\varphi||_1^2\big), \tag{2.10}$$

$$||(1+K_\varphi)A\varphi||_\infty \leq c_3\big(||\varphi_x||_\infty + ||\varphi||_1 + ||\varphi||_1^2\big), \tag{2.11}$$

$$\begin{aligned}||A\varphi - A\psi||_\infty &\\ \leq c_3\big(1 + ||\varphi_x||_\infty + ||\psi_x||_\infty &+ ||\varphi||_1 + ||\psi||_1 + ||\varphi||_1^2 + ||\psi||_1^2\big)\\ &\times \big(||\varphi_x - \psi_x||_\infty + ||\varphi - \psi||_1\big),\end{aligned} \tag{2.12}$$

$$\begin{aligned}||(1+K_\varphi)A\varphi - (1+K_\psi)A\psi||_\infty &\\ \leq c_3\big(1 + ||\varphi_x||_\infty + ||\psi_x||_\infty &+ ||\varphi||_1 + ||\psi||_1 + ||\varphi||_1^2 + ||\psi||_1^2\big)\\ &\times \big(||\varphi_x - \psi_x||_\infty + ||\varphi - \psi||_1\big).\end{aligned} \tag{2.13}$$

Step 3: *linear wave equation.* Consider the inhomogeneous initial-boundary value problem

$$\varphi_{tt} - \beta^2\varphi_{xx} = f, \quad 0 \leq x \leq L, \quad t \geq 0$$

with the boundary conditions (2.3)–(2.4) and the initial conditions (2.6), where $\beta = \sqrt{EI/I_\rho}$. Using d'Alembert's formula, we can express explicitly the solution of this problem:

$$\begin{aligned}\varphi(x,t) = \frac{1}{2}&\left(\int_0^{x+\beta t}\varphi^+(\xi)\,d\xi + \int_0^{x-\beta t}\varphi^-(\xi)\,d\xi\right)\\ &+ \frac{1}{2\beta}\int_0^t\int_{x-\beta(t-\tau)}^{x+\beta(t-\tau)}\tilde{f}(\xi,\tau)\,d\xi d\tau, \quad 0 \leq x \leq L, \quad t \geq 0.\end{aligned} \tag{2.14}$$

In this formula, $\varphi^+(x)$ $(x \geq 0)$, $\varphi^-(x)$ $(x \leq L)$, and $\tilde{f}(x,t)$ $(-\infty < x < \infty,\ t \geq 0)$ are defined by

$$\varphi^+(x+nL) = \begin{cases}\sigma^k\big(\varphi_0'(x) + \frac{1}{\beta}\varphi_1(x)\big) & : n = 2k,\\ \sigma^k\big(\varphi_0'(L-x) - \frac{1}{\beta}\varphi_1(L-x)\big) & : n = 2k+1\end{cases}$$
$$\text{for } 0 \leq x < L,\ k = 0,\ 1,\ 2,\ \dots,$$

$$\varphi^-(x+nL) = \begin{cases}\sigma^{|k|}\big(\varphi_0'(x) - \frac{1}{\beta}\varphi_1(x)\big) & : n = 2k,\\ \sigma^{|k|}\big(\varphi_0'(L-x) + \frac{1}{\beta}\varphi_1(L-x)\big) & : n = 2k-1\end{cases}$$
$$\text{for } 0 \leq x < L,\ k = 0,\ -1,\ -2,\ \dots,$$

$$\tilde{f}(x+nL,t) = \begin{cases}\sigma^{|k|}f(x,t) & : n = 2k,\\ -\sigma^{|k+1|}f(L-x,t) & : n = 2k+1\end{cases}$$
$$\text{for } 0 \leq x < L,\ t \geq 0,\ k = 0,\ \pm 1,\ \pm 2,\ \dots,$$

where

$$\sigma = -\frac{1-\alpha\beta}{1+\alpha\beta}\ \big(\in(-1,1)\big) \quad \text{with} \quad \alpha = \frac{l}{EI}\ (>0).$$

We can check that if $f(x,t) \in C(Q_\infty)$ and $\partial f/\partial x$ is a piecewise continuous function on Q_∞, then the right-hand side of (2.14) is in $C^2_{pw}(Q_\infty)$ (note that $\partial f/\partial t$ is not required to be piecewise continuous). Moreover, defining $N(\varphi(\cdot,t))$ to be

$$N(\varphi(\cdot,t)) = ||\varphi(\cdot,t)||_\infty + ||\varphi_x(\cdot,t)||_\infty + ||\varphi_t(\cdot,t)||_\infty,$$

it follows from (2.14) that the solution φ satisfies

$$N(\varphi(\cdot,t)) \le c_4\big(||\varphi_0||_\infty + ||\varphi_0'||_\infty + ||\varphi_1||_\infty\big) + c_4\int_0^t ||f(\cdot,\tau)||_\infty\, d\tau \tag{2.15}$$

for $t \ge 0$, where c_4 is a positive constant depending only on I_ρ, EI, L, and l.

Step 4: *existence.* Define $\varphi^{(n)}$ for $n = 0,\ 1,\ 2,\ \ldots$ by

$$\varphi^{(0)}(x,t) = \varphi_0(x) + t\varphi_1(x),$$
$$I_\rho\varphi_{tt}^{(n+1)} - EI\varphi_{xx}^{(n+1)} = (1+K_{\varphi^{(n)}})A\varphi^{(n)}$$

with the boundary conditions (2.3)–(2.4) and the initial conditions (2.6). Let

$$g^{(n)} = \frac{1}{I_\rho}(1+K_{\varphi^{(n)}})A\varphi^{(n)}.$$

Since $g^{(0)} \in C^1_{pw}(Q_\infty)$, $\varphi^{(n+1)}$ belongs to C^2_{pw} and satisfies (2.15) with $f = g^{(n)}$ for $n = 0,\ 1,\ \ldots$. From this estimate and (2.11), we can show that

$$N(\varphi^{(n)}(\cdot,t)) \le M_0, \quad 0 \le t \le t_0,\ n = 0,1,\ldots \tag{2.16}$$

for some positive constants M_0 and t_0 independent of n. Furthermore, letting $\psi^{(n)} = \varphi^{(n+1)} - \varphi^{(n)}$, $\psi^{(n+1)}$ then satisfies (2.15) with $\varphi_0 = \varphi_1 = 0$ and $f = g^{(n+1)} - g^{(n)}$. From this estimate, (2.13), and (2.16), we can see that

$$N(\psi^{(n+1)}(\cdot,t)) \le \tfrac{1}{2}\sup_{0\le\tau\le t_1} N(\psi^{(n)}(\cdot,\tau)), \quad 0 \le t \le t_1,\ n = 0,1,\ldots$$

for some positive constants $t_1 \le t_0$ independent of n. This implies that $\varphi^{(n)}$ is a Cauchy sequence in $C^1(Q_{t_1})$. The constants M_0, t_0, and t_1 depend only on ρ, I_ρ, EI, L, k_1, k_2, l, $||\varphi_0||_\infty$, $||\varphi_0'||_\infty$, $||\varphi_1||_\infty$, and $||\varphi_1'||_\infty$.

Let $\varphi \in C^1(Q_{t_1})$ be the limit of $\varphi^{(n)}$ and $g = I_\rho^{-1}(1+K_\varphi)A\varphi$. It follows from (2.13) that $g^{(n)}$ converges to g in $C(Q_{t_1})$. Hence, φ satisfies (2.14) with $f = g$. On the other hand, by the definitions of $A\varphi$ and K_φ, we see that $\partial g/\partial x \in C(Q_{t_1})$. Therefore, as stated in Step 3, φ is a C^2_{pw} solution of (2.9). The uniqueness of solutions follows easily from (2.15) and (2.13).

Global existence of the solution can be proved in exactly the same way as in the proof of Theorem 1 in [2] using the estimate $E(t;\varphi) \le E(0;\varphi)$, $t \ge 0$, of the energy E (see the proof of Lemma 3.2 below). □

The inequalities in the following lemma are used in the next section.

Lemma 2.2. *Let $\psi \in C^1_{pw}[0,L]$, and define*

$$p(x) = \int_0^x \{\cos\psi(\xi) - 1\}\, d\xi, \quad q(x) = \int_0^x \sin\psi(\xi)\, d\xi.$$

Then, for any $x \in [0,L]$, we have

$$|xp'(x) - p(x)| \le L^{3/2}\left(\int_0^L |\psi'(\xi)|^2\, d\xi\right)^{1/2},$$
$$|xq'(x) - q(x)| \le L^{3/2}\left(\int_0^L |\psi'(\xi)|^2\, d\xi\right)^{1/2}.$$

Proof. From the definition of p, it follows that for any $x \in [0,L]$,

$$\begin{aligned}|xp'(x) - p(x)| &= \left|\int_0^x \big(\cos\psi(x) - \cos\psi(\xi)\big)\, d\xi\right| \\ &\le \int_0^x |\psi(x) - \psi(\xi)|\, d\xi = \int_0^L \left|\int_\xi^x \psi'(\eta) d\eta\right| d\xi \\ &\le \int_0^L \int_0^L |\psi'(\eta)|\, d\eta d\xi \le L^{3/2}\left(\int_0^L |\psi'(\xi)|^2\, d\xi\right)^{1/2}.\end{aligned}$$

The inequality for $|xq'(x) - q(x)|$ can be verified in the same way as above. □

3. The main result and its proof.

The main result is as follows.

Theorem 3.1. *Let $\varphi(x,t)$ be the function stated in Proposition* 2.1. *Then there exist constants $M \ge 1$ and $\gamma > 0$ such that*

$$E(t;\varphi) \le Me^{-\gamma t}E(0;\varphi), \quad t \ge 0.$$

Proof. First of all, we introduce the functional

$$V(t;\varphi) = E(t;\varphi) + \epsilon F(t;\varphi) \quad (\epsilon > 0),$$

where $E(t;\varphi)$ is the energy functional defined in (1.15), and

$$F(t;\varphi) = \int_0^L \{\rho u_t(xu_x - u) + \rho w_t(xw_x - w) + I_\rho x\varphi_t\varphi_x\}\, dx.$$

Differentiating $V(t;\varphi)$ in t, we have (see Lemma 3.2 below)

$$\begin{aligned}\dot V(t;\varphi) = &- [k_1u_t^2 + k_2w_t^2 + l\varphi_t^2]_{x=L} \\ &- \epsilon E(t;\varphi) - \epsilon\int_0^L (\rho u_t^2 + \rho w_t^2)\, dx \\ &+ \tfrac{1}{2}L\epsilon[\rho u_t^2 + \rho w_t^2 + I_\rho\varphi_t^2 + (l^2/EI)\varphi_t^2]_{x=L} \\ &- \epsilon[k_1u_t(xu_x - u) + k_2w_t(xw_x - w)]_{x=L}\end{aligned} \tag{3.1}$$

("." stands for d/dt). Now using Lemma 2.2 and the inequality

$$\xi\eta \le \delta\xi^2 + \eta^2/(4\delta) \quad \text{for any real } \xi,\ \eta, \text{ and any } \delta > 0,$$

we obtain the following estimate on $\dot{V}(t;\varphi)$: for any $\epsilon > 0$ and $\delta > 0$,

$$\begin{aligned}\dot{V}(t;\varphi) \leq &-(\epsilon/2)\int_0^L (3\rho u_t^2 + 3\rho w_t^2 + I_\rho \varphi_t^2)\,dx \\ &- \epsilon \int_0^L (EI/2 - 2\delta)\varphi_x^2\,dx \\ &- \left\{k_1 - \left(\frac{L\rho}{2} + \frac{k_1 L^{3/2}}{4\delta}\right)\epsilon\right\} u_t(L,t)^2 \\ &- \left\{k_2 - \left(\frac{L\rho}{2} + \frac{k_2 L^{3/2}}{4\delta}\right)\epsilon\right\} w_t(L,t)^2 \\ &- \left\{l - \left(\frac{LI_\rho}{2} + \frac{Ll^2}{2EI}\right)\epsilon\right\} \varphi_t(L,t)^2.\end{aligned}$$

Therefore, if we choose δ, ϵ and the control gains such that

$$\begin{aligned}&k_1 > 0,\ k_2 > 0,\ l > 0, \\ &k_1 - \left(\frac{L\rho}{2} + \frac{k_1 L^{3/2}}{4\delta}\right)\epsilon > 0, \quad k_2 - \left(\frac{L\rho}{2} + \frac{k_2 L^{3/2}}{4\delta}\right)\epsilon > 0, \\ &l - \left(\frac{LI_\rho}{2} + \frac{Ll^2}{2EI}\right)\epsilon > 0, \quad EI/2 - 2\delta > 0,\end{aligned} \tag{3.2}$$

we have

$$\dot{V}(t;\varphi) \leq -\epsilon(1 - 4\delta/EI)E(t;\varphi), \quad t \geq 0. \tag{3.3}$$

On the other hand, it follows from Schwarz's inequality and Lemma 2.2 that

$$|F(t;\varphi)| \leq CE(t;\varphi), \quad t \geq 0,$$

where C is a constant depending only on ρ, I_ρ, EI, and L. This implies

$$(1 - C\epsilon)E(t;\varphi) \leq V(t;\varphi) \leq (1 + C\epsilon)E(t;\varphi). \tag{3.4}$$

Therefore, it follows from (3.3) that

$$\dot{V}(t;\varphi) \leq -\gamma V(t;\varphi), \quad t \geq 0, \tag{3.5}$$

where

$$\gamma = \epsilon\left(1 - \frac{4\delta}{EI}\right)\frac{1}{1 + C\epsilon} > 0.$$

If ϵ is chosen so as to satisfy, in addition to (3.2),

$$1 - C\epsilon > 0,$$

the inequalities (3.5) and (3.4) imply

$$E(t;\varphi) \leq \frac{1 + C\epsilon}{1 - C\epsilon} e^{-\gamma t} E(0;\varphi), \quad t \geq 0.$$

The proof is thus completed. □

Lemma 3.2. *For the function φ stated in Proposition* 2.1, *the equality* (3.1) *holds.*

Proof. In the proof, we often use the equality

$$\int_0^L f(x)\Big(\int_0^x g(\xi)\,d\xi\Big)\,dx = -\int_0^L \Big(\int_L^x f(\xi)\,d\xi\Big)g(x)\,dx, \tag{3.6}$$

where f and g are functions of $L^2(0,L)$.

Differentiating $V(t;\varphi)$ in t, we have

$$\dot V(t;\varphi) = \dot E(t;\varphi) + \epsilon \dot F(t;\varphi).$$

Let us calculate the right-hand side of the above. First of all, multiplying both sides of (2.1) by φ_t and integrating the result in x from 0 to L, we can see from (3.6), (2.2), integration by parts, and (2.4) that

$$\dot E(t;\varphi) = -[k_1u_t^2 + k_2w_t^2 + l\varphi_t^2]_{x=L}.$$

Next, we consider

$$\begin{aligned}\dot F(t;\varphi) = &\int_0^L \{\rho u_{tt}(xu_x - u) + \rho w_{tt}(xw_x - w) + I_\rho x\varphi_{tt}\varphi_x\}\,dx \\ &+ \int_0^L \{\rho u_t(xu_{xt} - u_t) + \rho w_t(xw_{xt} - w_t) + I_\rho x\varphi_t\varphi_{xt}\}\,dx.\end{aligned} \tag{3.7}$$

Substituting (2.1) into $I_\rho\varphi_{tt}$ of the above, and using (3.6), we see that the first term of the right-hand side of (3.7) is equal to

$$\begin{aligned}\int_0^L \Big\{&\rho u_{tt}(xu_x - u) + \rho w_{tt}(xw_x - w) \\ &- \rho u_{tt}\int_0^x \xi(\cos\varphi(\xi,t))_x\,d\xi - \rho w_{tt}\int_0^x \xi(\sin\varphi(\xi,t))_x\,d\xi \\ &+ \tfrac12 EIx(\varphi_x^2)_x - k_1u_t(L,t)x(\cos\varphi)_x - k_2w_t(L,t)x(\sin\varphi)_x\Big\}\,dx.\end{aligned} \tag{3.8}$$

Furthermore, from integration by parts and (2.2), we have that (3.8) is equal to

$$\frac12 EIL[\varphi_x^2]_{x=L} - \int_0^L \frac12 EI\varphi_x^2\,dx - [k_1u_t(xu_x - u) + k_2w_t(xw_x - w)]_{x=L}.$$

The second term of the right-hand side of (3.7) is calculated as

$$\begin{aligned}&\int_0^L \frac{x}{2}\{\rho(u_t^2)_x + \rho(w_t^2)_x + I_\rho(\varphi_t^2)_x\}\,dx - \int_0^L (\rho u_t^2 + \rho w_t^2)\,dx \\ &= \frac{L}{2}[\rho u_t^2 + \rho w_t^2 + I_\rho\varphi_t^2]_{x=L} - \frac12\int_0^L (3\rho u_t^2 + 3\rho w_t^2 + I_\rho\varphi_t^2)\,dx.\end{aligned}$$

Therefore, we arrive at

$$\begin{aligned}\dot F(t;\varphi) = &- E(t;\varphi) - \int_0^L (\rho u_t^2 + \rho w_t^2)\,dx \\ &+ (L/2)[\rho u_t^2 + \rho w_t^2 + I_\rho\varphi_t^2 + EI\varphi_x^2]_{x=L} \\ &- [k_1u_t(xu_x - u) + k_2w_t(xw_x - w)]_{x=L}.\end{aligned}$$

The proof is thus completed. □

Acknowledgments. The author is grateful to Professors D. L. Russell and M. Renardy for helpful discussions and their hospitality during a stay at Virginia Tech.

REFERENCES

[1] S. S. Antman, *Nonlinear Problems of Elasticity*, Springer-Verlag, New York, 1995.

[2] R. E. Caflisch and J. H. Maddocks, *Nonlinear dynamical theory of the elastica*, Proc. Roy. Soc. Edinburgh Sect. A, 99 (1984), pp. 1–23.

[3] G. Chen, M. C. Delfour, A. M. Krall, and G. Payre, *Modeling, stabilization and control of serially connected beams*, SIAM J. Control Optim., 25 (1987), pp. 526–546.

[4] G. Chen, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249–273.

[5] G. Chen, *A note on the boundary stabilization of the wave equation*, SIAM J. Control Optim., 19 (1981), pp. 106–113.

[6] K. Ito and N. Kunimatsu, *Semigroup model and stability of the structurally damped Timoshenko beam with boundary inputs*, Internat. J. Control, 54 (1991), pp. 367–391.

[7] J. U. Kim and Y. Renardy, *Boundary control of the Timoshenko beam*, SIAM J. Control Optim., 25 (1987), pp. 1417–1429.

[8] V. Komornik and E. Zuazua, *A direct method for the boundary stabilization of the wave equation*, J. Math. Pures Appl., 69 (1990), pp. 33–54.

[9] V. Komornik and E. Zuazua, *Stabilization frontière de l'équation des ondes: une méthode directe*, C. R. Acad. Sci. Paris Sér. I Math., 305 (1987), pp. 605–608.

[10] V. Komornik, *On the boundary stabilization of plates*, Pitman Res. Notes Math. Ser. 249, Longman Scientific and Technical, Harlow, UK, 1991, pp. 105–111.

[11] J. E. Lagnese and G. Leugering, *Uniform stabilization of a nonlinear beam by nonlinear boundary feedback*, J. Differential Equations, 91 (1991), pp. 355–388.

[12] J. E. Lagnese, *Boundary stabilization of linear elastodynamic systems*, SIAM J. Control Optim., 21 (1983), pp. 968–984.

[13] J. E. Lagnese, *Boundary Stabilization of Thin Plates*, SIAM Stud. Appl. Math. 10, SIAM, Philadelphia, 1989.

[14] J. E. Lagnese, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163–182.

[15] J. E. Lagnese, *Note on boundary stabilization of wave equations*, SIAM J. Control Optim., 26 (1988), pp. 1250–1256.

[16] J. E. Lagnese, *Uniform asymptotic energy estimates for solutions of the equations of dynamic plane elasticity with nonlinear dissipation at the boundary*, Nonlinear Anal., 16 (1991), pp. 35–54.

[17] I. Lasiecka and R. Triggiani, *Uniform exponential energy decay of wave equations in a bounded region with $L_2(0,\infty;L_2(\Gamma))$-feedback control in the Dirichlet boundary conditions*, J. Differential Equations, 66 (1987), pp. 340–390.

[18] G. Leugering, *On boundary feedback stabilisability of a viscoelastic beam*, Proc. Roy. Soc. Edinburgh Sect. A, 114 (1990), pp. 57–69.

[19] J.-P. Puel and M. Tucsnak, *Boundary stabilization for the von Kármán equations*, SIAM J. Control Optim., 33 (1995), pp. 255–273.

[20] J. P. Quinn and D. L. Russell, *Asymptotic stability and energy decay rates for solutions of hyperbolic equations with boundary damping*, Proc. Roy. Soc. Edinburgh Sect. A, 77 (1977), pp. 97–127.

[21] B. Rao, *Stabilization of Kirchhoff plate equation in star-shaped domain by nonlinear boundary feedback*, Nonlinear Anal., 20 (1993), pp. 605–626.

[22] D. Tataru, *Uniform decay rates and attractors for evolution PDE's with boundary dissipation*, J. Differential Equations, 121 (1995), pp. 1–27.

[23] S. W. Taylor, *Project Description in NSF Proposal*, manuscript, 1994.

[24] E. Zuazua, *Uniform stabilization of the wave equation by nonlinear boundary feedback*, SIAM J. Control Optim., 28 (1990), pp. 466–477.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 1, pp. 330–331

ERRATUM: A UNIQUENESS RESULT FOR THE LINEAR SYSTEM OF ELASTICITY AND ITS CONTROL THEORETICAL CONSEQUENCES*

ENRIQUE ZUAZUA†

Abstract. In this note we explain why the three-dimensional counterexample of section 7.2 of the paper under consideration [E. Zuazua, *SIAM J. Control Optim.,* 34 (1996), pp. 1473–1495] is incorrect. It concerns the existence of eigenfunctions for the Dirichlet Laplacian of a special form. We claimed the existence of particular three-dimensional domains in which these eigenfunctions do exist. However, in a recent joint work with G. Sweers [G. Sweers and E. Zuazua, *J. Elasticity,* to appear], we have proved that they may not exist in any bounded Lipschitz domain in dimensions $n \geq 3$.

PII. S0363012998334122

There is a mistake in the counterexample given in section 7.2 for dimension $n = 3$. Indeed, the function ϕ that we propose on page 1490 of [1] does not satisfy the Dirichlet homogeneous boundary condition in the second surface of formula (7.2). This function ϕ does satisfy the differential equation

$$-\mu\Delta\Phi - (\lambda + 2\mu)\frac{\partial^2\Phi}{\partial x_3^2} = \kappa^2\Phi \tag{1}$$

but may not be considered as a counterexample for the Dirichlet problem. The domains indicated in the figures of that section correspond to the analytical expressions

$$-\frac{\sqrt{\lambda + 2\mu}}{k}\arcsin\left(p(x_1, x_2)\right) \leq x_3 \leq \frac{\sqrt{\lambda + 2\mu}}{k}\arcsin\left(p(x_1, x_2)\right). \tag{2}$$

The function ϕ given in that section can be slightly modified to fulfill the Dirichlet boundary condition on the boundary of that domain. It is sufficient to take

$$\phi\left(x_1, x_2, x_3\right) = p\left(x_1, x_2\right) - \sin\left(\frac{k}{\sqrt{\lambda + 2\mu}} \mid x_3 \mid\right) \tag{3}$$

(we just replace x_3 by $\mid x_3 \mid$). Note, however, that this ϕ satisfies

$$-\mu\Delta\phi - (\lambda + \mu)\partial_3^2\phi = k^2\phi + \xi, \tag{4}$$

ξ being a measure with support on $x_3 = 0$. Thus, it does not quite solve the eigenvalue problem under consideration. In fact, in a recent joint work with G. Sweers [2] we prove that, in dimension $n = 3$, there is no Lipschitz domain Ω in which there is a nontrivial solution of (1) satisfying the homogeneous Dirichlet boundary conditions and being of the form $\Phi(x_1, x_2, x_3) = p(x_1, x_2) + q(x_3)$. This unexpected result is sharp since, as indicated in section 7.1, in dimension $n = 2$, there are polygonal domains in which the corresponding two-dimensional problem admits nontrivial eigenfunctions of the form $\Phi(x_1, x_2) = p(x_1) + q(x_2)$. The rest of the results of this paper remain true. Note, however, that, according to the new developments indicated above, Theorem 1.3 is true for all Lipschitz bounded domains when $n = 3$ and not only under the additional assumptions of Theorem 1.1.

*Received by the editors February 17, 1998; accepted for publication February 17, 1998; published electronically November 17, 1998.

http://www.siam.org/journals/sicon/37-1/33412.html

†Departamento de Matemática Aplicada, Universidad Complutense, 28040 Madrid, Spain (zuazua@eucmax.sim.ucm.es).

REFERENCES

[1] E. Zuazua, *A uniqueness result for the linear system of elasticity and its control theoretical consequences*, SIAM J. Control Optim., 34 (1996), pp. 1473–1495.

[2] G. Sweers and E. Zuazua, *On the non-existence of some special eigenfunctions for the Dirichlet Laplacian and the Lamé system*, J. Elasticity, to appear.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 2, pp. 333–352

EXISTENCE AND APPROXIMATION OF ROBUST SOLUTIONS OF VARIATIONAL INEQUALITY PROBLEMS OVER POLYTOPES*

GERARD VAN DER LAAN[†], DOLF TALMAN[‡], AND ZAIFU YANG[§]

Abstract. We study nonlinear variational inequality problems over polytopes from a viewpoint of stability and propose a new solution concept. Extending an earlier concept proposed by Yang [Z. Yang, *SIAM J. Control Optim.*, 34 (1996), pp. 491–506] on the unit simplex, we will introduce the concept of the robust stationary point, which is a refinement of the concept of the stationary point. Though a stationary point need not be robust, it is shown that every continuous function on a polytope has a robust stationary point. We develop a simplicial algorithm to compute a robust stationary point of a continuous function on a polytope. The algorithm can be briefly stated as follows. Starting with any point in the relative interior of a polytope, the algorithm generates a piecewise linear path which leads to an approximate robust stationary point of any a priori chosen accuracy within a finite number of steps. Moreover, we also discuss several numerical examples and apply the new concept to noncooperative games and economic equilibrium problems.

Key words. variational inequality problem, robust stationary point, polytopes, simplicial algorithm, fixed points, stability, Nash equilibrium, Walrasian equilibrium

AMS subject classifications. Primary, 49D35, 90A14; Secondary, 90C30, 90C33

PII. S0363012996309344

1. Introduction.

We study nonlinear stationary point problems from a view point of stability. Given an arbitrary polytope P in $\mathbb{R}^n$ and an arbitrary function $f : \mathrm{P} \mapsto \mathbb{R}^n$, the problem of stationary point (or variational inequality) for f on P is to find a point $x^* \in \mathrm{P}$ such that

$$(x^* - x)^\top f(x^*) \geq 0$$

for any point $x \in \mathrm{P}$. Such a point x^* is called a stationary point of f on P. This problem has many important applications in various fields, such as noncooperative game theory, economic equilibrium theory, fixed point theory, nonlinear optimization theory, and engineering. The concept of a robust stationary point is a refinement of the concept of a stationary point of a continuous function on the unit simplex and was essentially motivated from problems of economic equilibrium, noncooperative games, biology, and engineering (see, e.g., Myerson [13], van Damme [3], and Yamamoto [20]). It is well known (see Hartman and Stampacchia [9] and Eaves [4]) that a continuous function on a compact convex nonempty subset of $\mathbb{R}^n$ has at least one stationary point. In general, a continuous function on a compact convex nonempty subset of $\mathbb{R}^n$ has multiple stationary points, and some of them are undesirable from a viewpoint of

*Received by the editors September 16, 1996; accepted for publication (in revised form) February 9, 1998; published electronically November 23, 1998.
http://www.siam.org/journals/sicon/37-2/30934.html

[†] Department of Econometrics, Free University, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands (glaan@econ.vu.nl). The research of this author was part of the research program "Competition and Cooperation."

[‡] Department of Econometrics and CentER, Tilburg University, P. O. Box 90153, 5000 LE Tilburg, The Netherlands (talman@kub.nl). The research of this author was part of the research program "Competition and Cooperation."

[§] Faculty of Business Administration, Yokohama National University, 79-4 Tokiwadai, Hodogaya-ku, Yokohama 240-0067, Japan (zyang@business.ynu.ac.jp). The research of this author was supported by the Japan Society for Promotion of Science while he was a JSPS fellow at the Institute of Socio-Economic Planning, University of Tsukuba, Japan.

stability. Hence it is very important to eliminate those undesirable stationary points in the context of game theory and general equilibrium theory.

Based on these ideas we will apply the concept of robust stationary point to nonlinear stationary point problems on polytopes. Though a stationary point need not be robust, it is shown that every continuous function on a polytope has a robust stationary point. When we apply this new concept to game theoretic problems and economic equilibrium problems, this concept is both game-theoretically and economically meaningful. Furthermore, we develop a simplicial algorithm to find a robust stationary point for any continuous function on a polytope. This algorithm can be briefly described as follows. Starting from an arbitrarily chosen interior point of the polytope, the algorithm generates a piecewise linear (PL) path which leads to an approximate robust stationary point of any a priori chosen accuracy within a finite number of steps. The path traced by the algorithm corresponds to a sequence of θ-robust stationary points of the PL approximation of the function with respect to a well-chosen triangulation of the polytope, where $0 < \theta \leq 1$. This triangulation has the novelty that when the path generated by the algorithm approaches the boundary of the polytope, the mesh size of the triangulation along the path automatically converges to zero. This property induces that θ goes to zero for the θ-robust stationary points induced by the path when the path approaches the boundary. Hence the algorithm shares the basic properties of a simplicial homotopy algorithm (see Eaves [5]) and a simplicial restart algorithm (see van der Laan and Talman [11]), but it dispenses with an extra dimension required by simplicial homotopy algorithms. It is also worthwhile to stress that another motivation of this paper comes from variational inequality problems in which the function is not monotone. We only require the function to be continuous. The results we obtain in this paper considerably generalize those of Yang [21] on the unit simplex and of Talman and Yang [17] on the simplotope. For extensive treatments of simplicial algorithms we refer to Allgower and Georg [1], Todd [18], and Yang [22].

This paper is organized as follows. In section 2 we introduce the concept of robust stationary point on polytopes and prove the existence of a robust stationary point for every continuous function. We also give an interpretation of a robust stationary point in term of the complementary slackness condition. Section 3 presents a triangulation of the polytope which underlies the algorithm. In section 4 we give the path of points followed by the algorithm, prove the convergence of the algorithm under the assumption that the function f to be considered is continuous, and also derive the accuracy of an approximate robust stationary point. Section 5 describes the steps of the algorithm. In section 6 we discuss two important and practical applications of the concept of robust stationary point in the context of game theory and economic equilibrium theory.

2. Robust stationary points on polytopes.

We first introduce some notation. The set N denotes the set of all natural numbers. The set N_0 is equal to the union of N and $\{0\}$. The notion $I \subset J$ means I is a proper subset of J and $I \subseteq J$ means $I \subset J$ or $I = J$. The set I_n denotes the set of positive integers $\{1, \ldots, n\}$. The vector $e(i)$ is the ith unit vector of $\mathbf{R}^n$ for each $i \in I_n$. The notion $\mathbf{0}$ is the n-vector of all zeros. Consider an arbitrary full-dimensional polytope

$$\mathrm{P} = \{x \in \mathbf{R}^n \mid a^{i\top} x \leq b_i, \text{ for all } i \in I_m\}.$$

We assume that P is simple and that no constraints are redundant.

For each subset I of I_m, define

$$F(I) = \{x \in \mathrm{P} \mid a^{i\top} x = b_i \ \text{ for all } i \in I\}.$$

Then $F(I)$ is a face of P unless it is empty. Note that $F(\emptyset) = \mathrm{P}$. Let

$$\mathcal{I} = \{\, I \subseteq I_m \mid F(I) \ \text{ is a nonempty face of } \ \mathrm{P}\}.$$

Under the above assumption that $\dim(P) = n$, P is a simple polytope, and the linear inequalities defining P are nonredundant, we have the following observations:

(i) for each face F of P the set $I \in \mathcal{I}$ with $F = F(I)$ is unique and identical with the set $\{i \in I_m \mid a^{i\top} x = b_i \ \text{ for all } x \in F\}$;

(ii) $\dim(F(I)) = n - |I|$;

(iii) G is a facet of $F(I)$ if and only if $G = F(I \cup \{h\})$ for some $h \notin I$ with $I \cup \{h\} \in \mathcal{I}$.

For each $I \in \mathcal{I}$, define

$$F^*(I) = \left\{ x \in \mathbb{R}^n \mid x = \sum_{h \in I} \nu_h a^h, \ \nu_h \geq 0, \ \text{ for all } \ h \in I \right\},$$

with $F^*(\emptyset) = \{\mathbf{0}\}$. Now we introduce the concept of a robust stationary point on polytopes. Let a function $f : \mathrm{P} \mapsto \mathbf{R}^n$ be given.

Definition 2.1. *For given $\theta > 0$ a point $x \in \mathrm{P}$ is a θ-robust stationary point of f if*

(1) *x is an interior point of* P*;*

(2) *for some $I \in \mathcal{I}$, $f(x) = \sum_{h \in I_m} \mu_h a^h$ with $\mu_h \geq 0$ for all $h \in I$ and $\mu_h = 0$ for all $h \in I_m \setminus I$, if $\mu_l > \mu_k$, then $b_l - a^{l\top} x \leq \theta(b_k - a^{k\top} x)$.*

A geometric interpretation of the above definition will be given later. We first give an easy observation.

Lemma 2.2. *For any $c \in \mathbb{R}^n$ there exists $I \in \mathcal{I}$ with $|I| = n$ such that $c = \sum_{h \in I} \mu_h a^h$ with $\mu_h \geq 0$ for all $h \in I$.*

It follows immediately that Definition 2.1 is valid.

Definition 2.3. *A point $x^* \in \mathrm{P}$ is a robust stationary point of f on* P *if there exist sequences $\{\, \theta_l \mid l \in \mathsf{N}\,\}$ of positive numbers and $\{\, x(\theta_l) \mid l \in \mathsf{N}\,\}$ of θ_l-robust stationary points $x(\theta_l)$ of f such that*

$$\lim_{l \to \infty} \theta_l = 0 \quad \textit{and} \quad \lim_{l \to \infty} x(\theta_l) = x^*.$$

Now we derive some properties of stationary points. The following lemma is explicitly given in Burke and More [2] and is implicitly used in Talman and Yamamoto [16].

Lemma 2.4. *Let $f : \mathrm{P} \mapsto \mathbb{R}^n$ be a continuous function. Then $x^* \in \mathrm{P}$ is a stationary point of f on* P *if and only if $x^* \in F(I)$ and $f(x^*) \in F^*(I)$ for some $I \in \mathcal{I}$.*

The following result states that the concept of robust stationary point is a refinement of that of the stationary point.

Lemma 2.5. *Let $f : \mathrm{P} \mapsto \mathbb{R}^n$ be a continuous function. If $x^* \in \mathrm{P}$ is a robust stationary point of f on* P*, then x^* is also a stationary point of f on* P.

Proof. We need to consider two cases. If x^* lies in the interior of P, then it follows from Definitions 2.1 and 2.3 that $f(x^*) = \mathbf{0}$. Hence x^* is a stationary point of f by

Lemma 2.4. If x^* lies on the boundary of P, it easily follows from Definitions 2.1 and 2.3 and Lemma 2.4 that x^* is a stationary point of f on P. □

Now we are going to give an interpretation of a robust stationary point. For a stationary point $x^* \in F(I)$ for some $I \in \mathcal{I}$, we have that $\mu_i > 0$ implies $b_i - a^{i\top} x^* = 0$ for any i. One might think of this relation as the complementary slackness condition, as in the linear programming (LP) theory. In order for x^* to be a robust stationary point, it is reasonable to require that the higher μ_i is, the stronger this equality should be. That is to say, in a neighborhood of x^* for any $\theta > 0$ there should be a point $x(\theta)$ such that if $\mu_i(\theta) > \mu_j(\theta)$, then $x(\theta)$ must be θ times closer to $F(\{i\})$ than to $F(\{j\})$, i.e., $b_i - a^{i\top} x(\theta) \leq \theta(b_j - a^{j\top} x(\theta))$. In section 6 we shall apply the concept of robust stationary point to noncooperative game theory and economic equilibrium theory.

The next two examples demonstrate that the concept of the robust stationary point is indeed a proper refinement of the concept of the stationary point.

Example 2.6. *Take $n = 2$ and $a^1 = -a^3 = (1,0)^\top$, $a^2 = -a^4 = (0,1)^\top$, $b_1 = b_2 = 1$, and $b_3 = b_4 = 0$. Then $\mathrm{P} = C^2 = \{x \in \mathbb{R}^2 \mid x_1 \leq 1,\, x_2 \leq 1,\, -x_1 \leq 0, \text{ and } -x_2 \leq 0\}$. Let a continuous function $f : \mathrm{P} \mapsto \mathbb{R}^2$ be given by*

$$f(x) = (x_1 - 1, x_2 - 1)^\top.$$

It is not difficult to show that for any continuous function $f : C^2 \mapsto \mathbb{R}^2$, x^* is a stationary point of f if and only if it holds that

$$\begin{array}{lll} f_i(x^*) \leq 0 & \text{if} & x_i^* = 0, \\ f_i(x^*) = 0 & \text{if} & 0 < x_i^* < 1, \\ f_i(x^*) \geq 0 & \text{if} & x_i^* = 1. \end{array}$$

The set of stationary points of the function is equal to

$$\{(0,0)^\top,\ (1,1)^\top,\ (1,0)^\top,\ (0,1)^\top\}.$$

However, only $(0,0)^\top$ is a robust stationary point of f. To show this, let for $\theta \in (0,1)$

$$x_1(\theta) = \theta,\ x_2(\theta) = \theta^2.$$

Clearly, $x(\theta)$ lies in the interior of P for each $\theta \in (0,1)$. Note that

$$f(x(\theta)) = \mu_3 a^3 + \mu_4 a^4 = \mu_3(-1,0)^\top + \mu_4(0,-1)^\top,$$

with $\mu_3 > 0$ and $\mu_4 > 0$. In fact we have $\mu_4 = 1 - \theta^2 > \mu_3 = 1 - \theta$. It also holds that

$$b_4 - a^{4\top} x(\theta) = \theta^2 \leq \theta(b_3 - a^{3\top} x(\theta)) = \theta^2.$$

Hence for each $\theta \in (0,1)$, $x(\theta)$ is a θ-robust stationary point. It is easy to see that

$$\lim_{\theta \to 0^+} x(\theta) = (0,0)^\top.$$

It is also easy to check why all other stationary points are not robust stationary points. We leave it to the reader.

Example 2.7. *Let $f : C^2 \mapsto \mathbb{R}^2$ be given by*

$$f(x) = (x_2(1 - x_1)^2(x_2 - 1), x_1(1 - x_2)^2(x_1 - 1))^\top,$$

where the set C^2 is as given in Example 2.6.

The set of stationary points of this function is equal to

$$\begin{gathered}\{(0,x)^\top \mid 0 \le x \le 1\} \cup \{(1,x)^\top \mid 0 \le x \le 1\} \\ \cup\{(x,1)^\top \mid 0 \le x \le 1\} \cup \{(x,0)^\top \mid 0 \le x \le 1\}.\end{gathered}$$

It is remarkable that this function has only one stationary point $(0,0)^\top$ which is a robust stationary point. We will prove it. For $\theta \in (0,1)$, let

$$x(\theta) = (\theta, \theta^2)^\top.$$

Clearly, $x(\theta)$ is an interior point of C^2. For each $\theta \in (0,1)$ we have

$$f(x(\theta)) = \mu_3 a^3 + \mu_4 a^4 = \mu_3(-1,0)^\top + \mu_4(0,-1)^\top,$$

with $\mu_3 = \theta^2(1-\theta)^2(1-\theta^2) > 0$ and $\mu_4 = \theta(1-\theta^2)^2(1-\theta) > 0$. Since $\frac{\mu_4}{\mu_3} = \frac{(1+\theta)}{\theta} > 1$, $\mu_4 > \mu_3$. It also holds that

$$b_4 - a^{4\top} x(\theta) = \theta^2 \le \theta(b_3 - a^{3\top} x(\theta)) = \theta^2.$$

Hence $x(\theta)$ is a θ-robust stationary point. Since

$$\lim_{\theta \to 0^+} x(\theta) = (0,0)^\top,$$

the point $(0,0)^\top$ is a robust stationary point by definition. It is left to the reader to check that any other stationary point is not a robust stationary point.

We now come to the question of under what conditions there exists a robust stationary point for a function on a polytope. In the remainder of this section we prove that every continuous function $f : \mathrm{P} \mapsto \mathbb{R}^n$ has a robust stationary point, and hence continuity is enough to ensure the existence of a robust stationary point.

Define the function $\eta : \mathbb{R}^n \mapsto \mathbb{R}$ by

$$\eta(x) = \min_{i \in I_m} (b_i - a^{i\top} x).$$

Let $\Upsilon = \max_{x \in \mathrm{P}} \eta(x)$. Since P is a bounded full-dimensional polyhedron, it is easy to see that $\Upsilon > 0$. Take any $\omega \in (0, \Upsilon]$. For $I \in \mathcal{I}$ and $\theta \in [0, \frac{1}{2}]$, we define a^I and $b_I(\theta)$ by

$$\begin{aligned} a^I &= \sum_{h \in I} a^h, \\ b_I(\theta) &= \sum_{h \in I} b_h - \omega \sum_{k=n+1-|I|}^{n} \theta^k. \end{aligned} \tag{2.1}$$

Definition 2.8. *For each $\theta \in [0, \frac{1}{2}]$, the set $A(\theta)$ in $\mathbb{R}^n$ is given by*

$$A(\theta) = \{\, x \in \mathbb{R}^n \mid a^{I\top} x \le b_I(\theta), \ \textit{for any} \ \ I \in \mathcal{I} \,\}.$$

We remark that for each $\theta \in [0, \frac{1}{2}]$, $A(\theta)$ is a polytope which is a nonempty subset of P and $A(0) = \mathrm{P}$. Note that we take the interval $[0, \frac{1}{2}]$ for the purpose of the triangulation to be introduced in the coming section.

Let the collection of ordered indexed sets, $\mathcal{J}$, be defined by

$$\mathcal{J} = \{L = (L_1, \ldots, L_k) \mid \emptyset \subset L_1 \subset \cdots \subset L_k \subseteq I \ \text{ for each } \ I \in \mathcal{I}, \ k \in I_n \,\} \cup \{\emptyset\}.$$

For $L = (L_1, \dots, L_k) \in \mathcal{J}$ and $\theta \in (0, \frac{1}{2}]$, let

$$F(\theta; L) = \{x \in A(\theta) \mid a^{L_h\top} x = b_{L_h}(\theta) \ \text{ for all } h \in I_k\}.$$

Then $F(\theta; L)$ is a face of $A(\theta)$ with dimension equal to $n - k$ in case $L_1 \neq \emptyset$. Note that $F(\theta; \emptyset) = A(\theta)$. For each $L = (L_1, \dots, L_k) \in \mathcal{J}$, define

$$F^*(L) = \{x \in \mathbb{R}^n \mid x = \sum_{h=1}^{k} \mu_h a^{L_h},\ \mu_h \geq 0 \ \text{ for all } \ h \in I_k\}$$

with $F^*(\emptyset) = \{\mathbf{0}\}$.

Let $f : \mathrm{P} \mapsto \mathbb{R}^n$ be a continuous function. Then for each $\theta \in [0, \frac{1}{2}]$ there is a stationary point of f on $A(\theta)$, since $A(\theta)$ is a nonempty convex compact set and f is a continuous function. For each $\theta \in [0, \frac{1}{2}]$, let $x(\theta)$ denote a stationary point of f on $A(\theta)$. Then by Lemma 2.4 there exists a minimal face $F(\theta; L(\theta))$ for some $L(\theta) = (L_1(\theta), \dots, L_k(\theta)) \in \mathcal{J}$ such that $x(\theta) \in F(\theta; L(\theta))$ and $f(x(\theta)) \in F^*(L(\theta))$. Note that $L_k(\theta) \in \mathcal{I}$ by definition. So there exist $\mu_l \geq 0$ for all $l \in L_k(\theta)$ and $\mu_l = 0$ for all $l \in I_m \setminus L_k(\theta)$ such that

$$f(x(\theta)) = \sum_{h \in I_m} \mu_h a^h.$$

We can choose $\omega \in (0, \Upsilon]$ so small that for every $\theta \in [0, \frac{1}{2}]$ it holds that

$$b_q - a^{q\top} x(\theta) \leq \theta(b_p - a^{p\top} x(\theta))$$

for all $q \in L_k(\theta)$ and all $p \in I_m \setminus L_k(\theta)$. Now we have the following lemma.

LEMMA 2.9. *Let $f : \mathrm{P} \mapsto \mathbb{R}^n$ be a continuous function. Then for each $\theta \in (0, \frac{1}{2}]$ a stationary point of f on $A(\theta)$ is a θ-robust stationary point of f.*

Proof. Since $A(\theta)$ is a nonempty convex compact set and f is a continuous function, f has a stationary point $z^* \in A(\theta)$, i.e.,

$$(z^* - x)^\top f(z^*) \geq 0$$

for all $x \in A(\theta)$. Then there exists a minimal face $F(\theta; L)$ with $L = (L_1, \dots, L_k) \in \mathcal{J}$ such that $z^* \in F(\theta; L)$ and $f(z^*) \in F^*(L)$. So there exist $\mu_1 \geq 0$, …, $\mu_k \geq 0$ such that

$$f(z^*) = \sum_{h=1}^{k} \mu_h a^{L_h}.$$

Without loss of generality we may assume that all $\mu_h > 0$. Let t_1, …, t_k be a sequence of increasing positive integers such that $L_h = \{i_{t_1}, \dots, i_{t_h}\}$ for each $h \in I_k$. Note that $t_k \leq n$. Let $L_0 = \emptyset$. For each $h \in I_k$ and $l \in L_h \setminus L_{h-1}$, let $\nu_l = \sum_{i=h}^{k} \mu_i$. Note that $|L_h| > 0$ for all $h \in I_k$ and $|L_p| < |L_q|$ with $1 \leq p < q \leq k$ by definition. Hence we have $\nu_p > \nu_q$ for all $p \in L_i$ and $q \in L_j$ with $1 \leq i < j \leq k$ and

$$f(z^*) = \sum_{h \in L_k} \nu_h a^h.$$

We will prove $b_p - a^{p\top} z^* \leq \theta(b_q - a^{q\top} z^*)$. It is sufficient to restrict to the case in which $p \in L_i$ and $q \in L_{i+1}$ with $1 \leq i \leq k - 1$. We have to consider the following cases.

Case 1. If $|L_i| + 1 = |L_{i+1}| = 2$, we have

$$\begin{array}{rcl} a^{p\top} z^* & = & b_p - \omega\theta^n, \\ (a^p + a^q)^\top z^* & = & b_p + b_q - \omega(\theta^n + \theta^{n-1}). \end{array}$$

It is easy to see that

$$b_p - a^{p\top} z^* \leq \theta(b^q - a^{q\top} z^*).$$

Case 2. If $1 = |L_i|$ and $|L_{i+1}| \geq 2$, then we have

$$\begin{array}{rcl} a^{p\top} z^* & = & b_p - \omega\theta^n, \\ (a^p + a^q)^\top z^* & \leq & b_p + b_q - \omega(\theta^n + \theta^{n-1}). \end{array}$$

It follows that

$$b_q - a^{q\top} z^* \geq \omega\theta^{n-1}.$$

Hence we have

$$b_p - a^{p\top} z^* \leq \theta(b^q - a^{q\top} z^*).$$

Case 3. If $1 < |L_i| + 1 = |L_{i+1}|$, then we have

$$\begin{array}{rcl} \sum_{h\in L_i\setminus\{p\}} a^{h\top} z^* & \leq & \sum_{h\in L_i\setminus\{p\}} b_h - \omega \sum_{h=n+2-|L_i|} \theta^h, \\ \sum_{h\in L_i} a^{h\top} z^* & = & \sum_{h\in L_i} b_h - \omega \sum_{h=n+1-|L_i|} \theta^h. \end{array}$$

It follows that

$$b_p - a^{p\top} z^* \leq \omega\theta^{n+1-|L_i|}.$$

On the other hand we have

$$b_q - a^{q\top} z^* = \omega\theta^{n-|L_i|}.$$

Hence

$$b_p - a^{p\top} z^* \leq \theta(b_q - a^{q\top} z^*).$$

Case 4. If $1 < |L_i| + 1 < |L_{i+1}|$, then we have

$$\begin{array}{rcl} \sum_{h\in L_i} a^{h\top} z^* & = & \sum_{h\in L_i} b_h - \omega \sum_{h=n+1-|L_i|}^{n} \theta^h, \\ \sum_{h\in L_i\setminus\{p\}} a^{h\top} z^* & \leq & \sum_{h\in L_i\setminus\{p\}} b_h - \omega \sum_{h=n+2-|L_i|}^{n} \theta^h. \end{array}$$

This implies

$$b_p - a^{p\top} z^* \leq \omega\theta^{n+1-|L_i|}. \tag{2.2}$$

On the other hand, we have

$$\begin{array}{rcl} \sum_{h\in L_i} a^{h\top} z^* & = & \sum_{h\in L_i} b_h - \omega \sum_{h=n+1-|L_i|}^{n} \theta^h, \\ \sum_{h\in L_i\cup\{q\}} a^{h\top} z^* & \leq & \sum_{h\in L_i\cup\{q\}} b_h - \omega \sum_{h=n+2-|L_i|}^{n} \theta^h. \end{array}$$

This implies

$$b_q - a^{q\top} z^* \geq \omega\theta^{n-|L_i|}. \tag{2.3}$$

Now it follows from (2.2) and (2.3) that

$$\frac{b_p - a^{p\top} z^*}{b_q - a^{q\top}} \leq \frac{\omega\theta^{n+1-|L_i|}}{\omega\theta^{n-|L_i|}} = \theta.$$

Hence

$$b_p - a^{p\top} z^* \leq \theta(b_q - a^{q\top} z^*).$$

Case 5. For any $q \in I_m \setminus L_k$, we have $\nu_q = 0$. Take any $p \in L_k$. It is clear that $\nu_p > \nu_q$. Now it follows from the choice of ω that

$$b_p - a^{p\top} z^* \leq \theta(b_q - a^{q\top} z^*).$$

We completed the proof. □

THEOREM 2.10. *Let $f : \mathrm{P} \mapsto \mathbb{R}^n$ be a continuous function. Then f has at least one robust stationary point.*

Proof. Let $\{\theta_k\}_1^\infty$ be a sequence of positive real numbers strictly between zero and $\frac{1}{2}$ converging to zero. According to Lemma 2.9 there exists a θ_k-robust stationary point $x(\theta_k)$ for each $k \in \mathsf{N}$. Since P is a compact set, there exists a subsequence out of $\{x(\theta_k)\}_1^\infty$ converging to a cluster point $x^* \in \mathrm{P}$. It is clear that x^* is a robust stationary point of f on P. □

In the subsequent sections we will develop an algorithm to find a robust stationary point of any continuous function. This also gives a constructive proof of Theorem 2.10.

3. A continuous refining triangulation of polytopes. In this section we introduce a triangulation of the polytope which underlies the algorithm. Let $v = (v_1, \ldots, v_n)^\top$ be any point in the interior of the polytope P. The point v will be the starting point of the algorithm. Take a sufficiently small $\omega \in (0, \Upsilon]$ such that v is contained in the interior of $A(\frac{1}{2})$.

We say that $I \in \mathcal{J}$ conforms to $J \in \mathcal{J}$ if it holds that every component of I is also a component of J. Let $\{\, \theta_k \mid k \in \mathsf{N}\}$ be a strictly decreasing sequence of real numbers in $(0, \frac{1}{2}]$ converging to zero. Let θ_0 be a constant bigger than θ_1. For $L = (L_1, \ldots, L_k) \in \mathcal{J}$, let

$$F(\theta_0, \theta_1; L) = \{\, x | x = av + (1-a)z \ \text{ for some } \ z \in F(\theta_1; L) \ \text{ and } \ a \in [0,1] \,\}$$

and for $k \in \mathsf{N}$

$$F(\theta_k, \theta_{k+1}; L) = \{\, x | x = ay + (1-a)z \ \text{ for some } y \in F(\theta_k; L), \\ z \in F(\theta_{k+1}; L), \text{ and } \ a \in [0,1] \,\}.$$

For $L \in \mathcal{J}$ and $l \in \mathsf{N}$, we denote the union of $F(\theta_{i-1}, \theta_i; L)$ over $i = 1, \ldots, l$ by $\mathcal{F}(\theta_l; L)$. For $l \in \mathsf{N}_0$, the union of $F(\theta_l, \theta_{l+1}; L)$ over all $L \in \mathcal{J}$ is denoted by $\mathcal{F}(\theta_l, \theta_{l+1})$. For $L \in \mathcal{J}$, we denote the union of $F(\theta_k, \theta_{k+1}; L)$ over all $k \in \mathsf{N}_0$ by $\mathcal{F}(L)$. Notice that the dimension of $\mathcal{F}(L)$ is equal to $t = n - k + 1$ for $L = (L_1, \ldots, L_k) \in \mathcal{J}$. The union of $F(\theta_l, \theta_{l+1}; L)$ over all $L \in \mathcal{J}$ and all $l \in \mathsf{N}_0$ is equal to the interior of the polytope P. The subdivision of $\mathrm{P} = C^2$ of Example 2.6 with $\theta_k = 2^{-k}$ for $k \in \mathsf{N}$, $\omega = \frac{1}{2}$, and $v = (1/2, 1/2)^\top$ is depicted in Figure 1. Note that in the figure we only draw the subdivision for $\theta_1 = \frac{1}{2}$ and $\theta_2 = \frac{1}{4}$.

A simplicial subdivision underlying the algorithm must be such that every set $F(\theta_k, \theta_{k+1}; L)$ is subdivided into t-dimensional simplices. Such a triangulation can be

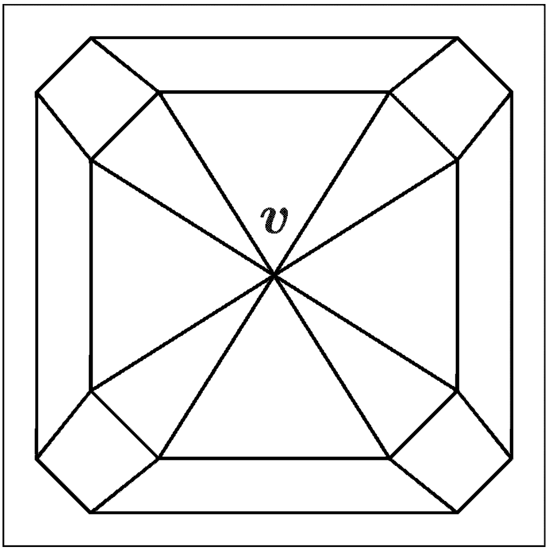


FIG. 1. *Subdivision of the interior of C^2.*

described as follows. For $L \in \mathcal{J}$, we denote $v(0,L) = v$, and for $l \in \mathsf{N}$, let $v(l,L)$ be a relative interior point (e.g., the barycenter) of $F(\theta_l; L)$. For $L \in \mathcal{J}$, if L consists of n components, then $F(\theta_l; L)$ is a vertex of $A(\theta_l)$. For general $L \in \mathcal{J}$, let $F(\theta_l; L(n))$ be a vertex of $F(\theta_l; L)$, i.e., $L(n)$ has n components and L conforms to $L(n)$. Moreover let $(L^1, L^2, \ldots, L^t) = \gamma(L, L(n))$ be a conformation between L and $L(n)$, i.e., $L^1 = L(n)$, $L^k \in \mathcal{J}$ for $k = 2, \ldots, t-1$, $L^t = L$, L^k conforms to L^{k-1} and has one component less than L^{k-1} for $k = 2, \ldots, t$. For given $k \in \mathsf{N}_0$, $L \in \mathcal{J}$, and $\gamma(L, L(n))$, the subset $F(\theta_k, \theta_{k+1}; L, \gamma(L, L(n)))$ of $F(\theta_k, \theta_{k+1}; L, \gamma(L, L(n)))$ of $F(\theta_k, \theta_{k+1}; L)$ is defined to be the convex hull of $v(k, L^1)$, $v(k, L^2)$, $\ldots$, $v(k, L^t)$, $v(k+1, L^1)$, $v(k+1, L^2)$, $\ldots$, and $v(k+1, L^t)$, so

$$F(\theta_k, \theta_{k+1}; L, \gamma(L, L(n))) = \left\{ x \in \mathrm{P} | x = v(k, L(n)) + \alpha q^0 + \sum_{j=1}^{t-1} \alpha_j q^j(\alpha), \right.$$

$$\left. 0 \leq \alpha \leq 1, \text{ and } 0 \leq \alpha_{t-1} \leq \cdots \leq \alpha_1 \leq 1 \right\},$$

where $q^0 = v(k+1, L(n)) - v(k, L(n))$, and for $j = 1, \ldots, t-1$, $0 \leq \alpha \leq 1$,

$$q^j(\alpha) = \alpha(v(k+1, J_{j+1}) - v(k+1, J_j)) + (1-\alpha)(v(k, J_{j+1}) - v(k, J_j)).$$

The dimension of $F(\theta_k, \theta_{k+1}; L, \gamma(L, L(n)))$ is equal to t, and $F(\theta_k, \theta_{k+1}; L)$ is the union of $F(\theta_k, \theta_{k+1}; L, \gamma(L, L(n)))$ over all conformations $\gamma(L, L(n))$ and over all index sets $L(n)$ conformed by L.

DEFINITION 3.1. *Let d be an arbitrary positive integer. For $k \in \mathsf{N}_0$, the set $G^d(k, k+1; L, \gamma(L, L(n)))$ is the collection of t-simplices $\sigma(a, \pi)$ with vertices y^1, $\ldots$, y^{t+1} in $F(\theta_k, \theta_{k+1}; L, \gamma(L, L(n)))$ such that*

(1) *$y^1 = v(k, L(n)) + a(0)d^{-1}q^0 + (a(0)+dk)^{-1} \sum_{j=1}^{t-1} a(j) q^j(d^{-1}a(0))$, where $a = (a(0), a(1), \ldots, a(n-1))^\top$ is a vector of integers such that $0 \leq a(0) \leq d-1$, and $a(n-1) = \cdots = a(t) = 0 \leq a(t-1) \leq \cdots \leq a(2) \leq a(1) \leq a(0) + dk$;*

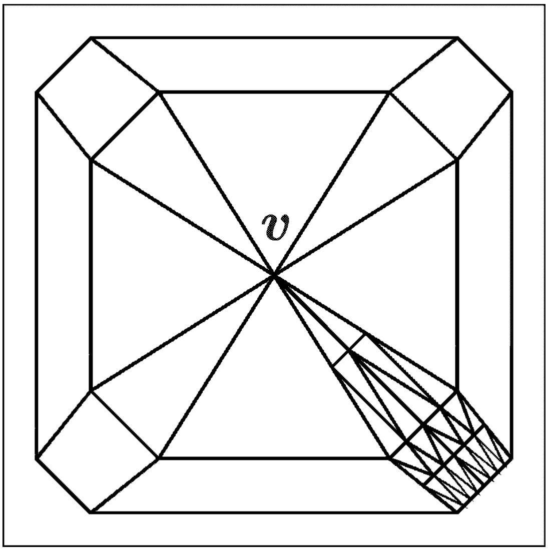


FIG. 2. *The P-triangulation of C^2 for $n = 2$.*

(2) *$\pi = (\pi_1, \ldots, \pi_t)$ is a permutation of $(0, 1, \ldots, t-1)$ such that $s < s'$ if for some $q \in \{1, \ldots, t-2\}$ it holds that $\pi_s = q$, $\pi_{s'} = q+1$, $a(q) = a(q+1)$ in case $q \geq 1$, and $a(0) + kd = a(1)$ in case $q = 0$;*

(3) *Let i be such that $\pi_i = 0$. Then*

$$\begin{aligned} y^{j+1} &= y^j + (a(0)+kd)^{-1} q^{\pi_j}(d^{-1}a(0)), \quad j = 1, \ldots, i-1, \\ y^{i+1} &= v(k, L(n)) + (a(0)+1)d^{-1}q^0 \\ &\quad + (a(0)+1+kd)^{-1} \sum_{j=1}^{t-1} a(j) q^j (d^{-1}(a(0)+1)) \\ &\quad + (a(0)+1+kd)^{-1} \sum_{j=1}^{i-1} q^{\pi_j}(d^{-1}(a(0)+1)), \\ y^{j+1} &= y^j + (a(0)+1+kd)^{-1} q^{\pi_j}(d^{-1}(a(0)+1)), \quad i < j \leq t. \end{aligned}$$

The set $G^d(k, k+1; L, \gamma(L, L(n)))$ is a simplicial subdivision of $F(\theta_k, \theta_{k+1}; L, \gamma(L, L(n)))$ with grid size d^{-1}. Moreover, the union $G^d(k, k+1; L)$ of $G^d(k, k+1; L, \gamma(L, L(n)))$ over all conformations $\gamma(L, L(n))$ and $L(n)$ conformed by L is a simplicial subdivision of $F(\theta_k, \theta_{k+1}; L)$. The union $G^d(k, k+1)$ of $G^d(k, k+1; L)$ over all sets $L \in \mathcal{J}$ induces a triangulation of $\mathcal{F}(\theta_k, \theta_{k+1})$. Taking the union $G^d(k)$ of $G^d(j, j+1)$ over $j = 0, 1, \ldots, k-1$, we obtain a simplicial subdivision of $A(\theta_k)$ with grid size d^{-1}. The union of $G^d(k)$ over all $k \in \mathsf{N}_0$ is a continuous refining simplicial subdivision of the interior of P and is called the P-triangulation of P. We remark that for $L \in \mathcal{J}$ the union $G^d(L)$ of $G^d(k, k+1; L)$ over $k = 0, 1, \ldots,$ is a simplicial subdivision of the set $\mathcal{F}(L)$. The P-triangulation of $\text{P} = C^2$ with $\theta_k = 2^{-k}$ for $k \in \mathsf{N}$, and $v = (1/2, 1/2)^\top$ with $\theta_k = 2^{-k}$ for $k \in \mathsf{N}$, and $v = (1/2, 1/2)^\top$ is illustrated in Figure 2. Note that in the figure we only draw a part of the P-triangulation.

As a norm we use the Euclidean norm $||\cdot||$ in $\mathbb{R}^n$. For a set B in $\mathbb{R}^n$, we define the diameter of B by

$$\mathrm{diam}(B) = \sup\{\, ||y^1 - y^2|| \mid y^1, y^2 \in B\,\}.$$

Then for given $k \in \mathsf{N}_0$ the mesh size of $G^d(k, k+1)$ is equal to

$$\delta_{k,d} = \sup\{\, \mathrm{diam}(\sigma) \mid \sigma \in G^d(k, k+1)\,\}.$$

Now we have the following observation.

LEMMA 3.2. *For the P-triangulation of* P *with grid size d^{-1}, it holds that*

$$\lim_{k\to\infty} \delta_{k,d} = 0.$$

The P-triangulation has the following property that the diameter of the simplices converges to zero when the boundary of P is approached.

4. The path of the algorithm. Now we discuss operation of the algorithm in the P-triangulation of the polytope P to approximate a robust stationary point of a continuous function f on P. Starting at the point v, the algorithm will generate a sequence of adjacent simplices of the P-triangulation in the set $\mathcal{F}(L)$ having L-complete common facets for varying $L \in \mathcal{J}$.

DEFINITION 4.1. *Let $f : \mathrm{P} \mapsto \mathbb{R}^n$ be a continuous function. For given $L = (L_1, \ldots, L_k) \in \mathcal{J}$ and $s = t$ or $t-1$, where $t = n-k+1$, an s-simplex σ with vertices $y^1, \ldots, y^{s+1}$ is L-complete if the system of linear equations*

$$\sum_{i=1}^{s+1} \lambda_i \begin{pmatrix} f(y^i) \\ 1 \end{pmatrix} - \sum_{j=1}^{k} \mu_j \begin{pmatrix} a^{L_j} \\ 0 \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix} \tag{4.1}$$

has a solution λ_i^*, $i \in I_{s+1}$, μ_j^*, $j \in I_k$, satisfying $\lambda_i^* \geq 0$, $i \in I_{s+1}$, $\mu_j^* \geq 0$, $j \in I_k$.

Notice that the system (4.1) has $s+1+k$ columns, so when $s = t-1$, the system has $n+1$ columns, and for $s = t$ one column more. A solution λ_i^*, $i \in I_{s+1}$, μ_j^*, $j \in I_k$, will be denoted by (λ^*, μ^*).

Nondegeneracy assumption For $s = t-1$ the system (4.1) has a unique solution (λ^*, μ^*) with $\lambda_i^* > 0$, $i \in I_t$, and $\mu_j^* > 0$, $j \in I_k$, and for $s = t$ at most one component of (λ^*, μ^*) is equal to zero. We remark that this assumption can easily be dropped if we use the lexicographic pivoting method to solve system (4.1); see e.g., Todd [18] or Yang [22].

Under this nondegeneracy assumption $\sigma^0 = \{\, v\,\}$ is L^0-complete with $L^0 = (L_1^0, \ldots, L_n^0) \in \mathcal{J}$, where $L_h^0 = \{i_1, \ldots, i_h\}$ for $h \in I_n$ such that the system of linear equations

$$f(v) = \sum_{h=1}^{n} \mu_h a^{L_h^0}$$

has solutions $\mu_h > 0$ for all $h \in I_n$.

The algorithm now starts with σ^0 for $L = L^0$ and follows a sequence of adjacent t-simplices in $\mathcal{F}(L)$ for varying L, $L \in \mathcal{J}$, such that their common facets are L-complete. In this way within a finite number of steps either the algorithm reaches a point $\bar{x}$ in an n-dimensional simplex for which $\bar{f}_j(\bar{x}) = 0$ for every $j \in I_n$, where $\bar{f}$ is the PL approximation of f with respect to the P-triangulation, or for $k = 1, 2, \ldots,$ the algorithm finds a $J(k)$-complete simplex in $F(\theta_k; J(k))$ for some $J(k) \in \mathcal{J}$. Let $\{ \theta_t \mid t \in \mathsf{N} \}$ be given as in section 3. Now we have the following result.

Lemma 4.2. *For some $l \in \mathsf{N}$ and $L = (L_1, \ldots, L_k) \in \mathcal{J}$, let σ with vertices $y^1, \ldots, y^t$ be an L-complete $(t-1)$-simplex lying in $F(\theta_l; L)$. Let (λ^*, μ^*) be the corresponding unique solution of system (4.1). Then $x = \sum_{i=1}^t \lambda_i^* y^i$ is a θ_l-robust stationary point of the PL approximation $\bar{f}$ of f with respect to the P-triangulation. Moreover, x is a stationary point of $\bar{f}$ on $A(\theta_l)$.*

Proof. It follows from (4.1) that at $x = \sum_{i=1}^t \lambda_i^* y^i$

$$\bar{f}(x) = \sum_{h=1}^{k} \mu_h^* a^{L_h},$$

where $\mu_h^* > 0$ for $h = 1, \ldots, k$. So $x \in F(\theta_l; L)$ and $\bar{f}(x) \in F^*(L)$. According to Lemma 2.4 this implies that x is a stationary point of $\bar{f}$ on $A(\theta_l)$. By applying Lemma 2.9 we know that x is a θ_l-robust stationary point of $\bar{f}$. □

The next lemma shows that a θ_l-robust stationary point of $\bar{f}$ is an approximate θ_l-robust stationary point of f.

Lemma 4.3. *Let $\eta_{l,d} = \sup\{ \operatorname{diam}(f(\sigma)) \mid \sigma \in G^d(l-1, l) \}$. Let x be a θ_l-robust stationary point of the PL approximation $\bar{f}$ of f with respect to the P-triangulation with grid size d^{-1} obtained by the algorithm, so that $x \in F(\theta_l; L)$ for some $L \in \mathcal{J}$. Then $f(x)$ lies in the $\eta_{l,d}$-neighborhood of $F^*(L)$, i.e., there is a $y \in F^*(L)$ such that $||y - f(x)|| \leq \eta_{l,d}$.*

Proof. Let $y^1, \ldots, y^t$ be the vertices of a $(t-1)$-simplex of $G^d(l-1, l)$ in $F(\theta_l; L)$ containing x. Then $\bar{f} = \sum_{j=1}^t \lambda_j^* f(y^j)$ lies in $F^*(L)$, where $\lambda_1^*, \ldots, \lambda_t^*$ are convex combination coefficients such that $x = \sum_{j=1}^t \lambda_j^* y^j$. Hence we have

$$\begin{aligned} ||\bar{f}(x) - f(x)|| &= ||\textstyle\sum_{j=1}^t \lambda_j^* f(y^j) - f(x)|| \\ &= ||\textstyle\sum_{j=1}^t \lambda_j^* (f(y^j) - f(x))|| \\ &\leq \textstyle\sum_{j=1}^t \lambda_j^* ||f(y^j) - f(x)|| \\ &\leq \eta_{l,d}. \quad \square \end{aligned}$$

Given a function $f : \mathrm{P} \mapsto \mathbb{R}^n$, let us define a function $g : \mathrm{P} \mapsto \mathrm{P}$ by

$$g(x) = argmin\{(x + f(x) - y)^\top (x + f(x) - y) \mid y \in \mathrm{P}\}.$$

Now we have the following well-known result (see Hartman and Stampacchia [9] and Eaves [4]).

Theorem 4.4. *Let $f : \mathrm{P} \mapsto \mathbb{R}^n$ be a continuous function. Then x^* is a stationary point of f on* P *if and only if x^* is a fixed point of g on* P.

There are several ways of proving the existence of a stationary point in the above theorem. For example, one way is to use Brouwer's fixed point theorem. Of course, we could also use degree theory, as we could use degree theory to prove Brouwer's theorem.

We are now going to discuss the convergence properties of the algorithm. First we consider the case in which the algorithm converges to a boundary point of P. Since P is compact and f is continuous on P, the error $\eta_{l,d}$ tends to zero as the mesh size $\delta_{l,d}$ goes to zero when l goes to infinity. Let x^l be a θ_l-robust stationary point of $\bar{f}$

and $\eta_{l,d}$ the error in Lemma 4.3. Suppose that the algorithm generates the sequence $\{\, x^h \mid h \in \mathsf{N} \,\}$ of approximate θ_l-robust stationary points of f which therefore has a cluster point x^*. For simplicity of notation we can assume that this sequence itself converges to x^*. We are now ready to state the following theorem.

THEOREM 4.5. *Let* $\{\, x^l \mid l \in \mathsf{N} \,\}$ *be a sequence generated by the algorithm with* $x^l \in F(\theta_l; J(l))$ *for each* $l \in \mathsf{N}$. *Then the sequence* $\{\, x^l \mid l \in \mathsf{N} \,\}$ *has a cluster point* x^* *which is a robust stationary point of* f *on* P.

Proof. To prove the theorem, we first extend the domain of the PL approximation $\bar{f}$ of f. Recall from Lemma 3.2 that for a given positive integer d, the mesh size $\delta_{l,d}$ converges to zero as l goes to infinity. We can take $\bar{f}(x)$ to be $f(x)$ if x lies on the boundary of P, since f is a continuous function. Hence $\bar{f}$ is also a continuous function on P. From Lemma 4.2 we know that for each $l \in \mathsf{N}$, x^l is a θ_l-robust stationary point of $\bar{f}$. By definition x^* is a robust stationary point of $\bar{f}$ on P. Now we are going to show that for any given $\epsilon > 0$, there exists a positive integer M such that for $l \in \mathsf{N}$ with $l > M$ there is a θ_l-robust stationary point $y^l \in A(\theta_l)$ of f on P which is in the ϵ-neighborhood of x^l.

Let

$$U(\theta) = \begin{cases} A(\theta) & \text{for } \theta \in [0, \theta_1], \\ A(\theta_1)(\theta - 1)/(\theta_1 - 1) + \{\, v \,\}(\theta - \theta_1)/(1 - \theta_1) & \text{for } \theta \in [\theta_1, 1]. \end{cases}$$

Observe that $U(\theta)$ is a convex and compact set contained in $A(\theta_1)$ for any $\theta \in [\theta_1, 1]$. As θ decreases from 1 to 0, the set $U(\theta)$ first expands from the starting point v to the set $A(\theta_1)$ and then to the whole set P. Let $Y(\theta)$ denote the set of stationary points of f on $U(\theta)$ for $\theta \in [0, 1]$. Notice that for each $\theta \in (0, \theta_1]$, x is a θ-robust stationary point if x belongs to the set $Y(\theta)$.

For each $\theta \in [0, 1]$, define a function $g_\theta : \mathrm{P} \mapsto U(\theta)$ by

$$g_\theta(x) = argmin\{\, (x + f(x) - y)^\top (x + f(x) - y) \mid y \in U(\theta) \,\}.$$

Since $U(\theta)$ is a convex and compact set and f is continuous, the function g_θ is a continuous function. By Theorem 4.4, $x \in Y(\theta)$ if and only if $x = g_\theta(x)$. Define a homotopy function $H : \mathrm{P} \times [0, 1] \mapsto \mathbb{R}^n$ by

$$H(x, \theta) = x - g_\theta(x).$$

The function H is also a continuous function. Now let

$$H^{-1}(\mathbf{0}) = \{\, (x, \theta) \in \mathrm{P} \times [0, 1] \mid H(x, \theta) = \mathbf{0} \,\}.$$

Let $Z(\theta)$ denote the set of stationary points of $\bar{f}$ on $U(\theta)$ for each $\theta \in [0, 1]$. Similarly, a continuous function G with respect to $\bar{f}$ can be defined as

$$G : \mathrm{P} \times [0, 1] \mapsto \mathbb{R}^n$$

such that

$$Z(\theta) = \{\, x \in \mathrm{P} \mid G(x, \theta) = \mathbf{0} \,\},\ \theta \in [0, 1].$$

Let

$$G^{-1}(\mathbf{0}) = \{\, (x, \theta) \in \mathrm{P} \times [0, 1] \mid G(x, \theta) = \mathbf{0} \,\}.$$

For each $l \in \mathsf{N}$, let

$$\xi^l = (x^l, \theta_l).$$

It is clear that $\lim_{l\to\infty} \xi^l = \xi^* = (x^*, 0)$. Define

$$N(\epsilon) = \{\, (x, \theta) \in \mathrm{P} \times [0,1] \ \mid\ ||(x,\theta) - (z,\alpha)|| < \epsilon \ \text{ for some } \ (z,\alpha) \in H^{-1}(\mathbf{0})\}.$$

This implies that

$$||H(\psi)|| > 0 \ \text{ for any } \ \psi \in \mathrm{P} \times [0,1] \setminus N(\epsilon).$$

Notice that $N(\epsilon)$ is open, so the set $\mathrm{P} \times [0,1] \setminus N(\epsilon)$ is compact. It follows from the compactness that the minimum can be attained. There exists $\nu > 0$ such that

$$\min\{\, ||H(\psi)|| \ \mid\ \psi \in \mathrm{P} \times [0,1] \setminus N(\epsilon) \,\} > \nu.$$

This implies that if a point $\psi \in \mathrm{P} \times [0,1]$ satisfies

$$||H(\psi)|| \leq \nu, \tag{4.2}$$

then ψ must be in $N(\epsilon)$. Because H and G are uniformly continuous on $\mathrm{P} \times [0,1]$ and $\bar{f}$ is the LP approximation of f with respect to the P-triangulation, we can prove that

$$||H(\psi) - G(\psi)|| < \epsilon \tag{4.3}$$

for any $\psi = (x, \theta) \in \mathrm{P} \times [0,1]$ under the condition that the diameter of the simplices in which x lies is small enough, say, smaller than $\Delta > 0$. Given any $\epsilon > 0$, because of $\bar{f}$ being a PL approximation of f, it holds that

$$||f(x) - \bar{f}(x)|| < \epsilon,$$

for any $x \in \mathrm{P}$ when the diameter of the simplices in which x lies, is small enough. For any given $x \in \mathrm{P}$ and any $\theta \in [0,1]$, define a function $h_\theta : \mathbb{R}^n \mapsto U(\theta)$ by

$$h_\theta(x + z) = argmin\{(x + z - y)^\top (x + z - y) \mid y \in U(\theta)\}.$$

It is clear that h_θ is a Lipschitz continuous function. Furthermore, it is easy to see that for all $p, q \in \mathbb{R}^n$

$$||h_\theta(x + p) - h_\theta(x + q)|| \leq ||p - q||.$$

Then it follows immediately that

$$\begin{array}{rcl} ||H(\psi) - G(\psi)|| & = & ||h_\theta(x + f(x)) - h_\theta(x + \bar{f}(x))|| \\ & \leq & ||f(x) - \bar{f}(x)|| \\ & < & \epsilon. \end{array}$$

Lemma 3.2 states that given a positive integer d as l goes to infinity, the mesh size $\delta_{l,d}$ converges to zero. This implies that there exists a positive integer M such that for every $l \in \mathsf{N}$ with $l > M$, it holds that

$$\delta_{l,d} < \Delta.$$

Since for any $l \in \mathsf{N}$ with $l > M$, $\psi^l \in G^{-1}(\mathbf{0})$, i.e., $G(\psi^l) = \mathbf{0}$, it follows from (4.3) that

$$||H(\psi^l)|| < \epsilon.$$

By (4.2) ψ^l must be in $N(\epsilon)$. This implies that for any $l \in \mathsf{N}$ with $l > M$ there is $\psi^l \in H^{-1}(\mathbf{0})$ which is in the ϵ-neighborhood of ξ^l. Without loss of generality we may assume that $\psi^l = (y^l, \theta_l)$. This is what we claimed.

Now let us take a sequence of positive real numbers $\{\, \epsilon_l \mid l \in \mathsf{N} \,\}$ with limit zero. Then for any l there exists a y^{i_l} being a θ_{i_l}-robust stationary point of f with $||x^{i_l} - y^{i_l}|| \le \epsilon_l$. Since $\lim_{l\to\infty} x^{i_l} = x^*$ and $\lim_{l\to\infty} \epsilon_l = 0$, we have

$$\lim_{l\to\infty} y^{i_l} = x^*.$$

Hence x^* is a robust stationary point of f on P. □

Now we consider the case in which the algorithm converges to a simplex in the interior of the polytope. This means that the algorithm terminates with an n-dimensional simplex σ with vertices $y^1, \ldots, y^{n+1}$ within a finite number of steps. Let $\bar{x} = \sum_{i=1}^{n+1} \lambda_i^* y^i$. Then for each $k \in I_n$ it holds that $\bar{f}_k(\bar{x}) = 0$. It is clear that $\bar{x}$ is a robust stationary point of $\bar{f}$ on P. If the accuracy of approximation is not satisfactory, the algorithm can be restarted at the point $\bar{x}$ with a smaller grid size d^{-1} to find a better approximate robust stationary point, hopefully within a small number of steps. In this case we may assume that the algorithm generates a sequence $\{\, \bar{x}^h \mid h \in \mathsf{N} \,\}$, where $\bar{x}^h$ is the robust stationary point of $\bar{f}$ on P corresponding to the grid size d_h^{-1} for a strictly increasing sequence of positive integers $\{\, d_h \mid h \in \mathsf{N} \,\}$. It is readily seen that for every $k \in \mathsf{N}_0$, the mesh size δ_{k,d_h} tends to zero when h goes to infinity. Therefore the sequence $\{\, \bar{x}^h \mid h \in \mathsf{N} \,\}$ has a subsequence converging to a point being a robust stationary point of f on P. For a subset B of R^n, int(B) denotes the relative interior of B. We now have the following corollary.

COROLLARY 4.6. *Let $\bar{x}^h \in int(\mathrm{P})$ be the robust stationary point of $\bar{f}$ generated by the algorithm for the P-triangulation with grid size d_h^{-1} for $h \in \mathsf{N}$. Suppose that $\{d^h \mid h \in \mathsf{N}\}$ is a strictly increasing sequence of positive integers. Then the sequence $\{\bar{x}^h \mid h \in \mathsf{N}\}$ has a cluster point which is a robust stationary point of f on* P.

5. The steps of the algorithm. As described in the last section, starting at the point v, the algorithm will generate a unique PL path leading to an approximate solution by making alternating LP pivot steps in system (4.1) and replacement steps in the underlying triangulation. When, with respect to some simplex $\sigma(a, \pi)$ with vertices $y^1, \ldots, y^{t+1}$ in some $G^d(l, l+1; L, \gamma(L, L(n)))$ for some $l \in \mathsf{N}_0$ and $\gamma(L, L(n))$, the variable λ_q, for some q, $1 \le q \le t+1$, becomes zero through an LP pivot step in (4.1), then the replacement step determines the unique t-simplex $\bar{\sigma}(\bar{a}, \bar{\pi})$ in $F(\theta_l, \theta_{l+1}; L, \gamma(L, L(n)))$ sharing with σ a common facet τ opposite the vertex y^q, unless this facet lies on the boundary of the set $F(\theta_l, \theta_{l+1}; L, \gamma(L, L(n)))$. If τ does not lie on the boundary of $F(\theta_l, \theta_{l+1}; L, \gamma(L, L(n)))$, then $\bar{\sigma}(\bar{a}, \bar{\pi})$ can be obtained from a and π as given in Table 1, where $E(i-1)$ is the ith unit vector in $\mathbb{R}^n$ for $i \in I_n$.

The algorithm continues with $\bar{\sigma}$ by making an LP pivot step in (4.1) with $(f^\top(\bar{y}), 1)^\top$, where $\bar{y}$ is the vertex of $\bar{\sigma}$ opposite the facet τ. When the L-complete facet τ of the simplex $\sigma(a, \pi)$ in $G^d(l, l+1; L, \gamma(L, L(n)))$ is not a facet of another t-simplex in $G^d(l, l+1; L, \gamma(L, L(n)))$, then τ lies on the boundary of $F(\theta_l, \theta_{l+1}; L, \gamma(L, L(n)))$. According to Definition 3.1 we have the following lemma.

Table 1

Parameters of $\bar\sigma$ if the vertex y^q of $\sigma(a,\pi)$ is replaced.

	$\bar\pi$	$\bar a$
$q = 1$	$(\pi_2, ..., \pi_t, \pi_1)$	$a + E(\pi_1)$
$1 < q < t+1$	$(\pi_1, ..., \pi_{q-2}, \pi_q, \pi_{q-1}, \pi_{q+1} ..., \pi_t)$	a
$q = t+1$	$(\pi_t, \pi_1, ..., \pi_{t-1})$	$a - E(\pi_t)$

Lemma 5.1. *Let $\sigma(a,\pi)$ be a t-simplex in $F(\theta_l, \theta_{l+1}; L, \gamma(L, L(n)))$. The facet τ of σ opposite the vertex y^q, $1 \leq q \leq t+1$, lies on the boundary of this set if and only if one of the following cases occurs:*

(i) *$q = 1$, $\pi_1 = 0$, and $a(0) = d - 1$;*

(ii) *$1 < q < t+1$, $\pi_q = h+1$, $\pi_{q-1} = h$ for some $h \in \{0\} \cup I_{t-2}$, and $a(h) = a(h+1)$ in case $h \geq 1$, and $a(0) + kd = a(1)$ in case $h = 0$;*

(iii) *$q = t+1$, $\pi_t = 0$, and $a(0) = 0$;*

(iv) *$q = t+1$, $\pi_t = t-1$, and $a(t-1) = 0$.*

Suppose that the algorithm generates the simplex $\sigma(a,\pi)$ as given in Lemma 5.1 and λ_q becomes zero after making an LP pivot step in (4.1). Then the facet τ of σ opposite the vertex y^q is L-complete. In case (i) the facet τ lies in the face $F(\theta_{l+1}; L)$ of $A(\theta_{l+1})$ and the algorithm reaches a θ_{l+1}-robust stationary point $\bar x = \sum_{i=1}^{t+1} \lambda_i y^i$ of $\bar f$ lying in $F(\theta_{l+1}; L)$. If l is large enough, then $\bar x$ is an approximate robust stationary point of f. Otherwise, the algorithm continues with $\bar\sigma$ by making an LP pivot step in (4.1) with $(f^\top(\bar y), 1)^\top$, where $\bar y$ is the vertex of $\bar\sigma$ opposite the facet τ and $\bar\sigma$ in $F(\theta_{l+1}, \theta_{l+2}; L, \gamma(L, L(n)))$ is obtained according to Table 1.

In case (ii), and if $h = 0$, τ is a facet of the t-simplex $\bar\sigma = \sigma(a,\pi)$ in $F(\theta_l, \theta_{l+1}; L, \bar\gamma(L, \bar L(n)))$ with $\bar L(n)$ and $\bar\gamma$ defined as follows. Let $L^1 = L(n) = (L_1, \ldots, L_n)$. When $L^2 = (L_1, \ldots, L_{n-1})$, we have $\bar L(n) = (L_1, \ldots, L_{n-1}, \bar L_n)$ with $\bar L_n = L_{n-2} \cup (L(n) \setminus L_{n-1})$. When $L^2 = (L_2, \ldots, L_n)$, let $\bar L(n) = (\bar L_1, L_2, \ldots, L_n)$ with $\bar L_1 = L_2 \setminus L_1$. Finally, if $L^2 = (L_1, \ldots, L_i, L_{i+2}, \ldots, L_n)$ for some $i \in I_{n-3}$, we have $\bar L(n) = (L_1, \ldots, L_i, \bar L_{i+1}, L_{i+2}, \ldots, L_n)$ with $\bar L_{i+1} = L_i \cup (L_{i+2} \setminus L_{i+1})$. Then $\bar\gamma(L, \bar L(n)) = (\bar L(n), L^2, \ldots, L^t)$. In case (i), and if $h \geq 1$, the facet τ is a facet of the t-simplex $\bar\sigma = \sigma(a,\pi)$ in $F(\theta_l, \theta_{l+1}; L)$ lying in the subset $F(\theta_l, \theta_{l+1}; L, \bar\gamma(L, L(n)))$ with

$$\bar\gamma(L, L(n)) = (L^1, \ldots, L^h, \bar L^{h+1}, L^{h+2}, \ldots, L^t),$$

where $\bar L^{h+1} \in \mathcal{J}$, $\bar L^{h+1} \neq L^{h+1}$, is uniquely determined by the properties that $\bar L^{h+1}$ conforms to L^h, has one component less than L^h, and is conformed by L^{h+2}. In both subcases of case (ii) the algorithm proceeds by making an LP pivot step in (4.1) with $(f^\top(\bar y), 1)^\top$, where $\bar y$ is the vertex of the new t-simplex $\bar\sigma$ opposite the facet τ.

In case (iii) the facet τ lies in the face $F(\theta_l; L)$ of $A(\theta_l)$ and the algorithm continues with $\bar\sigma$ by making an LP pivot step in (4.1) with $(f^\top(\bar y), 1)^\top$, where $\bar y$ is the vertex $\bar\sigma$ opposite the vertex τ and σ in $F(\theta_{l-1}, \theta_l; L, \gamma(L, L(n)))$ is obtained from Table 1.

In case (iv) the facet lies in the set $F(\theta_l, \theta_{l+1}; L^{t-1})$ of $\mathcal{F}(L)$. More precisely, τ is the $(t-1)$-simplex $\sigma(a, \bar\pi)$ in $F(\theta_l, \theta_{l+1}; \bar L, \bar\gamma(\bar L, L(n)))$, where $\bar L = L^{t-1}$, $\bar\gamma(\bar L, L(n)) = (L^1, \ldots, L^{t-1})$. The algorithm now continues by making an LP pivot step in (4.1) with $(-(a^{L_h})^\top, 0)^\top$, where L_h is the unique component of L^{t-1} but not in L^t.

Finally, if, through an LP pivot step in (4.1), μ_i becomes zero for some $i \in I_k$ and $k = 1$, then the algorithm terminates with the approximate robust stationary point $\bar x = \sum_{j=1}^{n+1} \lambda_j y^i$. In case the accuracy is not satisfactory, the algorithm can restart at the point $\bar x$. When $k > 1$, then the simplex $\sigma(a,\pi)$ is a facet of a unique $(t+1)$-simplex

$\bar{\sigma}$ in $\mathcal{F}(\bar{L})$ with $\bar{L} = (L_1, \ldots, L_{i-1}, L_{i+1}, \ldots, L_k)$. To be precise, $\bar{\sigma} = \sigma(a, \bar{\pi})$ lies in $F(\theta_l, \theta_{l+1}; \bar{L}, \bar{\gamma}(\bar{L}, L(n)))$, where $\bar{\pi} = (\pi_1, \ldots, \pi_t, t)$ and $\bar{\gamma}(\bar{L}, L(n)) = (L^1, \ldots, L^t, \bar{L})$. The algorithm now proceeds by making an LP pivot step in (4.1) with $(f^\top(\bar{y}), 1)^\top$, where $\bar{y}$ is the vertex of σ opposite the facet σ.

This completes the description of how to follow a sequence of adjacent simplices of varying dimension in the P-triangulation of the polytope P.

6. Special cases and applications to games and economics. We now turn to discussing the cases when the polytope P is not full-dimensional. Moreover we also show the applications of the concept of robust stationary point in game theory and economic equilibrium theory. First we discuss the cases when the polytope P is not full-dimensional. If P is a lower-dimensional polytope we may assume that P can be described as

$$\mathrm{P} = \{x \in \mathbb{R}^n \mid {a^i}^\top x \le b_i,\ i \in I_m, \quad \text{and} \quad {c^j}^\top x = d_j,\ j \in I_{m^1}\}$$

and P is simple and no constraints are redundant, while $\dim(P) = n - m^1$, for some m^1, $0 \le m^1 \le n$. In this case for each $I \in \mathcal{I}$, the cone $F^*(I)$ is defined by

$$F^*(I) = \left\{ x \in \mathbb{R}^n \mid x = \sum_{i \in I} \nu_h a^i + \sum_{j \in I_{m^1}} \mu_j c^j,\ \nu_i \ge 0 \ \text{ for } \ i \in I,\ \mu_j \in \mathbb{R} \ \text{ for } j \in I_{m^1} \right\},$$

where $F(I)$ is a nonempty face of P and defined as

$$F(I) = \{x \in \mathrm{P} \mid {a^i}^\top x = b_i \quad \text{for} \quad i \in I\}.$$

The dimension of such a face $F(I)$ is equal to $n - m^1 - |I|$. Now the definition of robust stationary points is the same as in Definition 2.3 except that the definition of θ-robust stationary points is adapted as follows.

DEFINITION 6.1. *For given $\theta > 0$ a point $x \in \mathrm{P}$ is a θ-robust stationary point of f if*

(1) *x is a relative interior point of* P*;*

(2) *for some $I \in \mathcal{I}$, $f(x) = \sum_{h \in I_m} \mu_h a^h + \sum_{j \in I_{m^1}} \nu_j c^j$ with $\mu_h \ge 0$ for all $h \in I$ and $\mu_h = 0$ for all $h \in I_m \setminus I$, and $\nu_j \in \mathbb{R}$ for all $j \in I_{m^1}$, if $\mu_l > \mu_k$, then $b_l - {a^l}^\top x \le \theta(b_k - {a^k}^\top x)$.*

We leave it to the reader to prove the existence of a robust stationary point for any continuous function on a lower-dimensional polytope P in $\mathbb{R}^n$. Now the algorithm is the same as described in sections 4 and 5 except that the LP pivot steps are made in the following system for given $L = (L_1, \ldots, L_k) \in \mathcal{J}$:

$$\sum_{i=1}^{s+1} \lambda_i f(y^i) - \sum_{j=1}^{k} \mu_j a^{L_j} + \sum_{h=1}^{m^1} \nu_h c^h = \mathbf{0},$$

$$\sum_{i=1}^{s+1} \lambda_i = 1,$$

$$\lambda_i \ge 0,\ i \in I_{s+1};\ \mu_j \ge 0,\ j \in I_k;\ \nu_h \in \mathbb{R},\ h \in I_{m^1},$$

where $s = t$ or $t - 1$ with $t = n - m^1 - k + 1$.

It should be noted that when we study the stability and refinements of Nash equilibria or Walrasian equilibria, we often have to deal with problems over lower-dimensional polytopes.

Very special cases of the polytope P are the $(n-1)$-dimensional unit simplex $S^n = \{x \in \mathbb{R}^n_+ \mid \sum_{i=1}^n x_i = 1\}$ or the simplotope. We rewrite $S^n = \{x \in \mathbb{R}^n \mid {a^i}^\top x \le$

TABLE 2
Payoffs of the game.

Player 2

Player 1

	ϕ_1	ϕ_2	ϕ_3
ψ_1	$(1,1)$	$(0,0)$	$(-9,-9)$
ψ_2	$(0,0)$	$(0,0)$	$(-7,-7)$
ψ_3	$(-9,-9)$	$(-7,-7)$	$(-7,-7)$

$0, i \in I_n,$ and $\sum_{i=1}^{n} x_i = 1\}$, where $a^i = -e(i)$ for all $i \in I_n$. In case the polytope is the unit simplex S^n, Definition 6.1 coincides with Definition 2.1 in Yang [21]. The simplotope is the Cartesian product of, say, n, $n \in \mathsf{N}$, unit simplices S^{n_j}, $n_j \in \mathsf{N}$ and $j \in I_n$. In case the polytope is the simplotope, then the concept of robust stationary point is reduced to that in Talman and Yang [17]. We recall that concept here. The definition of robust stationary points is the same as in Definition 2.3, but the form of θ-robust stationary points becomes much simpler. Let $\mathcal{S} = \prod_{h=1}^{n} S^{n_h}$, where S^{n_h} is the $(n_h - 1)$-dimensional unit simplex for each $h \in I_n$.

DEFINITION 6.2. *Let $f : \mathcal{S} \mapsto \prod_{h=1}^{n} \mathbb{R}^{n_h}$ be a function. For given $\theta > 0$ a point $x \in \mathcal{S}$ is a θ-robust stationary point of f if*

(1) *x is a relative interior point of $\mathcal{S}$;*

(2) *$f_{h,i}(x) < f_{h,j}(x)$ implies $x_{h,i} \leq \theta x_{h,j}$ for $i, j \in I_{n_h}$ and $h \in I_n$.*

Furthermore, when the simplotope is the strategy space of a noncooperative finite n-person game, the concept of robust stationary point coincides with the well-known concept of proper Nash equilibrium in Myerson [13] provided that the function on the simplotope is defined as the marginal expected payoff of the game. We suggest the interested reader see Myerson [13] and van Damme [3] for the game-theoretic interpretation of the above concept. Let us illustrate this by a well-known example of Myerson [13].

EXAMPLE 6.3. *We consider a bimatrix game with two players. Each player has three pure strategies. The payoffs are given in Table* 2.

This game has three Nash equilibria: (ψ_1, ϕ_1), (ψ_2, ϕ_2), and (ψ_3, ϕ_3). Among these equilibria, (ψ_1, ϕ_1) is the only proper Nash equilibrium. The marginal expected payoff function is given by $f : S^3 \times S^3 \mapsto \mathbb{R}^3 \times \mathbb{R}^3$ with

$$f(x) = (f_{1,1}(x), f_{1,2}(x), f_{1,3}(x); f_{2,1}(x), f_{2,2}(x), f_{2,3}(x))^\top,$$

where

$$\begin{array}{rcl} f_{1,1}(x) & = & x_{2,1} - 9x_{2,3}, \\ f_{1,2}(x) & = & -7x_{2,3}, \\ f_{1,3}(x) & = & -9x_{2,1} - 7x_{2,2} - 7x_{2,3}, \\ f_{2,1}(x) & = & x_{1,1} - 9x_{1,3}, \\ f_{2,2}(x) & = & -7x_{1,3}, \\ f_{2,3}(x) & = & -9x_{1,1} - 7x_{1,2} - 7x_{1,3}. \end{array}$$

This function has three stationary points: $(1,0,0;1,0,0)^\top$, $(0,1,0;0,1,0)^\top$, and $(0,0,1;0,0,1)^\top$, corresponding to the three Nash equilibria given above, respectively. The point $(1,0,0;1,0,0)^\top$ is the only robust stationary point which corresponds to the only proper Nash equilibrium.

Finally let us apply the concept of robust stationary point to the standard exchange economy model with linear production. For detailed discussions, we refer to

Scarf [15] and Koopmans [10]. In such an economy, there are, say, n commodities, a finite number of production activities or firms, and a finite number of consumers, each of whom initially has a certain amount of commodities. Exchange of commodities is based on relative prices. All agents exchange their commodities to maximize their utility under their budget constraints, and all firms run their activities to achieve their maximal profits. This economy can be captured by an excess demand function $z : S^n \mapsto \mathbb{R}^n$ and an $n \times (n+k)$ matrix $A = [a^1, \ldots, a^{k+n}] = [a^1, \ldots, a^k, -e(1), \ldots, -e(n)]$, which satisfies the following standard conditions:

(i) z is continuous;

(ii) $p^\top z(p) = 0$ (Walras' law);

(iii) $Ay \geq \mathbf{0}$ and $y \in \mathbb{R}^{n+k}_+$ imply $y_i = 0$ for all $i \in I_{n+k}$. (No production without input.)

Now we explain the matrix A in some detail. For each $j \in I_k$, the jth column a^j of A represents an activity. It says that if activity j is operated at level one, $|a^j_i|$ units of commodity i are supplied as output for $a^j_i \geq 0$ or required as input for $a^j_i < 0$. The last n columns of A imply an assumption of free disposal. The Walrasian equilibrium is defined as follows. A pair $(p^*, y^*) \in S^n \times \mathbb{R}^{n+k}_+$ is an equilibrium if it satisfies $z(p^*) \leq Ay^*$ and $p^{*\top} A \leq \mathbf{0}$. The interpretation is that at equilibrium the demand is met by supply for all commodities and no activity makes positive profit. Note that at equilibrium we have $p^{*\top} A y^* = 0$. This means that an activity having a deficit ($p^{*\top} a^j < 0$) is not producing ($y^*_j = 0$) while an activity in operation ($y^*_j > 0$) runs at balance ($p^{*\top} a^j = 0$). It is shown in Scarf [15] that the economy has at least one equilibrium.

Define $S_A = \{p \in S^n \mid p^\top A \leq \mathbf{0}\}$. It is shown by Eaves [7] (see also van den Elzen [8]) that S_A is an $(n-1)$-dimensional polytope and (p^*, y^*) is an equilibrium if and only if p^* is a stationary point of the function z on S_A. It is easy to see that the set S_A can be expressed as $S_A = \{x \in \mathbb{R}^n \mid p^\top A \leq \mathbf{0}, \sum_{i=1} x_i = 1\}$. If there is a redundant constraint, we just delete it. So we may assume that all constraints describing S_A are nonredundant. Applying the concept of robust stationary point, we have the following economically meaningful concept which refines the Walrasian equilibrium concept. Let $m = k + n$ and 1^n be the n-vector of ones.

Definition 6.4. *For given $\theta \in (0,1)$ a point $p \in S_A$ is a θ-robust Walrasian equilibrium of z if*

(1) *p is a relative interior point of S_A;*

(2) *for some $I \in \mathcal{I}$, $z(p) = \sum_{h \in I_m} y_h a^h + \beta 1^n$ with $y_h \geq 0$ for all $h \in I$ and $y_h = 0$ for all $h \in I_m \setminus I$, and $\beta \in \mathbb{R}$, if $y_l > y_k$, then $a^{l\top} p \geq \theta a^{k\top} p$.*

A robust Walrasian equilibrium p^* is the limit of a sequence of θ_t-robust Walrasian equilibria (p^t) as θ_t converges to zero. We are now going to explain the above concept in terms of economics. First note that for each $h \in I_m$ the parameter $y_h \geq 0$ represents the level of activity h. At θ-robust equilibrium, all production activities are making deficits ($p^\top a^h < 0$ for all h). A θ-robust equilibrium requires that if $y_l > y_k$, then $(0 >) \, a^{l\top} p \geq \theta a^{k\top} p \, (> a^{k\top} p)$. It says that when all firms are making deficits, the higher an activity level of a firm is, the lower the per unit deficit of that firm must be. In other words, the higher deficit a firm makes, the less that firm should produce. Hence as θ converges to zero, this adjustment mechanism will eventually bring the negative profit $y_h a^{h\top} p$ of each firm h close to zero and a robust Walrasian equilibrium state will be reached. This reveals a thought similar to the classical Walrasian tâtonnement adjustment process. We should point out the difference between the classical Walrasian tâtonnement adjustment process and the algorithm proposed in this paper. The former process is known to be nonconvergent, but the latter algorithm is globally convergent and also admits a nice economic interpretation as described above.

REFERENCES

[1] E.L. Allgower and K. Georg, *Numerical Continuation Methods*, Springer-Verlag, Berlin, 1990.

[2] J.V. Burke and J.J. Moré, *Exposing constraints*, SIAM J. Optim., 4 (1994), pp. 573–595.

[3] E. van Damme, *Stability and Perfection of Nash Equilibria*, Springer-Verlag, Berlin, 1987.

[4] B.C. Eaves, *On the basic theory of complementarity*, Math. Programming, 1 (1971), pp. 68–75.

[5] B.C. Eaves, *Homotopies for computation of fixed points*, Math. Programming, 3 (1972), pp. 1–22.

[6] B.C. Eaves, *Computing stationary points*, Math. Programming Stud., 7 (1978), pp. 1–14.

[7] B.C. Eaves, *Thought on computing market equilibrium with SLCP*, in The Computation and Modelling of Economic Equilibria, A.J.J. Talman and G. van der Laan, eds., North-Holland, Amsterdam, 1987, pp. 1–17.

[8] A. van den Elzen, *Adjustment Processes for Exchange Economies and Non-cooperative Games*, Springer-Verlag, Berlin, 1993.

[9] P. Hartman and G. Stampacchia, *On some nonlinear elliptic differential functional equations*, Acta Math., 115 (1966), pp. 271–310.

[10] T.C. Koopmans, *Activity Analysis of Production and Allocation*, Wiley, New York, 1951.

[11] G. van der Laan and A.J.J. Talman, *A restart algorithm for computing fixed points without an extra dimension*, Math. Programming, 20 (1979), pp. 33–48.

[12] G. van der Laan and A.J.J. Talman, *Simplicial approximation of solutions to the nonlinear complementarity problem with lower and upper bounds*, Math. Programming, 38 (1987), pp. 1–15.

[13] R.B. Myerson, *Refinements of Nash equilibrium concepts*, Internat. J. Game Theory, 8 (1978), pp. 73–80.

[14] H. Scarf, *The approximation of fixed points of a continuous mapping*, SIAM J. Appl. Math., 15 (1967), pp. 1328–1343.

[15] H. Scarf, *Computation of Economic Equilibria*, Yale University Press, New Haven, CT, 1973.

[16] A.J.J. Talman and Y. Yamamoto, *A simplicial algorithm for stationary point problems on polytopes*, Math. Oper. Res., 14 (1989), pp. 383–399.

[17] A.J.J. Talman and Z. Yang, *On the existence and computation of robust stationary points on the simplotope*, FEW manuscript, Tilburg University, The Netherlands, 1996.

[18] M.J. Todd, *The Computation of Fixed Points and Applications*, Lecture Notes in Econom. and Math. Systems 124, Springer-Verlag, Berlin, 1976.

[19] Y. Yamamoto, *A unifying model based on retraction for fixed point algorithms*, Math. Programming, 28 (1984), pp. 192–197.

[20] Y. Yamamoto, *A path-following procedure to find a proper equilibrium of finite games*, Internat. J. Game Theory, 22 (1993), pp. 49–59.

[21] Z. Yang, *A simplicial algorithm for computing robust stationary points of a continuous function on the unit simplex*, SIAM J. Control Optim., 34 (1996), pp. 491–506.

[22] Z. Yang, *Simplicial Fixed Point Algorithms and Applications*, Thesis Publisher, Amsterdam, 1996.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 2, pp. 353–375

A VARIABLE METRIC PROXIMAL POINT ALGORITHM FOR MONOTONE OPERATORS*

J. V. BURKE† AND MAIJIAN QIAN‡

Abstract. The proximal point algorithm (PPA) is a method for solving inclusions of the form $0 \in T(z)$, where T is a monotone operator on a Hilbert space. The algorithm is one of the most powerful and versatile solution techniques for solving variational inequalities, convex programs, and convex-concave mini-max problems. It possesses a robust convergence theory for very general problem classes and is the basis for a wide variety of decomposition methods called *splitting methods.* Yet the classical PPA typically exhibits slow convergence in many applications. For this reason, acceleration methods for the PPA algorithm are of great practical importance. In this paper we propose a variable metric implementation of the proximal point algorithm. In essence, the method is a Newton-like scheme applied to the Moreau–Yosida resolvent of the operator T. In this article, we establish the global and linear convergence of the proposed method. In addition, we characterize the superlinear convergence of the method. In a companion work, we establish the superlinear convergence of the method when implemented with Broyden updating (the nonsymmetric case) and BFGS updating (the symmetric case).

Key words. maximal monotone operator, proximal point methods, variable metric, global convergence, convergence rates

AMS subject classifications. Primary, 90C25; Secondary, 49J45, 47H05, 49M45

PII. S0363012992235547

1. Introduction.

The proximal point algorithm (PPA) is one of the most powerful and versatile solution techniques for problems of convex programming and minimax convex-concave programming. It possesses a robust convergence theory for very general problem classes in finite- and infinite-dimensions (e.g., see [11, 16, 21, 22, 23, 28, 32, 41, 40]) and is the basis for a wide variety of decomposition methods called *splitting methods* (e.g., see [4, 9, 12, 43, 44]). Yet, the classical PPA typically exhibits slow convergence in many applications. For this reason, acceleration methods for the PPA are of great practical importance. In this paper we propose a variable metric implementation of the proximal point algorithm. Our approach extends and refines results that originally appeared in [38] and is in the spirit of several recent articles [3, 7, 10, 18, 20, 24, 25, 36]. However, there is a fundamental difference between the method presented here and those studied in [3, 7, 10, 18, 20, 24, 25, 36]. This difference has a profound impact on the methodology applied in this article. All previous work on this topic (except [38]) applies exclusively to monotone operators that arise as the subdifferential of a finite-valued, finite dimensional convex function. The results of this article apply to general monotone operators on a Hilbert space. The resulting difference in methodology roughly corresponds to the difference between methods for function minimization and methods for solving systems of equations.

There are both advantages and disadvantages to the more general approach. The advantages are that the method applies to a much broader class of problems. This is

*Received by the editors August 12, 1992; accepted for publication (in revised form) September 30, 1997; published electronically November 23, 1998.
http://www.siam.org/journals/sicon/37-2/23554.html

†Department of Mathematics, Box 354350, University of Washington, Seattle, WA 98195–4350 (burke@math.washington.edu). The research of this author was supported by National Science Foundation grant DMS-9303772.

‡Department of Mathematics, California State University, Fullerton, CA 92834 (mqian@fullerton.edu).

so not only because the theory is developed in the Hilbert space setting, but, more important, because many monotone operators cannot be represented as the subdifferential of a finite-valued, finite dimensional convex function. General monotone operators do not possess many of the rich structural properties associated with the subdifferential of a convex function (e.g., subdifferentials of convex functions are the only maximal cyclically monotone operators [33]). In addition, in the case where the operator is the subdifferential of a convex function, we do not require the usual assumption that the underlying function be finite-valued.

The disadvantages of our general approach arise from the fact that the method cannot make use of the additional structure present when the operator is the subdifferential of a convex function. This complicates both the structure of the method and its analysis. Of particular note in this regard is the complexity of our global convergence result. If the operator is the subdifferential of a convex function, then solving the inclusion $0 \in T(x)$ is equivalent to minimizing the underlying convex function. The global convergence of a method is then typically driven by a line-search routine (e.g., see [3, 7, 10, 18, 20, 24, 25, 36]). In the general setting we do not have direct recourse to this strategy. This complicates both the structure of the algorithm and its convergence theory. Nonetheless, the proof technique developed in this paper can be refined in the convex programming setting, thereby significantly simplifying both the global and the local convergence results [5, 6].

Notwithstanding these differences in methodology, our approach is still nicely motivated by recalling the behavior of the PPA in the context of convex programming:

$$\min_{z\in\mathcal{H}} f(z)\ , \tag{1}$$

where $\mathcal{H}$ is a Hilbert space and $f\colon\mathcal{H}\mapsto\mathbb{R}\cup\{+\infty\}$ is a lower semicontinuous convex function that is not identically $+\infty$. Define the Moreau–Yosida regularization of f to be the function $f_\lambda\colon\mathcal{H}\mapsto\mathbb{R}$ given by

$$f_\lambda(\bar{z}) := \min_{z\in\mathcal{H}}\left\{\lambda f(z) + \frac{1}{2}\|z-\bar{z}\|^2\right\}.$$

The set of solutions to (1) corresponds precisely to the set of points at which f_λ attains its minimum value. The function f_λ is continuously Fréchet differentiable [28, Proposition 7.d]. The PPA applied to (1) is *approximately* the steepest descent algorithm applied to f_λ [11]. This analogy immediately suggests that a variable metric approach could be applied to the function f_λ to accelerate the method. This idea was first studied in [38] and is the basis of the acceleration techniques described in [3, 7, 10, 18, 20, 24, 25, 36].

In [3], Bonnans et al. develop methods along an algorithmic pattern originally suggested by Qian in [38]. This pattern circumvents many of the difficulties associated with a variable metric approach applied directly to the function f_λ. The key is to employ a matrix secant update based on the function f instead of f_λ. The local convergence results in [3, Section 3] require some smoothness assumptions. In particular, linear convergence is established when the function f is differentiable with Lipschitz continuous derivative, and superlinear convergence is established when f is twice strictly Fréchet differentiable at a unique solution $\bar{z}$, where the second derivative is positive definite (we speak only of quotient or q-rate of convergence).

In [18, 20, 24, 25], the authors apply the bundle concept for nonsmooth convex minimization [17] to approximate the Moreau–Yosida regularization f_λ and its derivative. Variable metric updates, in particular, quasi-Newton updates, are then applied using these approximate values. The superlinear convergence results in the papers [18, 20, 24] require either strong smoothness assumptions on the function f (such as the Lipschitz continuity of ∇f) or that the regularization parameter λ diverges to $+\infty$. In [20], Lemaréchal and Sagastizábal propose a clever *reversal* quasi-Newton formula which uses the value of the gradient of f_λ at a variety of points other than those strictly obtained by the iterates. This promising idea deserves further theoretical and numerical study.

In [10] and [36], the authors develop an approach based on Newton's method for semismooth functions as developed in [30, 31, 37, 34]. Properly speaking, these methods are neither an adaptation of the PPA algorithm nor a variable metric method. Nonetheless, the flavor of both of these methodologies is present. In order to obtain superlinear convergence, smoothness hypotheses are again required; however, these hypotheses are of a somewhat more technical nature. Specifically, it is required that

(a) the function f be *semismooth* at a unique solution to (1) [37],

(b) every element of the set-valued mapping

$$\partial_B^2 f(z) := \{\lim_{y_k \to z} \nabla^2 f_\lambda(y_k) \,:\, y_k \to x, \nabla f(y_k) \text{ exists for all } k = 1, 2, \ldots\}$$

be nonsingular at the unique solution $\bar{z}$, and

(c) the sequence of Hessian approximates $\{V_k\}$ used to generate the iterates $\{z_k\}$ satisfy

$$\lim_{k \to \infty} \operatorname{dist}(V_k, \partial_B^2 f(z_k)) = 0 \ . \tag{2}$$

One can show that the semismoothness hypotheses are satisfied in many cases of interest when f is finite-valued. Moreover, by Rademacher's theorem on the differentiability of Lipschitz continuous functions, it follows that the set-valued mapping $\partial_B^2 f(z)$ is always well-defined and compact-valued in the finite dimensional, finite-valued case, with the nonsingularity property being closely tied to the usual hypothesis of strong convexity. Although the limiting hypotheses on the V_k's is a bit strong, it is not entirely unreasonable in the absence of differentiability. In [36], Qi and Chen propose a very nice preconditioning technique wherein an exact value for the gradient of a *shifted* Moreau–Yosida regularization can be computed from inexact values for the gradient of f_λ. This technique is similar in spirit to the *reversal* quasi-Newton formula found in [20]. Both of these techniques should prove useful in numerical implementations.

The algorithm presented in this paper is most closely related to the methods proposed by Chen and Fukushima [7] and Mifflin, Sun, and Qi [25]. However, there are several fundamental distinctions, the foremost of which is that the methods in [7, 25] are restricted to finite dimensional finite-valued convex programming problems. Within this framework, these authors use bundle strategies to approximate f_λ and its gradient and establish the global convergence of their methods with the aid of a line search routine. Chen and Fukushima establish global and linear convergence results along with a generalization of the Dennis–Moré characterization theorem for superlinear convergence [14]. One of the most important features of the Chen–Fukushima algorithm is that the line search is based on the function f rather than approximations to the function f_λ. This is very important in practice since obtaining sufficiently accurate approximations to the function f_λ is usually quite time consuming. Their linear and superlinear convergence results blend bundle techniques with the theory of nonsmooth equations. Consequently, the convergence hypotheses are reminiscent of those employed in [10] and [36]; in particular, they require semismoothness, CD-regularity, and the strong approximation property (2). In [6], the methods of this paper are applied to the Chen–Fukushima algorithm to obtain the superlinear convergence of the method when BFGS matrix secant updating is employed.

In [25], Mifflin, Sun, and Qi obtain the first superlinear convergence result for a variable metric proximal point algorithm using the BFGS matrix secant update in the setting of finite dimensional finite-valued convex programming. Their proposed algorithm uses a line search based on approximations to the function f_λ and requires that the function f_λ is strongly convex with ∇f_λ Fréchet differentiable at the unique global solution to the convex program. In addition it is assumed that the iterates satisfy a certain approximation property involving the gradient ∇f_λ. In section 4 of this paper, we discuss how these hypotheses are related to those that are also required in our convergence analysis.

In this paper, we provide a general theory for a variable metric proximal point algorithm (VMPPA) applied to maximal monotone operators from a Hilbert space to itself. In the important special case of convex programming, where T is taken to be the subdifferential of the function f, we do not assume that f is finite-valued or differentiable on the whole space. However, to obtain superlinear convergence, we do require certain smoothness hypotheses at a unique global solution $\bar{z}$. These smoothness hypotheses differ from those assumed in [3, 18, 20, 24] since they are imposed on the operator T^{-1} rather than T. In this regard, they are reminiscent of the hypotheses employed in [25]. The choice of smoothness hypotheses has deep significance in the context of convex programming. Differentiability hypotheses on $T = \partial f$ imply the second-order differentiability of f, whereas differentiability hypotheses on $T^{-1} = (\partial f)^{-1}$ are related to the standard strong second-order sufficiency conditions of convex programming [40, Proposition 2] and thus reduce to the standard hypotheses used in local analysis of convergence. In particular, the differentiability of $(\partial f)^{-1}$ does not imply that ∂f is single-valued or differentiable, nor does it imply that f is finite-valued.

Our smoothness hypotheses also differ from those that appear in [7, 10, 36]. These methods rely on the theory of nonsmooth equations and require hypotheses such as semismoothness and nonsingularity of the elements of $\partial_B^2 f$. In addition, the proof theory for these methods specifically requires that the underlying convex function be finite-valued in a neighborhood of the unique solution to (1) (again, these methods assume that the function is finite-valued on all of $\mathbb{R}^n$). This limits direct application to constrained problems since in the constrained case solutions typically lie on the boundary of the constraint region (i.e., on the boundary of the domain of the essential objective function).

Throughout the paper we illustrate many of the ideas and results by applying them to the case of convex programming. Our purpose here is not only to show how the results can be applied, but also to ground them in the familiar surroundings of this concrete application. Further details on the application of these results to the case of convex programming can be found in [5].

The paper is structured as follows. We begin with a review of the classic proximal point algorithm in section 2. The VMPPA is introduced in section 3. This section contains the approximation criteria that must be satisfied at each iteration. Two criteria are presented. The first is required to obtain global convergence and the second is required to accelerate the local convergence of the method. This division into global and local criteria is one of the recurring themes of the paper. On the global level the method behaves like a steepest descent method, while at the local level it becomes more Newton-like. This feature is common to most general purpose methods in nonlinear programming, such as the nonmonotone descent methods, the dogleg method, and trust-region methods. In section 4 we discuss the smoothness hypotheses required for the local analysis. We also extend some of the differentiability results appearing in [19, 35] to maximal monotone operators. In section 5, we study the operators $\mathcal{N}_k$ associated with the Newton-like iteration proposed in section 3. The focus of this section is to provide conditions under which the operators $\mathcal{N}_k$ are nonexpansive at a solution to the inclusion $0 \in T(z)$. A global convergence result paralleling Rockafellar's 1976 result [41] is given in section 6. In section 7 we study local convergence rates. Linear convergence is established under a Lipschitz continuity assumption on T^{-1}, and a characterization of superlinear convergence for the VMPPA is also given. This characterization is modeled on the landmark characterization of superlinear convergence of variable metric methods in nonlinear programming due to Dennis and Moré [14]. In [6], we use this characterization result to establish the superlinear convergence of the method when the derivatives are approximated using the BFGS and Broyden updating strategies.

A word about our notation is in order. We denote the closed unit ball in the Hilbert space $\mathcal{H}$ by $\mathbb{B}$. Then the ball with center a and radius r is denoted by $a + r\mathbb{B}$. Given a set $Z \subset \mathcal{H}$ and an element $z \in \mathcal{H}$, the distance of z to Z is $\operatorname{dist}(z, Z) = \inf\{\|z - z'\| : z' \in Z\}$.

Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two Hilbert spaces. Given a *multifunction* (also referred to as a *mapping* or an *operator* depending on the context) $T : \mathcal{H}_1 \rightrightarrows \mathcal{H}_2$, the *graph* of T, $\operatorname{gph} T$, is the subset of the product space $\mathcal{H}_1 \times \mathcal{H}_2$ defined by $\operatorname{gph} T = \{(z, w) \in \mathcal{H}_1 \times \mathcal{H}_2 | w \in T(z)\}$. The *domain* of T is the set $\operatorname{dom} T := \{z \in \mathcal{H}_1 | T(z) \neq \emptyset\}$. The identity mapping will be denoted by I. The *inverse* of an operator T is defined by $T^{-1}(w) := \{z \in \mathcal{H}_1 | (z, w) \in \operatorname{gph} T\}$.

Given a lower semicontinuous convex function $f : \mathcal{H} \to \mathbb{R} \bigcup \{+\infty\}$, the *conjugate* of f is defined by $f^*(z^*) = \sup_{z \in \mathcal{H}}\{\langle z^*, z\rangle - f(z)\}$.

2. Monotone operators and the classic algorithm.

Given a real Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$, we say that the multifunction $T : \mathcal{H} \rightrightarrows \mathcal{H}$ is *monotone* if for every z and z' in $\operatorname{dom} T$, and $w \in T(z)$ and $w' \in T(z')$, we have $\langle z - z', w - w'\rangle \geq \kappa\|z - z'\|^2$ for some $\kappa \geq 0$. If $\kappa > 0$, then T is said to be *strongly monotone with modulus κ*. The monotone operator T is said to be *maximal* if its graph is not properly contained in the graph of any other monotone operator. An important example of a monotone operator is the *subgradient* of a convex function (see Minty [27] and Moreau [28]).

We are concerned with solving inclusions of the form

$$0 \in T(z), \tag{3}$$

where T is a maximal monotone operator. In the case of the convex programming problem (1), the operator T is the subdifferential of the convex function f, and the inclusion (3) characterizes the points z at which f attains its minimum value. A wide variety of other problems can be cast in this framework, e.g., variational inequalities, complementary problems, and mini-max problems. Existence results for inclusion (3) can be found in [41].

In 1962, Minty [27] showed that, when the operator T is maximal monotone, the *Moreau–Yosida resolvent* of T,

$$P_\lambda = (I + \lambda T)^{-1} \text{ with } \lambda > 0,$$

is single-valued and nonexpansive on $\mathcal{H}$. This result suggests that a solution to the inclusion $0 \in T(z)$ can be iteratively approximated by the recursion $z^{k+1} = P_\lambda(z^k)$. One can modify this scheme by varying the scalar λ and by choosing the iterates z^{k+1} to be an approximate solution to the equation $(I + \lambda_k T)(z) = z^k$. The PPA applies precisely these ideas. The algorithm, starting from any point z^0, generates a sequence $\{z^k\}$ in $\mathcal{H}$ by the approximation rule

$$z^{k+1} \approx (I + c_k T)^{-1}(z^k) \ . \tag{4}$$

The principal difficulty in applying the PPA lies in executing the operators $P_k = (I + c_k T)^{-1}$. In the case of convex programming, the iteration (4) reduces to the iteration

$$z^{k+1} \approx \arg\min_{z\in\mathcal{H}} \left\{ c_k f(z) + \frac{1}{2}\|z - z^k\|^2 \right\} \ .$$

Notice that executing the algorithm exactly (i.e., with "=" instead of "≈" in the above algorithm) can be as difficult as solving the original problem directly. Hence it is critical that the convergence results are obtained under the assumption of approximation.

In [22] and [23], Martinet proved the convergence of the *exact* PPA for certain cases of the operator T with fixed $c_k \equiv c$. The first theorem on the convergence of the general PPA was proved by Rockafellar [41] in 1976. His theorem not only insures the global convergence under an approximating rule, but also describes the global behavior when the inclusion $0 \in T(z)$ has no solution.

The convergence rate of the PPA depends on properties of the operator T, the choice of the sequence $\{c_k\}$, and the accuracy of the approximation in (4). The first rate of convergence results were also obtained by Rockafellar [41] in 1976, under the assumption that the solution set is a singleton $\{\bar{z}\}$. He proved that if the sequence $\{c_k\}$ is bounded away from 0, and $T^{-1}(w)$ is bounded by a linear function of $\|w\|$ when w is near 0, then the rate of convergence is at least linear. Luque [21] extended Rockafellar's theorem to the case where $T^{-1}(0)$ is not required to be a singleton, and showed that such an estimate of the convergence rate is tight.

3. The algorithm and approximation criteria. The algorithm proposed in this section is a Newton-like iteration for solving the resolvent equation $z = P_\lambda(z)$. In the context of the convex programming problem, the iteration takes the form

$$z^{k+1} = z^k - H_k \nabla f_\lambda(z^k) \ ,$$

where the operator H_k is used to approximate second-order properties of the function f_λ. If f_λ is twice differentiable with $[\nabla^2 f_\lambda(z^k)]^{-1}$ bounded, then for Newton's method one sets $H_k = [\nabla^2 f_\lambda(z^k)]^{-1}$. However, in general, f_λ is only known to be differentiable with Lipschitz continuous gradient [28]. Thus, in the finite dimensional case, the Hessian $\nabla^2 f_\lambda(x)$ is guaranteed to exist only on a dense subset by Rademacher's theorem. Further results on the second-order properties of f_λ can be found in [19, 35, 42].

It is well known that the negative gradient $-\nabla f_\lambda(z^k)$ is the unique element w^k solving the problem

$$\min_{w\in\mathcal{H}} \left\{ \lambda f(z^k + w) + \frac{1}{2}\|w\|^2 \right\}$$

or, equivalently, satisfying the inclusion

$$0 \in \lambda \partial f(z^k + w^k) + w^k \ . \tag{5}$$

The PPA for a general maximal monotone operator T can be formally derived from (5) by replacing λ, z^k, and ∂f by c_k, z^k, and T, respectively, to obtain

$$0 \in c_k T(z^k + w^k) + w^k \ ,$$

or equivalently,

$$w^k = [(I + c_k T)^{-1} - I](z^k) \ ,$$

where equality follows from the fact that w^k is unique. This motivates us to define the operator

$$D_k := (I + c_k T)^{-1} - I \ . \tag{6}$$

This operator provides the analogue of the direction of steepest descent in the operator setting.

The algorithm we propose for solving the inclusion $0 \in T(z)$ can be succinctly stated as follows.

The Variable Metric Proximal Point Algorithm.

Let $z^0 \in \mathcal{H}$ and $c_0 \geq 1$ be given. Having z^k, set

$$z^{k+1} := z^k + H_k w^k, \quad \text{where} \ \ w^k \approx D_k(z^k),$$

and choose $c_{k+1} \geq 1$.

As mentioned in the previous section, it is critical that the convergence results are obtained under the assumption that $D_k(z^k)$ can only be approximated. We use the following approximation criteria:

$$(\mathcal{G}) \qquad \|w^k - D_k(z^k)\| \leq \min\left\{1, \frac{1}{\|H_k\|}\right\} \epsilon_k \ \ \text{with} \ \ \sum_{k=0}^{\infty} \epsilon_k < \infty$$

and

$$(\mathcal{L}) \qquad \|w^k - D_k(z^k)\| \leq \delta_k \|w^k\| \ \ \text{with} \ \ \lim_{k \to \infty} \delta_k = 0 \ .$$

The approximation criterion $(\mathcal{G})$ is used to establish global convergence properties, while criterion $(\mathcal{L})$ is used to obtain local rates of convergence.

Although these criteria are used in the proof of convergence, they are impractical from the perspective of implementation. In their stead, we provide criteria that are implementable. To obtain these criteria we recall the following result from Rockafellar [41].

Proposition 1 (see [41, Proposition 3]). *Let $S_k(w) := T(z^k + w) + \frac{1}{c_k} w$. Then $0 \in S_k(w^k) \Leftrightarrow w^k = D_k(z^k)$. Moreover, for all $w \in \mathcal{H}$ we have the bound*

$$\|w - D_k(z^k)\| \leq c_k \text{dist}\,(0, S_k(w)) \ . \tag{7}$$

Proposition 1 yields the following alternative approximation criteria for the w^k's. Since this result is an immediate consequence of Proposition 1, its proof is omitted.

Proposition 2. *Consider the following acceptance criteria for the w^k's:*

$$(\mathcal{G}') \qquad \operatorname{dist}(0, S_k(w^k)) \le \min\left\{1, \frac{1}{\|H_k\|}\right\}\frac{\epsilon_k}{c_k} \quad \textit{with} \quad \sum_{k=0}^{\infty}\epsilon_k < \infty$$

and

$$(\mathcal{L}') \qquad \operatorname{dist}(0, S_k(w^k)) \le \frac{\delta_k}{c_k}\|w^k\| \quad \textit{with} \quad \lim_{k\to\infty}\delta_k = 0\ .$$

We have $(\mathcal{G}')$ implies $(\mathcal{G})$ and $(\mathcal{L}')$ implies $(\mathcal{L})$.

Remark. Note that to satisfy either $(\mathcal{G}')$ or $(\mathcal{L}')$ it is not necessary to find an element of $S_k(w^k)$ of least norm.

Before leaving this section we recall from [41] a few properties of the operators D_k and $P_k := D_k + I$ that are essential in the analysis to follow.

Proposition 3 (see [41, Proposition 1]).

a) *The operator D_k can be expressed as*

$$D_k = -\left(I + T^{-1}\frac{1}{c_k}\right)^{-1}, \tag{8}$$

and for any $z \in \mathcal{H}$, $-\frac{1}{c_k}D_k(z) \in T(P_k(z))$.

b) *For any $z, z' \in \mathcal{H}$, $\langle P_k(z) - P_k(z'), D_k(z) - D_k(z')\rangle \le 0$.*

c) *For any $z, z' \in \mathcal{H}$, $\|P_k(z) - P_k(z')\|^2 + \|D_k(z) - D_k(z')\|^2 \le \|z - z'\|^2$.*

Remark. An important consequence of part c) above is that the operators P_k and D_k are Lipschitz continuous with Lipschitz constant 1; that is, they are nonexpansive. Henceforth, we make free use of this fact.

4. On the differentiability of T^{-1} and D_k.

Just as Newton's method for minimization locates roots of the gradient, one can view the VMPPA as a Newton-like method for locating roots of the operator D_k. This perspective motivates our approach to the local convergence analysis. For this analysis, we require that the operator T^{-1} possesses certain smoothness properties. These properties in turn imply the smoothness of the operators D_k. Smoothness hypotheses are used in the convergence analysis in much the same way as they are used in the convergence analysis for Newton's method. For example, recall that to ensure the quadratic convergence of Newton's method one requires the derivative at a solution to be both locally Lipschitz and nonsingular. Nonsingularity ensures that the iterates are well-defined and can be bounded, while the Lipschitzian hypothesis guarantees that the error in the linearization is quadratically bounded (see [29, sections 3.2.12 and 10.2.2]). We make use of similar properties in our analysis.

In order to discuss the smoothness of T^{-1} and D_k, we recall various notions of differentiability for multivalued functions from the literature. For a more thorough treatment of these ideas in the context of monotone operators, we refer the reader to [1, 19, 26, 35, 42].

Definition 4. *We say that an operator $\Psi : \mathcal{H} \rightrightarrows \mathcal{H}$ is Lipschitz continuous at a point $\bar{w}$ (with modulus $\alpha \ge 0$) if the set $\Psi(\bar{w})$ is nonempty and there is a $\tau > 0$ such that*

$$\Psi(w) \subset \Psi(\bar{w}) + \alpha\|w - \bar{w}\|\mathbb{B} \qquad \textit{whenever } \|w - \bar{w}\| \le \tau\ .$$

We say that Ψ is differentiable at a point $\bar{w}$ if $\Psi(\bar{w})$ consists of a single element $\bar{z}$ and there is a continuous linear transformation $J : \mathcal{H} \to \mathcal{H}$ such that for some $\delta > 0$,

$$\emptyset \ne \Psi(w) - \bar{z} - J(w - \bar{w}) \subset o(\|w - \bar{w}\|)\mathbb{B} \qquad \textit{whenever } \|w - \bar{w}\| \le \delta\ .$$

We then write $J = \nabla\Psi(\bar{w})$.

Remarks. 1) These definitions of Lipschitz continuity and differentiability for multifunction are taken from [41, pp. 885 and 887] (also see [2, p. 41]). Note that these notions of Lipschitz continuity and differentiability correspond to the usual notions when Ψ is single-valued.

2) Rockafellar [41, Theorem 2] was the first to use Lipschitz continuity to establish rates of convergence for the PPA.

3) When the set $\Psi(\bar{w})$ is restricted to be a singleton $\{\bar{z}\}$, the differentiability of Ψ at $\bar{w}$ implies the Lipschitz continuity of Ψ at $\bar{w}$. Moreover, one can take $\alpha(\tau) \to \|J\|$ as $\tau \to 0$. This observation is verified in [41, Proposition 4].

4) It follows from the definition of monotonicity that if T is a maximal monotone operator, then the operator $\nabla T(z)$ is positive semidefinite whenever it exists.

We now give a result that relates the differentiability of a multivalued function to the differentiability of its inverse. The proof is omitted since it parallels the proof of a similar result for single-valued functions.

LEMMA 5. *Assume that* $\Psi : \mathcal{H} \rightrightarrows \mathcal{H}$ *is differentiable at* $\bar{z}$ *with* $\Psi(\bar{z}) = \{\bar{w}\}$ *and* $\nabla\Psi(\bar{z}) = J$ *with* J^{-1} *bounded. Also assume that* Ψ^{-1} *is Lipschitz continuous at* $\bar{w}$ *with* $\Psi^{-1}(\bar{w}) = \{\bar{z}\}$. *Then* Ψ^{-1} *is differentiable at* $\bar{w}$ *with* $\nabla\Psi^{-1}(\bar{w}) = J^{-1}$.

In the two examples that follow, we examine the concepts introduced in Definition 4 when the operator in question is the subdifferential of a convex function. The first example illustrates that ∂f^{-1} can be Lipschitz continuous but not differentiable at the origin, while in the second example ∂f^{-1} is differentiable at the origin, but ∂f is not differentiable on $(\partial f)^{-1}(0)$.

EXAMPLE 6. *Let*

$$f(z) := \begin{cases} 0 & \text{if } z < 0, \\ z & \text{if } z \geq 0, \end{cases} \qquad \text{and} \qquad T(z) := \partial f(z) = \begin{cases} 0 & \text{if } z < 0, \\ [0,1] & \text{if } z = 0, \\ 1 & \text{if } z > 0\ . \end{cases}$$

$$\text{Then} \qquad T^{-1}(y) = \begin{cases} \emptyset & \text{if } y < 0 \text{ or } y > 1, \\ (-\infty, 0] & \text{if } y = 0, \\ \{0\} & \text{if } y \in (0,1), \\ [0,\infty) & \text{if } y = 1\ . \end{cases}$$

T^{-1} *is Lipschitz continuous at* 0 *but is not differentiable at* 0.

EXAMPLE 7. *Let*

$$f(z) := \begin{cases} -z & \text{if } z < 0, \\ z^{5/3} & \text{if } z \geq 0, \end{cases} \qquad \text{and} \qquad T(z) := \partial f(z) = \begin{cases} -1 & \text{if } z < 0, \\ [-1,0] & \text{if } z = 0, \\ \frac{5}{3}z^{2/3} & \text{if } z > 0. \end{cases}$$

$$\text{Then} \qquad T^{-1}(y) = \begin{cases} \emptyset & \text{if } y < -1, \\ (-\infty, 0] & \text{if } y = -1, \\ \{0\} & \text{if } y \in (-1,0), \\ \frac{3}{5}y^{3/2} & \text{if } y \geq 0\ . \end{cases}$$

T^{-1} *is differentiable at* 0 *with* $J = 0$, *but* T *is not differentiable on* $T^{-1}(0)$.

The superlinear convergence result of section 7 requires the assumption that the operator T^{-1} be differentiable at the origin. Although this is a severe restriction on the applicability of these results, it turns out that in the case of convex programming it is a consequence of the standard second-order sufficiency conditions for constrained mathematical programs. This and related results were established by Rockafellar in [40, Proposition 2]. In this context, it is important to note that the second-order sufficiency condition is the standard hypothesis used in the mathematical programming literature to ensure the rapid local convergence of numerical methods. So, at least in the context of constrained convex programming, such a differentiability hypothesis is not as severe an assumption as one might at first suspect. To the contrary, it is a bit weaker than the standard hypothesis employed for such results. For the sake of completeness, we recall a portion of Rockafellar's result below.

Theorem 8. *Consider the convex programming problem* (1)*, where* $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *is given by*

$$f(z) = \begin{cases} f_0(z) & \text{if } f_i(z) \leq 0 \text{ for } i = 1, 2, \ldots, m, \\ +\infty & \text{otherwise,} \end{cases}$$

with $f_i : \mathbb{R}^n \to \mathbb{R}$ *convex for* $i = 0, 1, \ldots, m$. *Suppose that the following conditions are satisfied:*

(i) *The functions* f_i *for* $i = 0, 1, \ldots, m$ *are* $k \geq 2$ *times continuously differentiable in a neighborhood of a point* $\bar{z} \in \mathbb{R}^n$.

(ii) *There is a Kuhn–Tucker vector* $\bar{y} \in \mathbb{R}^m$ *for* $\bar{z}$ *such that* $\bar{y}_i > 0$ *for* $i \in I(\bar{z}) = \{i : f_i(\bar{z}) = 0,\ i = 1, 2, \ldots, m\}$.

(iii) *The gradients* $\{\nabla f_i(\bar{z}) : i \in I(\bar{z})\}$ *are linearly independent.*

(iv) *The matrix* $H = \nabla^2 f_0(\bar{z}) + \sum_{i=1}^m \bar{y}_i \nabla^2 f_i(\bar{z})$ *satisfies* $u^T H u > 0$ *for every nonzero* $u \in \mathbb{R}^n$ *such that* $\nabla f_0(\bar{z})^T u = 0$, *and* $\nabla f_i(\bar{z})^T u = 0$ *for* $i \in I(\bar{z})$.

Then the operator ∂f^{-1} *is* $(k-1)$ *times continuously differentiable in a neighborhood of the origin.*

Remark. Theorem 8 follows by applying the implicit function theorem to the Kuhn–Tucker conditions for the parameterized problems $\min\{f(z) - \langle w, z\rangle\}$ in a neighborhood of $w = 0$. The relationship to ∂f^{-1} comes from the fact that $\partial f^{-1}(w) = \operatorname{argmin}\{f(z) - \langle w, z\rangle\}$. Rockafellar establishes the result only for $k = 2$. The extension to $k > 2$ follows trivially from the implicit function theorem.

We now examine the differentiability properties of the mapping D_k. Two results in this direction are given. The first uses (8) to relate the differentiability of the operators T^{-1} and D_k, while the second uses the definition of D_k given in (6) to relate the differentiability of the operators T and D_k.

Proposition 9. *Let* $T : \mathcal{H} \rightrightarrows \mathcal{H}$ *be maximal monotone and* $\lambda > 0$. *Define*

$$D(z) = -\left(I + T^{-1}\frac{1}{\lambda}\right)^{-1}(z) \ . \tag{9}$$

Let $\bar{z} \in \mathcal{H}$ *and set* $\bar{w} = D(\bar{z})$ *and* $\bar{y} = -\frac{1}{\lambda}\bar{w}$. *The operator* T^{-1} *is differentiable at* $\bar{y}$ *with* $[I + \frac{1}{\lambda}\nabla(T^{-1})(\bar{y})]^{-1}$ *bounded if and only if the operator* D *is differentiable at* $\bar{z}$ *with* $(\nabla D(\bar{z}))^{-1}$ *bounded. In either case, we have*

$$\nabla D(\bar{z}) = -\left[I + \frac{1}{\lambda}\nabla(T^{-1})(\bar{y})\right]^{-1} \ . \tag{10}$$

Proof. First assume that T^{-1} is differentiable at $\bar{y}$ with $\nabla(T^{-1})(\bar{y})$ bounded. The differentiability of T^{-1} at $\bar{y}$ clearly implies that of D^{-1} at $\bar{w}$ with

$$\nabla[D^{-1}](\bar{w}) = -\left(I + \frac{1}{\lambda}\nabla[T^{-1}](\bar{y})\right) .$$

Since D is Lipschitzian with $D(\bar{z}) = \bar{w}$, Lemma 5 implies that D is differentiable at $\bar{z}$ with derivative given by (10). Since $\nabla[D^{-1}](\bar{w}) = (\nabla D(\bar{z}))^{-1}$, we conclude that the latter is bounded.

Conversely, assume that D is differentiable at $\bar{z}$ with $(\nabla D(\bar{z}))^{-1}$ bounded. We show that D^{-1} is single-valued and Lipschitzian at $\bar{w}$. The result will then follow from Lemma 5.

Let $\delta > 0$ be as in Definition 4 for $\nabla D(\bar{z})$. Since D is single-valued and $\nabla D(\bar{z})$ is surjective (it is invertible), we may apply a standard open mapping result from functional analysis (e.g., [8, Theorem 15.5]) to obtain the existence of a $\rho > 0$ and a $0 < \hat{\delta} < \delta$ such that

$$\bar{w} + \rho\mathbb{B} \subset D(\bar{z} + \hat{\delta}\mathbb{B}) . \tag{11}$$

Hence for each $w \in \bar{w} + \rho\mathbb{B}$ and $z \in D^{-1}(w) \cap (\bar{z} + \hat{\delta}\mathbb{B}) \neq \emptyset$ we have

$$w = \bar{w} + \nabla D(\bar{z})(z - \bar{z}) + o(\|z - \bar{z}\|) . \tag{12}$$

Since $(\nabla D(\bar{z}))^{-1}$ is bounded, there is a $\kappa > 0$ such that

$$\|w - \bar{w}\| + o(\|z - \bar{z}\|) = \|\nabla D(\bar{z})(z - \bar{z})\| \geq \kappa\|z - \bar{z}\| .$$

Hence, by reducing ρ and $\hat{\delta}$ if necessary, we may assume that

$$\|w - \bar{w}\| \geq \frac{\kappa}{2}\|z - \bar{z}\| \geq \frac{\kappa}{2}\|w - \bar{w}\|$$

for $w \in \bar{w} + \rho\mathbb{B}$, where the second inequality follows since D is nonexpansive. Therefore, we can assume that $o(\|z - \bar{z}\|) = o(\|w - \bar{w}\|)$ for all $w \in \bar{w} + \rho\mathbb{B}$ and $z \in D^{-1}(w) \cap (\bar{z} + \hat{\delta}\mathbb{B})$. By substituting this into (12) and rearranging, we obtain

$$\begin{aligned} & z = \bar{z} + (\nabla D(\bar{z}))^{-1}(w - \bar{w}) + o(\|w - \bar{w}\|) \\ & \text{for all } w \in \bar{w} + \rho\mathbb{B} \text{ and } z \in D^{-1}(w) \cap (\bar{z} + \hat{\delta}\mathbb{B}). \end{aligned} \tag{13}$$

We now show that (13) implies the existence of an $\epsilon > 0$ such that $D^{-1}(\bar{w} + \epsilon\mathbb{B}) \subset \bar{z} + \hat{\delta}\mathbb{B}$. Indeed, if this were not the case, then there would exist sequences $\{w_i\}$ and $\{z_i\}$ such that $z_i \in D^{-1}(w_i)$, $\|z_i - \bar{z}\| > \hat{\delta}$, and $w_i \to \bar{w}$. Since D^{-1} is itself maximal monotone, its images are convex; hence, by (11), there exists a sequence $\{\hat{z}_i\}$ with $\hat{z}_i \in D^{-1}(w_i)$ and $\|\hat{z}_i - \bar{z}\| = \hat{\delta}$ for all $i = 1, 2, \ldots$. But then (13) implies that

$$\hat{z}_i = \bar{z} + (\nabla D(\bar{z}))^{-1}(w_i - \bar{w}) + o(\|w_i - \bar{w}\|)$$

for all $i = 1, 2, \ldots$. This contradicts the fact that $w_i \to \bar{w}$ and $\|\hat{z}_i - \bar{z}\| = \hat{\delta}$ for all $i = 1, 2, \ldots$, and so such an $\epsilon > 0$ must exist. This fact combined with (13) implies that D^{-1} is Lipschitzian at $\bar{w}$ with $D^{-1}(\bar{w}) = \{\bar{z}\}$. Lemma 5 now applies to yield the result. □

Proposition 10. *Let D be defined as in* (9). *Let $\bar{z} \in \mathcal{H}$ and set $\bar{y} = (I + D)(\bar{z})$. The operator T is differentiable at $\bar{y}$ with $[I + \lambda \nabla T(\bar{y})]^{-1}$ bounded if and only if the operator D is differentiable at $\bar{z}$ with $[I + \nabla D(\bar{z})]^{-1}$ bounded. In either case we have the formula*

$$\nabla D(\bar{z}) = [I + \lambda \nabla T(\bar{y})]^{-1} - I \ .$$

Proof. Replace D by $P := I + D = (I + \lambda T)^{-1}$ and observe that D is differentiable at $\bar{z}$ with $[I + \nabla D(\bar{z})]^{-1}$ bounded if and only if P is differentiable at $\bar{z}$ with $[\nabla P(\bar{z})]^{-1}$ bounded. The proof now follows the same argument as in the proof of Proposition 9 with D replaced by P, T^{-1} replaced by T, and $\bar{w}$ replaced by $\bar{y}$. □

Propositions 9 and 10 say quite different things about the differentiability of D_k. To illustrate this difference, observe that in Example 7 the operator T is not differentiable at 0, while T^{-1} and D are differentiable at 0. On the other hand, if we take $T = \partial f$ with $f(x) = |x|^3$, then T^{-1} is not differentiable at 0, while T and D are differentiable at 0. It is also important to note that even if neither T nor T^{-1} is differentiable, D may be differentiable. But, in this case, we know from Propositions 9 and 10 that if D is differentiable and neither T nor T^{-1} is differentiable, then both $\nabla D(\bar{z})$ and $\nabla P(\bar{z})$ have to be singular or have unbounded inverses. For a further discussion of these issues in the context of finite dimensional convex programming, see [35].

When T is assumed to be the subdifferential of a convex function f, Propositions 9 and 10 can be refined by making use of the relation $\partial f^{-1} = \partial f^*$, where f^* is the convex conjugate of f [39, Corollary 12A]. This allows us to extend [35, Theorem 1] and [35, Theorem 2] to the Hilbert space setting (also see [19, Theorem 3.1]). However, some caution in terminology is required since f^* is not necessarily twice differentiable in the classical sense at points where ∂f^* is differentiable in the sense of Definition 4. Indeed, ∂f^* may be multivalued arbitrarily close to a point of differentiability. The best way to interpret this result is through Alexandrov's theorem [1], which states that at almost every point $\bar{z}$ in the interior of the domain of a convex function $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{\infty\}$ there is a quadratic function $q_{\bar{z}}$ such that $f(x) = q_{\bar{z}}(x) + o(\|x - \bar{z}\|^2)$. In [19] and [35], the matrix $\nabla^2 q_{\bar{z}}$ is called a *generalized Hessian* and is denoted $Hf(x)$. Note that the existence of a generalized Hessian at the point $\bar{z}$ guarantees that f is strictly differentiable at $\bar{z}$. Moreover, if $\partial f(x)$ is single-valued in a neighborhood of a point $\bar{z}$ at which $Hf(\bar{z})$ exists, then $\nabla^2 f(\bar{z})$ exists and equals $Hf(\bar{z})$. We extend this terminology to the Hilbert space setting with the following definition.

Definition 11. *Let $\phi : \mathcal{H} \mapsto \mathbb{R} \cup \{\infty\}$ be a function on the Hilbert space $\mathcal{H}$. We say that ϕ is twice differentiable in the generalized sense at a point $\bar{z} \in \mathcal{H}$ if there is a continuous quadratic functional $q_{\bar{z}}$ such that $\phi(x) = q_{\bar{z}}(x) + o(\|x - \bar{z}\|^2)$. The operator $\nabla^2 q_{\bar{z}}$ is called a generalized Hessian of ϕ at $\bar{z}$ and is denoted by $H\phi(\bar{z})$.*

With this terminology in hand, we apply Propositions 9 and 10 to the case of convex programming. The proofs of these results are not required since they are a direct translation of Propositions 9 and 10 into the terminology of convex programming.

Corollary 12. *Let $f : \mathcal{H} \to \mathbb{R} \bigcup \{+\infty\}$ be lower semicontinuous and convex. Let $\bar{z} \in \mathcal{H}$ and set $\bar{w} = \nabla f_\lambda(\bar{z})$ and $\bar{y} = \frac{1}{\lambda}\bar{w}$. Then f_λ is twice (Fréchet) differentiable at $\bar{z}$ with $[\nabla^2 f_\lambda(\bar{z})]^{-1}$ bounded if and only if f^* has a generalized Hessian at $\bar{y}$ with $[I + \frac{1}{\lambda} Hf^*(\bar{y})]^{-1}$ bounded. In either case we have*

$$\nabla^2 f_\lambda(\bar{z}) = \left[I + \frac{1}{\lambda} Hf^*(\bar{y}) \right]^{-1} \ .$$

Corollary 13. *Let $f : \mathcal{H} \to \mathbb{R}\bigcup\{+\infty\}$ be lower semicontinuous and convex. Let $\bar{z} \in \mathcal{H}$ and set $\bar{y} = \bar{z} - \nabla f_\lambda(\bar{z})$. Then f_λ is twice (Fréchet) differentiable at $\bar{z}$ with $[I + \nabla^2 f_\lambda(\bar{z})]^{-1}$ bounded if and only if f is twice differentiable in the generalized sense at $\bar{y}$ with $[I + \lambda H f(\bar{y})]^{-1}$ bounded. In either case we have*

$$\nabla^2 f_\lambda(\bar{z}) = I - [I + \lambda H f(\bar{y})]^{-1} \ .$$

Remark. As observed earlier, the generalized Hessian is necessarily positive semidefinite. This observation can be used to further refine the statement of Corollaries 12 and 13.

5. Newton operators. In this section we study the operators associated with the variable metric proximal point iteration:

$$\mathcal{N}_k := I + H_k D_k = P_k + (H_k - I)D_k \ . \tag{14}$$

This notation emphasizes the fact that these operators produce Newton-like iterates. Just as in the case of the classical Newton's method for equation solving [29, section 12.6], one of the keys to the convergence analysis is to show that these operators are contractive with respect to the solution set $T^{-1}(0)$. Clearly the operators $\mathcal{N}_k$ are single-valued. Moreover, fixed points of the operators $\mathcal{N}_k$ are solutions to the inclusion $0 \in T(z)$ since

$$0 \in T(z) \ \Leftrightarrow \ P_k(z) = z \ \Leftrightarrow \ D_k(z) = 0 \ \Leftrightarrow \ \mathcal{N}_k(z) = z.$$

Thus, conditions that ensure that the operators $\mathcal{N}_k$ are nonexpansive with respect to $T^{-1}(0)$ are important for the global analysis of the variable metric proximal point iteration. To obtain this property, we impose the following conditions on the linear transformations $\{H_k\}$:

(H1) Each H_k is a continuous linear transformation with continuous inverse.

(H2) There is a nonempty closed bounded subset Γ of $T^{-1}(0)$ such that

$$\|(H_k - I)D_k(z^k)\| \le \gamma_k \|D_k(z^k)\| \quad \text{for all } k,$$

where

$$\gamma_k := \frac{\|D_k(z^k)\|}{2\sigma_k + 3\|D_k(z^k)\|} \quad \text{with } \sigma_k = \sup\{\|z^k - z\| \ : \ z \in \Gamma\}.$$

Remark. The set Γ in (H2) is used to guarantee the boundedness of the sequence $\{z^k\}$. By taking $\Gamma = \{\bar{z}\}$, one can show that every weak cluster point of the sequence $\{z^k\}$ is an element of $T^{-1}(0)$. It was observed by Iusem [13] that if $T^{-1}(0)$ is bounded and one takes $\Gamma = T^{-1}(0)$, then the sequence $\{z^k\}$ has a weak limit $z^\infty \in T^{-1}(0)$ (see Theorem 17 and [41, Theorem 1]).

Hypothesis (H1) is standard and is automatically satisfied in the finite dimensional case. On the other hand, hypothesis (H2) is quite technical and requires careful examination. This hypothesis is problematic since it specifies that the matrices H_k satisfy a condition that depends on the unknown values σ_k and $\|D_k(z^k)\|$. We will show that in certain cases it is possible to satisfy (H2) without direct knowledge of these unknown values. This is done in two steps. First it is shown in Lemma 14 that if T^{-1} is Lipschitz continuous or differentiable at the origin, then γ_k is bounded below by a positive constant (which can be taken to be 1/6 as $\|D_k(z^k)\|$ approaches zero).

Then, in Lemma 15, it is shown that (H2) is satisfied if a related condition in terms of H_k and w^k is satisfied. Taken together, these results imply that at least locally (H2) can be satisfied by checking a condition based on known quantities.

Further insight into hypothesis (H2) can be gained by considering the case in which T^{-1} is differentiable at the origin. In this case H_k is intended to approximate $-(\nabla D_k(0))^{-1} = (I + c_k^{-1} J)$, where $J = \nabla(T^{-1})(0)$ (by Proposition 9). Hence, if $H_k \approx -(\nabla D_k(0))^{-1}$, then $(H_k - I) \approx c_k^{-1} J$. Therefore, one can guarantee that (H2) is satisfied by choosing c_k sufficiently large and $H_k \approx I$. This fact is used in [6] to establish the superlinear convergence of the method when the H_k's are obtained via matrix secant updating techniques.

The purpose of hypothesis (H2) is to globalize what is essentially a local algorithm (Newton's method). In the context of convex programming, one commonly obtains global convergence properties with the aid of a line search routine applied to the objective function f or its regularization f_λ. However, in the operator setting there is no natural underlying objective function to which a line search can be applied. This is a key difference between the approach taken in this paper and those in [3, 7, 10, 18, 20, 24, 36]. In the convex programming setting, the global convergence of the VMPPA is driven by a line search routine applied to the objective function f (or its regularization f_λ). In the operator setting, hypothesis (H2) replaces the line search and the associated hypotheses needed to make the line search strategy effective (such as the finite-valuedness of the objective function f and the boundedness of the sequence $\{H_k\}$). On the other hand, when it is known that the operator T is the subdifferential of a finite-valued finite dimensional convex function, then the algorithm of this paper can be modified to include the line search routine of Chen and Fukushima [7], thereby avoiding the need for hypothesis (H2) [6].

We now show three cases where the γ_k's are bounded away from zero.

LEMMA 14. *Suppose $T^{-1}(0)$ is nonempty.*

(i) *If the operator T is strongly monotone with modulus κ, then $T^{-1}(0) = \{\bar{z}\}$,*

$$\|z^k - \bar{z}\| \le \left(1 + \frac{1}{\kappa c_k}\right) \|D_k(z^k)\| \ ,$$

and $\gamma_k \ge \frac{1}{5+\frac{2}{\kappa c_k}} \ge \frac{1}{5+2/\kappa}$ for all k.

(ii) *If the operator T^{-1} is Lipschitz continuous at the origin with modulus α, then*

$$\operatorname{dist}(z^k, T^{-1}(0)) \le \left(1 + \frac{\alpha}{c_k}\right) \|D_k(z^k)\| \ , \tag{15}$$

for all k such that $\|D_k(z^k)\| \le \tau$, where τ is given in Definition 4. Moreover, if $T^{-1}(0) = \{\bar{z}\}$, then $\gamma_k \ge \frac{1}{5+2\alpha/c_k} \ge \frac{1}{5+2\alpha}$ for all k such that $\|D_k(z^k)\| \le \tau$.

(iii) *If T^{-1} is differentiable at the origin with derivative J, then $T^{-1}(0) = \{\bar{z}\}$, there is a $\delta > 0$ such that for all k with $\|D_k(z^k)\| \le \tau$ we have*

$$\|z^k - \bar{z}\| \le \left(1 + \frac{\|J\|}{c_k} + \sigma(\|D_k(z^k)\|)\right) \|D_k(z^k)\| \ ,$$

and $\gamma_k \ge \frac{1}{5+2\frac{\|J\|}{c_k}+\sigma(\|D_k(z^k)\|)}$ for all k, where $\sigma(\tau) \to 0$ as $\tau \to 0$.

Proof.

(i) If T is strongly monotone with modulus κ, then $\|z - z'\| \le \frac{1}{\kappa}\|w - w'\|$ for any z, z', w, w' such that $w \in T(z)$ and $w' \in T(z')$. That is, T^{-1} is single-valued and Lipschitz continuous. Let $z = P_k(z^k)$ and $z' = \bar{z}$, where $\{\bar{z}\} = T^{-1}(0)$. By Proposition 3 a) we have $-\frac{1}{c_k}D_k(z^k) \in T(P_k(z^k))$. Hence

$$\|z^k - \bar{z}\| \leq \|z^k - P_k(z^k)\| + \|P_k(z^k) - \bar{z}\| \leq \left(1 + \frac{1}{\kappa c_k}\right)\|D_k(z^k)\| \ ,$$

since $D_k = P_k - I$. By the definition of γ_k,

$$\begin{aligned}\gamma_k &= \frac{\|D_k(z^k)\|}{2\|z^k - \bar{z}\| + 3\|D_k(z^k)\|} \\ &\geq \frac{\|D_k(z^k)\|}{2(1 + \frac{1}{\kappa c_k})\|D_k(z^k)\| + 3\|D_k(z^k)\|} \geq \frac{\kappa c_k}{5\kappa c_k + 2} \ .\end{aligned}$$

This establishes the result since $c_k \geq 1$ for all k.

(ii) If $\|D_k(z^k)\| \leq \tau$, Definition 4 implies that

$$T^{-1}\left(-\frac{1}{c_k}D_k(z^k)\right) \subset T^{-1}(0) + \alpha\left\|\frac{1}{c_k}D_k(z^k)\right\|\mathbb{B} = T^{-1}(0) + \frac{\alpha}{c_k}\|D_k(z^k)\|\mathbb{B} \ ,$$

or

$$\left(I + T^{-1}\frac{1}{c_k}\right)(-D_k(z^k)) + D_k(z^k) \subset T^{-1}(0) + \frac{\alpha}{c_k}\|D_k(z^k)\|\mathbb{B} \ .$$

Since $D_k(z^k) = -(I + T^{-1}\frac{1}{c_k})^{-1}(z^k)$, we have $z^k \in (I + T^{-1}\frac{1}{c_k})(-D_k(z^k))$, and so

$$z^k \in T^{-1}(0) - D_k(z^k) + \frac{\alpha}{c_k}\|D_k(z^k)\|\mathbb{B}.$$

Hence (15) holds. If $T^{-1}(0) = \{\bar{z}\}$, then the lower bound on γ_k follows as in part (i).

(iii) This result follows as in part (ii) using the second remark after Definition 4. □

When $w^k \approx D_k(z^k)$, one can establish the inequality in hypothesis (H2) from a related condition on the vectors w^k. A specific technique for accomplishing this is given in the following lemma.

LEMMA 15. *Let $\xi, \hat{\gamma}_k, \delta_k \in \mathbb{R}_+$ be such that*

$$0 \leq \xi < 1,\ \delta_k \leq \min\left\{1, \|H_k\|^{-1}\right\}\frac{3}{7}(1-\xi)\hat{\gamma}_k,\ \text{and } \hat{\gamma}_k \leq \frac{1}{3}, \tag{16}$$

and let H_k be a continuous linear transformation from $\mathcal{H}$ to itself. If $z^k, w^k \in \mathcal{H}$ satisfy

$$\|(I - H_k)w^k\| \leq \xi\hat{\gamma}_k\|w^k\| \text{ and } \|w^k - D_k(z^k)\| \leq \delta_k\|w^k\|, \tag{17}$$

then $\|(I - H_k)D_k(z^k)\| \leq \hat{\gamma}_k\|D_k(z^k)\|$. Therefore, if (H1) *and criterion* ($\mathcal{L}$) *are satisfied, and if ξ and the sequence $\{(\hat{\gamma}_k, \delta_k)\} \subset \mathbb{R}^2$ satisfy* (16)*, with $\hat{\gamma}_k \leq \gamma_k$ for all k* (*where γ_k is defined in* (H2))*, then hypothesis* (H2) *is satisfied.*

Proof. From (16) and (17), we have

$$\|w^k\| \leq \|D_k(z^k)\| + \|w^k - D_k(z^k)\| \leq \|D_k(z^k)\| + \frac{3}{7}(1-\xi)\hat{\gamma}_k\|w^k\|;$$

hence

$$\|w^k\| \le \frac{1}{1-\frac{3}{7}(1-\xi)\hat{\gamma}_k}\|D_k(z^k)\| \ .$$

Again by (17),

$$\begin{aligned}\|(I-H_k)D_k(z^k)\| &\le \|(I-H_k)w^k\| + \|H_k\|\|w^k - D_k(z^k)\| + \|w^k - D_k(z^k)\| \\ &\le \xi\hat{\gamma}_k\|w^k\| + (\|H_k\|+1)\delta_k\|w^k\| \le \left(\xi + \frac{6}{7}(1-\xi)\right)\hat{\gamma}_k\|w^k\| \\ &\le \frac{\xi+\frac{6}{7}(1-\xi)}{1-\frac{3}{7}(1-\xi)\hat{\gamma}_k}\hat{\gamma}_k\|D_k(z^k)\| \le \hat{\gamma}_k\|D_k(z^k)\|\end{aligned}$$

since the inequality $\hat{\gamma}_k \le \frac{1}{3}$ implies that $\frac{\xi+\frac{6}{7}(1-\xi)}{1-\frac{3}{7}(1-\xi)\hat{\gamma}_k} = \frac{6+\xi}{7-3(1-\xi)\hat{\gamma}_k} \le 1$. □

We conclude this section by showing that the operators $\mathcal{N}_k$ are nonexpansive with respect to the set $T^{-1}(0)$.

PROPOSITION 16. *Assume* $T^{-1}(0)$ *is nonempty. If the sequence of linear transformations* $\{H_k\}$ *satisfies hypotheses* (H1) *and* (H2)*, then for all k we have* $\|H_kD_k(z^k)\| \le \frac{3}{2}\|D_k(z^k)\|$ *and*

$$\|\mathcal{N}_k(z^k)-\bar{z}\|^2 + \frac{\gamma_k^2}{4}\|D_k(z^k)\|^2 \le \|z^k-\bar{z}\|^2 \quad \textit{for all } \bar{z}\in\Gamma. \tag{18}$$

Proof. Let $\bar{z}\in\Gamma$. From the definitions of P_k and $\mathcal{N}_k$, we have

$$\|P_k(z^k)-\bar{z}\| = \|\mathcal{N}_k(z^k)-(H_k-I)D_k(z^k)-\bar{z}\| \ge |\|\mathcal{N}_k(z^k)-\bar{z}\|-\|(H_k-I)D_k(z^k)\||; \tag{19}$$

hence

$$\|P_k(z^k)-\bar{z}\|^2 \ge \|\mathcal{N}_k(z^k)-\bar{z}\|^2+\|(H_k-I)D_k(z^k)\|^2-2\|(H_k-I)D_k(z^k)\|\|\mathcal{N}_k(z^k)-\bar{z}\| \ . \tag{20}$$

From hypothesis (H2), we have

$$\|H_kD_k(z^k)\| \le \|D_k(z^k)\| + \|(H_k-I)D_k(z^k)\| \le (1+\gamma_k)\|D_k(z^k)\| \le \frac{3}{2}\|D_k(z^k)\| \ .$$

Hence

$$\|\mathcal{N}_k(z^k)-\bar{z}\| \le \|z^k-\bar{z}\| + \|H_kD_k(z^k)\| \le \sigma_k + \frac{3}{2}\|D_k(z^k)\| \ .$$

Then, again by hypothesis (H2),

$$\|(H_k-I)D_k(z^k)\| \le \gamma_k\|D_k(z^k)\| = \frac{\|D_k(z^k)\|^2}{2\sigma_k+3\|D_k(z^k)\|} \le \frac{\|D_k(z^k)\|^2}{2\|\mathcal{N}_k(z^k)-\bar{z}\|} \ . \tag{21}$$

Thus, from (20) and (21),

$$\|P_k(z^k)-\bar{z}\|^2 \ge \|\mathcal{N}_k(z^k)-\bar{z}\|^2 + \|(H_k-I)D_k(z^k)\|^2 - \|D_k(z^k)\|^2 \ . \tag{22}$$

Letting $z=z^k$ and $z'=\bar{z}$ in Proposition 3 c) yields

$$\|P_k(z^k)-\bar{z}\|^2 + \|D_k(z^k)\|^2 \le \|z^k-\bar{z}\|^2 \ . \tag{23}$$

From (22) and (23) we have

$$\|\mathcal{N}_k(z^k) - \bar{z}\|^2 + \|(H_k - I)D_k(z^k)\|^2 \le \|z^k - \bar{z}\|^2 \ . \tag{24}$$

We now consider $\alpha_k = \frac{\|(H_k - I)D_k(z^k)\|}{\|D_k(z^k)\|}$. If $\alpha_k \ge \frac{\gamma_k}{2}$, then (18) holds by (24). Suppose that $\alpha_k < \frac{\gamma_k}{2}$. From (19), we have

$$\|P_k(z^k) - \bar{z}\| \ge \|\mathcal{N}_k(z^k) - \bar{z}\| - \frac{\gamma_k}{2}\|D_k(z^k)\| \ .$$

Therefore, by (23),

$$\|\mathcal{N}_k(z^k) - \bar{z}\| \le \sqrt{\|z^k - \bar{z}\|^2 - \|D_k(z^k)\|^2} + \frac{\gamma_k}{2}\|D_k(z^k)\| \ .$$

Using the inequality $\sqrt{a^2 - b^2} \le a - \frac{b^2}{2a}$ for $a > b > 0$,

$$\|\mathcal{N}_k(z^k) - \bar{z}\| \le \|z^k - \bar{z}_k\| - \frac{\|D_k(z^k)\|^2}{2\|z^k - \bar{z}\|} + \frac{\gamma_k}{2}\|D_k(z^k)\| \ .$$

But $\frac{\|D_k(z^k)\|}{2\|z^k - \bar{z}\|} \ge \gamma_k$; thus

$$\|\mathcal{N}_k(z^k) - \bar{z}\| \le \|z^k - \bar{z}\| - \frac{\gamma_k}{2}\|D_k(z^k)\|$$

or

$$\|\mathcal{N}_k(z^k) - \bar{z}\| + \frac{\gamma_k}{2}\|D_k(z^k)\| \le \|z^k - \bar{z}\| \ . \tag{25}$$

From (25) we again obtain (18). □

6. Global convergence. The statement and proof of the global convergence result given below parallels the development given by Rockafellar in [41, Theorem 1] for the classical PPA.

THEOREM 17. *Let $\{z^k\}$ be any sequence generated by the VMPPA under criterion $(\mathcal{G})$ (or $(\mathcal{G}')$). Suppose that the solution set $T^{-1}(0)$ is nonempty and the sequence of linear transformations $\{H_k\}$ satisfies the hypotheses* (H1) *and* (H2). *Then the sequence $\{z^k\}$ is bounded, each weak cluster point of this sequence is an element of $T^{-1}(0)$, and $\lim_k D_k(z^k) = 0$. If it is also assumed that $T^{-1}(0)$ is bounded and $\Gamma = T^{-1}(0)$ in* (H2), *then there is a $\bar{z} \in T^{-1}(0)$ such that $\{z^k\}$ converges weakly to $\bar{z}$.*

In order to establish this result we require the following technical lemma, whose proof is straightforward and so is omitted.

LEMMA 18. *Suppose the nonnegative sequences $\{\epsilon_k\}$ satisfy $\sum_{k=0}^{\infty} \epsilon_k < +\infty$. If $\{u_k\}$ is a nonnegative sequence satisfying $u_{k+1} \le \epsilon_k + u_k$, then $\{u_k\}$ is a Cauchy sequence.*

Proof of Theorem 17. We begin by showing that the limit $\lim_k \|z^k - \bar{z}\| = \mu(\bar{z})$ exists for every $\bar{z} \in \Gamma$. To this end let $\bar{z} \in \Gamma$ and observe that the definition of $\mathcal{N}_k$ and Proposition 16 imply that

$$\begin{aligned}\|z^{k+1} - \bar{z}\| &= \|z^{k+1} - \mathcal{N}_k(z^k) + \mathcal{N}_k(z^k) - \bar{z}\| \le \|z^{k+1} - \mathcal{N}_k(z^k)\| + \|\mathcal{N}_k(z^k) - \bar{z}\| \\ &\le \|H_k\|\|w^k - D_k(z^k)\| + \|z^k - \bar{z}\| \le \epsilon_k + \|z^k - \bar{z}\| \ .\end{aligned}$$

Therefore, Lemma 18 implies that the sequence $\{\|z^k - \bar{z}\|\}$ is Cauchy, and so $\mu(\bar{z})$ exists for every $\bar{z} \in \Gamma$. An immediate consequence of the existence of these limits is the boundedness of the sequences $\{z^k\}$ and σ_k.

We now show that the sequence $\{D_k(z^k)\}$ converges strongly to the origin. Indeed, if this is not the case, then there is a subsequence $J \subset \{1, 2, \ldots\}$ such that $\inf_J \|D_k(z^k)\| = \beta_1 > 0$. This in turn implies that $\inf_J \gamma_k = \beta_2 > 0$ since otherwise $\lim_J \|D_k(z^k)\| = 0$ due to the boundedness of the sequence $\{\sigma_k\}$. Let $\bar{z} \in \Gamma$. By Proposition 16,

$$\begin{aligned}
&\frac{\gamma_k^2}{4}\|D_k(z^k)\|^2 - \|z^k - \bar{z}\|^2 + \|z^{k+1} - \bar{z}\|^2 \le \|z^{k+1} - \bar{z}\|^2 - \|\mathcal{N}_k(z^k) - \bar{z}\|^2 \\
&= \langle z^{k+1} - \mathcal{N}_k(z^k), z^{k+1} - \bar{z} + \mathcal{N}_k(z^k) - \bar{z}\rangle \\
&\le \|z^{k+1} - \mathcal{N}_k(z^k)\|(\|z^{k+1} - \bar{z}\| + \|\mathcal{N}_k(z^k) - \bar{z}\|) \\
&\le \|H_k\|\|w^k - D_k(z^k)\|(\|z^{k+1} - \bar{z}\| + \|z^k - \bar{z}\|) \le \epsilon_k(\|z^{k+1}\| + 2\|\bar{z}\| + \|z^k\|) = \epsilon_k C_k \ ,
\end{aligned}$$

with $\{C_k\}$ bounded, where the final inequality follows from criterion $(\mathcal{G})$. Hence

$$\frac{\gamma_k^2}{4}\|D_k(z^k)\|^2 \le \|z^k - \bar{z}\|^2 - \|z^{k+1} - \bar{z}\|^2 + \epsilon_k C_k,$$

whereby we obtain the contradiction

$$\begin{aligned}
0 < \frac{\beta_1^2\beta_2^2}{4} &\le \limsup_J \frac{\gamma_k^2}{4}\|D_k(z^k)\|^2 \\
&\le \lim_J(\|z^k - \bar{z}\|^2 - \|z^{k+1} - \bar{z}\|^2 + \epsilon_k C_k) = \mu(\bar{z}) - \mu(\bar{z}) + 0 = 0 \ .
\end{aligned}$$

Therefore, $\lim_k \|D_k(z^k)\| = 0$.

Next let $J \subset \{1, 2, \ldots\}$ be such that the subsequence $\{z^k\}_J$ converges weakly to z^∞, i.e., z^∞ is a weak cluster point of the sequence $\{z^k\}$. We show that z^∞ must be an element of $T^{-1}(0)$. From Proposition 3 a), we have that $-\frac{1}{c_k}D_k(z^k) \in T(P_k(z^k))$ for all k; hence $0 \le \langle z - P_k(z^k), w + \frac{1}{c_k}D_k(z^k)\rangle$, or equivalently, $\langle z - z^k - D_k(z^k), w + \frac{1}{c_k}D_k(z^k)\rangle \ge 0$ for all k and z, w with $w \in T(z)$. Taking the limit over J yields the inequality $\langle z - z^\infty, w\rangle \ge 0$ for all z, w with $w \in T(z)$. Since T is maximal monotone, we get $0 \in T(z^\infty)$.

Under the assumption that $\Gamma = T^{-1}(0)$, the argument showing that there is no more than one weak cluster point of $\{z^k\}$ is identical to the one given by Rockafellar in [41, Theorem 1]. □

Remark. To ensure the strong convergence of the sequence $\{z^k\}$, one again requires a growth condition on the inverse mapping T^{-1} in a neighborhood of the origin. Rockafellar has shown that Lipschitz continuity of T^{-1} at the origin suffices for this purpose [41, Theorem 2]. Other conditions can be found in the work of Luque [21, Proposition 1.2]. The results of Rockafellar and Luque are easily extended to the VMPPA.

7. Convergence rates.

7.1. Linear convergence. Just as in Rockafellar [41, Theorem 2], we require that the operator T^{-1} is Lipschitz continuous at the origin in order to establish that the convergence rate is at least linear.

THEOREM 19. *Let $\{z^k\}$ be any sequence generated by the VMPPA satisfying both criteria $(\mathcal{G})$ and $(\mathcal{L})$ for all k. Assume that T^{-1} is Lipschitz continuous at the* *origin with modulus α and the solution set $T^{-1}(0)$ is a singleton $\{\bar{z}\}$. If the sequence $\{H_k\}$ satisfies the hypotheses* (H1) *and* (H2) *with $\delta_k\|H_k\| \to 0$, then the sequence $\{z^k\}$ strongly converges to the solution and there is an index $\bar{k}$ such that*

$$\|z^{k+1} - \bar{z}\| \leq \sigma_k \|z^k - \bar{z}\| \quad \textit{for all } k \geq \bar{k} \ ,$$

where σ_k satisfies $\limsup_{k\to\infty} \sigma_k < 1$. That is, the convergence rate is linear.

Proof. By Theorem 17, we have $\|D_k(z^k)\| \to 0$. Hence, Part (ii) of Lemma 14 implies that $\{z^k\}$ converges strongly to $\bar{z}$. We now establish the linear rate.

Let $\tau > 0$ be as in Definition 4, and let $\tilde{k}$ be such that $\|\frac{1}{c_k} D_k(z^k)\| \leq \tau$ for all $k \geq \tilde{k}$. By Proposition 3 a) and the Lipschitz continuity of T^{-1} at 0, we have

$$\|P_k(z^k) - \bar{z}\| \leq \frac{\alpha}{c_k} \|D_k(z^k)\| \ . \tag{26}$$

Hence relation (14) and hypothesis (H2) yield

$$\begin{aligned} \|\mathcal{N}_k(z^k) - \bar{z}\| &= \|P_k(z^k) + (H_k - I)D_k(z^k) - \bar{z}\| \\ &\leq \|P_k(z^k) - \bar{z}\| + \gamma_k \|D_k(z^k)\| \ . \end{aligned} \tag{27}$$

Let $a_k := \frac{\alpha}{c_k} + \gamma_k$. Using (26) and (27),

$$\|\mathcal{N}_k(z^k) - \bar{z}\| \leq \left(\frac{\alpha}{c_k} + \gamma_k\right) \|D_k(z^k)\| = a_k \|D_k(z^k)\| \ . \tag{28}$$

Let $\gamma := \frac{1}{2(5+2\alpha)}$. By Proposition 16 and Lemma 14 we have, for $k \geq \tilde{k}$, that

$$\|\mathcal{N}_k(z^k) - \bar{z}\|^2 + \gamma^2 \|D_k(z^k)\|^2 \leq \|z^k - \bar{z}\|^2 \ . \tag{29}$$

By (28) and (29), when $k \geq \tilde{k}$,

$$\|\mathcal{N}_k(z^k) - \bar{z}\|^2 \leq {a_k}^2 \|D_k(z^k)\|^2 \leq \frac{{a_k}^2}{\gamma^2} \|z^k - \bar{z}\|^2 - \frac{{a_k}^2}{\gamma^2} \|\mathcal{N}_k(z^k) - \bar{z}\|^2 \ . \tag{30}$$

Let $\mu_k := \frac{a_k}{\sqrt{a_k^2 + \gamma^2}}$. From (30) we have

$$\|\mathcal{N}_k(z^k) - \bar{z}\| \leq \mu_k \|z^k - \bar{z}\| \ . \tag{31}$$

By (31), criterion ($\mathcal{L}$) (or ($\mathcal{L}'$)), and Proposition 3 c),

$$\begin{aligned} &\|z^{k+1} - \bar{z}\| \leq \|z^{k+1} - \mathcal{N}_k(z^k)\| + \|\mathcal{N}_k(z^k) - \bar{z}\| \\ &\leq \delta_k \|H_k\| \|w^k\| + \mu_k \|z^k - \bar{z}\| \leq \frac{\delta_k \|H_k\|}{1 - \delta_k} \|D_k(z^k)\| + \mu_k \|z^k - \bar{z}\| \\ &\leq \left(\frac{\delta_k \|H_k\|}{1 - \delta_k} + \mu_k\right) \|z^k - \bar{z}\| = \sigma_k \|z^k - \bar{z}\| \ , \end{aligned}$$

where $\sigma_k := \frac{\delta_k \|H_k\|}{1-\delta_k} + \mu_k$. Since there is a $\tilde{\delta} > 0$ such that $\mu_k < 1 - \tilde{\delta}$ for any k, and $\delta_k \|H_k\| \to 0$, we have $\sigma_k < 1$ for k sufficiently large. Moreover, we have $\limsup_{k\to\infty} \sigma_k = \limsup_{k\to\infty} \mu_k \leq 1 - \tilde{\delta}$. □

7.2. Superlinear convergence. We now give an analogue of Dennis and Moré's [14] characterization theorem for the superlinear convergence of variable metric methods in nonlinear programming that applies to the VMPPA. This result is used in [6] to establish the superlinear convergence of the VMPPA when the Broyden (nonsymmetric case) or the BFGS (symmetric case) updating formula is used to generate the matrices H_k.

THEOREM 20. *Let $\{z^k\}$ be any sequence generated by the VMPPA satisfying criterion $(\mathcal{L})$ for all k. Suppose that the operator T^{-1} is differentiable at the origin with $T^{-1}(0) = \{\bar{z}\}$ and $\nabla T^{-1}(0) = J$. If* $\lim_k \|D_k(z^k)\| = 0$, *then $\{z^k\}$ converges to the solution $\bar{z}$ superlinearly if and only if*

$$\frac{[I - (I + \frac{1}{c_k}J)H_k^{-1}](z^{k+1} - z^k)}{\|z^{k+1} - z^k\|} \to 0 \quad as\ k \to \infty\ . \tag{32}$$

Remark. By Proposition 9 we have $\nabla D(\bar{z}) = -(I + \frac{1}{c}J)^{-1}$. Consequently, condition (32) can be recast in the more familiar form given in [15, Theorem 8.2.4]. Note that the assumption in (32) on the sequence $\{H_k\}$ is much weaker than assuming that this sequence converges. Specific choices of the linear transformations H_k satisfying (32) are discussed in [6].

The proof of Theorem 20 requires the following lemma.

LEMMA 21. *Under the conditions in Theorem* 20 *we have*

(a) $T^{-1}(\frac{-1}{c_k}D_k(z^k)) - \bar{z} - J(\frac{-1}{c_k}D_k(z^k)) \subset o(\|z^k - \bar{z}\|)\mathbb{B}$, *and*

(b) $(I + \frac{1}{c_k}J)H_k^{-1}(z^{k+1} - \mathcal{N}_k(z^k)) \in o(\|z^k - \bar{z}\|)\mathbb{B}$,

for all k sufficiently large.

Proof. For part (a), let $\delta > 0$ be such that

$$T^{-1}(w) - Jw - \bar{z} \subset o(\|w\|)\mathbb{B} \tag{33}$$

whenever $\|w\| < \delta$. Let $\bar{k}_1$ be such that whenever $k > \bar{k}_1$, $\|D_k(z^k)\| \leq \delta$. Then, by (33) and Proposition 3 c), when $k > \bar{k}_1$,

$$T^{-1}\left(\frac{-1}{c_k}D_k(z^k)\right) - \bar{z} - J\left(\frac{-1}{c_k}D_k(z^k)\right) \subset o(\|D_k(z^k)\|)\mathbb{B} \subset o(\|z^k - \bar{z}\|)\mathbb{B}\ .$$

We now prove (b). Note that $\mathcal{N}_k(z^k) = (I + H_kD_k)(z^k)$; hence by criterion $(\mathcal{L})$

$$\begin{aligned}\left\|\left(I + \frac{1}{c_k}J\right)H_k^{-1}(z^{k+1} - \mathcal{N}_k(z^k))\right\| &= \left\|\left(I + \frac{1}{c_k}J(w^k - D_k(z^k)\right)\right\| \\ &\leq (1 + \|J\|)\|w^k - D_k(z^k)\| \leq \delta_k(1 + \|J\|)\|w^k\| \\ &\leq \frac{\delta_k(1 + \|J\|)}{1 - \delta_k}\|D_k(z^k)\|\ . \end{aligned} \tag{34}$$

Therefore by (34) and Proposition 3 c),

$$\left(I + \frac{1}{c_k}J\right)H_k^{-1}(z^{k+1} - \mathcal{N}_k(z^k)) \in o(\|D_k(z^k)\|)\mathbb{B} \subset o(\|z^k - \bar{z}\|)\mathbb{B}. \qquad \square$$

Proof of Theorem 20. Let $\tilde{z}^{k+1} := \mathcal{N}_k(z^k) = (I + H_kD_k)(z^k)$. By Proposition 3 a) we have $\tilde{z}^{k+1} = z^k - H_k(I + T^{-1}\frac{1}{c_k})^{-1}(z^k)$. Hence

$$\begin{aligned} z^k &\in \left(I + T^{-1}\frac{1}{c_k}\right)[H_k^{-1}(z^k - \tilde{z}^{k+1})] \\ &= H_k^{-1}(z^k - \tilde{z}^{k+1}) + T^{-1}\left[\frac{1}{c_k}H_k^{-1}(z^k - \tilde{z}^{k+1})\right]\ , \end{aligned}$$

or equivalently,

$$\begin{aligned}
z^{k+1} - \bar{z} &= z^k - \bar{z} + (z^{k+1} - z^k) \\
&\in \left[T^{-1}\left(\frac{1}{c_k} H_k^{-1}(z^k - \tilde{z}^{k+1})\right) - \bar{z} + (z^{k+1} - z^k) + H_k^{-1}(z^k - \tilde{z}^{k+1})\right] \\
&= \left[T^{-1}\left(\frac{1}{c_k} H_k^{-1}(z^k - \tilde{z}^{k+1})\right) - \bar{z} - J\left(\frac{1}{c_k} H_k^{-1}(z^k - \tilde{z}^{k+1})\right)\right] \\
&\quad + \left[I - \left(I + \frac{1}{c_k} J\right) H_k^{-1}\right](z^{k+1} - z^k) \\
&\quad + \left(I + \frac{1}{c_k} J\right) H_k^{-1}(z^{k+1} - \tilde{z}^{k+1}) \\
&= \left[T^{-1}\left(\frac{-1}{c_k} D_k(z^k)\right) - \bar{z} - J\left(\frac{-1}{c_k} D_k(z^k)\right)\right] \\
&\quad + \left[I - \left(I + \frac{1}{c_k} J\right) H_k^{-1}\right](z^{k+1} - z^k) \\
&\quad + \left(I + \frac{1}{c_k} J\right) H_k^{-1}(z^{k+1} - \tilde{z}^{k+1}) \ .
\end{aligned} \tag{35}$$

By Lemma 21 the first and third of the three terms appearing on the right-hand side of this inclusion can be bounded by an expression of the form $o(\|z^k - \bar{z}\|)\mathbb{B}$. If (32) holds, then $[I - (I + \frac{1}{c_k} J)H_k^{-1}](z^{k+1} - z^k) \in o(\|z^{k+1} - z^k\|)\mathbb{B}$. Therefore there are positive sequences $\{\alpha_{1k}\}$ and $\{\alpha_{2k}\}$, each converging to zero such that, for $k > \bar{k}_1$,

$$\begin{aligned}
\|z^{k+1} - \bar{z}\| &\le \alpha_{1k}\|z^{k+1} - z^k\| + \alpha_{2k}\|z^k - \bar{z}\| \\
&\le \alpha_{1k}(\|z^k - \bar{z}\| + \|z^{k+1} - \bar{z}\|) + \alpha_{2k}\|z^k - \bar{z}\| \\
&= \alpha_{1k}\|z^{k+1} - \bar{z}\| + (\alpha_{1k} + \alpha_{2k})\|z^k - \bar{z}\|.
\end{aligned}$$

Let $\bar{k}_2 > \bar{k}_1$ be such that $\alpha_{1k} < \frac{1}{2}$ for all $k > \bar{k}_2$. Then, denoting $\frac{\alpha_{1k} + \alpha_{2k}}{1 - \alpha_{1k}}$ by τ_k,

$$\|z^{k+1} - \bar{z}\| \le \frac{\alpha_{1k} + \alpha_{2k}}{1 - \alpha_{1k}}\|z^k - \bar{z}\| = \tau_k\|z^k - \bar{z}\|$$

whenever $k > \bar{k}_2$, and $\tau_k \to 0$ as $k \to \infty$. Therefore $\{z^k\}$ converges to $\bar{z}$ superlinearly.

Conversely, suppose that

$$\lim_{k \to \infty} \frac{\|z^{k+1} - \bar{z}\|}{\|z^k - \bar{z}\|} = 0 \ . \tag{36}$$

Divide (35) by $\|z^k - \bar{z}\|$ and let $k \to \infty$. From (36) and Lemma 21 we obtain

$$\frac{[I - (I + \frac{1}{c_k} J)H_k^{-1}](z^{k+1} - z^k)}{\|z^k - \bar{z}\|} \to 0 \quad \text{as } k \to \infty \ .$$

However, from (36) we have

$$\frac{\|z^k - \bar{z}\|}{\|z^{k+1} - z^k\|} \le \frac{\|z^k - \bar{z}\|}{\|z^k - \bar{z}\| - \|z^{k+1} - \bar{z}\|} = \frac{1}{1 - \frac{\|z^{k+1} - \bar{z}\|}{\|z^k - \bar{z}\|}} \to 1$$

as $k \to \infty$. Hence (32) holds. □

8. Concluding remarks. In this paper, we introduced a new PPA for solving the inclusion $0 \in T(x)$, where T is an arbitrary maximal monotone operator. The global convergence of the algorithm is demonstrated with an inexact solution at each step. This is important in practice, since solving for the exact solution at each step is impractical and may in fact be almost as difficult as solving the original problem. If it is assumed that T^{-1} is Lipschitz continuous at the origin, then the method is shown to be linearly convergent. If it is further assumed that T^{-1} is differentiable at the origin, then the classical characterization of superlinear convergence due to Dennis and Moré also holds for the VMPPA. In [6], this characterization of superlinear convergence is applied to establish the super-linear convergence of the method when certain matrix secant updating strategies are employed to generate the matrices H_k. In [5], we give some of the implementation details in the case of convex programming. We show how to apply the method to solve the associated primal, dual, and Lagrangian saddle point problems. In particular, it is shown how the bundle technique [17] can be applied to satisfy the approximation criteria ($\mathcal{L}$) and ($\mathcal{G}$) in both the primal and saddle point solution techniques. Preliminary numerical results comparing these three approaches are also presented.

Acknowledgments. The authors would like to thank the reviewers for their thorough work. Their comments and suggestions have greatly contributed to our exposition. In particular, we would like to thank Professor Alfredo Iusem for observing an error in an earlier version of Theorem 17 and for his suggested revision of this result when the set $T^{-1}(0)$ is assumed to be bounded.

REFERENCES

[1] A.D. Alexandrov, *The existence almost everywhere of the second differential of a convex function and some associated properties of convex surfaces*, Ucenye Zapiski Leningr. Gos. Univ. Ser. Mat., 37 (1939), pp. 3–35 (in Russian).

[2] J.P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.

[3] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C. Sagastizábal, *A family of variable metric proximal point methods*, Math. Programming, 68 (1995), pp. 15–47.

[4] L.M. Bregman, *The method of successive projection for finding a common point of convex sets*, Soviet Mathematics Doklady, 162 (1965), pp. 487–490.

[5] J.V. Burke and M. Qian, *On the local super-linear convergence of a matrix secant implementation of the variable metric proximal point algorithm for monotone operators*, in Reformulation—Nonsmooth, Piecewise Smooth, Semi-smooth, and Smoothing Methods, L. Qi and M. Fukushima, eds., Kluwer Academic Publishers, Norwell, MA, 1998, pp. 317–334.

[6] J.V. Burke and M. Qian, *On the super-linear convergence of the variable metric proximal point algorithm using Broyden and BFGS matrix secant updating*, Math. Programming, 1999, to appear.

[7] X. Chen and M. Fukushima, *Proximal Quasi-Newton Methods for Nondifferentiable Convex Optimization*, Math. Programming, 1999, to appear.

[8] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, New York, 1980.

[9] J. Eckstein, *Splitting Methods for Monotone Operators with Application to Parallel Optimization*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1989.

[10] M. Fukushima and L. Qi, *A globally and superlinearly convergent algorithm for nonsmooth convex minimization*, SIAM J. Optim., 6 (1996), pp. 1106–1120.

[11] O. Güler, *New proximal point algorithms for convex minimization*, SIAM J. Optim., 2 (1992), pp. 649–664.

[12] S. Han, *A decomposition method and its application to convex programming*, Math. Oper. Res., 14 (1989), pp. 237–248.

[13] A. Iusem, private communication, IMPA, Rio de Janeiro, Brazil, 1996.

[14] J.E. Dennis, Jr., and J.J. Moré, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 549–560.

[15] J.E. Dennis, Jr., and R.B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[16] G. Kassay, *The proximal points algorithm for reflexive Banach spaces*, Studia Univ. Babes-Bolyai Math., 30 (1930), pp. 9–17.

[17] C. Lemaréchal, *Bundle methods in nonsmooth optimization*, in Nonsmooth Optimization, C. Lemaréchal and R. Mifflin, eds., Pergamon Press, Oxford, 1978.

[18] C. Lemaréchal and C. Sagastizábal, *An approach to variable metric bundle methods*, in IFIP Proceedings, Systems Modeling and Optimization, J. Henry and J.P. Yuan, eds., Springer, Berlin, 1994, pp. 144–162.

[19] C. Lemaréchal and C. Sagastizábal, *Practical aspects of the Moreau–Yosida regularization: Theoretical preliminaries*, SIAM J. Optim., 7 (1997), pp. 367–385.

[20] C. Lemaréchal and C. Sagastizábal, *Variable metric bundle methods: From conceptual to implementable forms*, Math. Programming, 76 (1997), pp. 393–410.

[21] F.J. Luque, *Asymptotic convergence analysis of the proximal point algorithm*, SIAM J. Control Optim., 22 (1984), pp. 277–293.

[22] B. Martinet, *Regularisation d'inequations variationelles par approximations successive*, Revue Française d'Informatique et de Recherche Opérationelle, 4 (1970), pp. 154–158.

[23] B. Martinet, *Determination approachée d'un point fixe d'une application pseudo-contractante. cas de l'application prox*, Comptes Rendus de l'Académie des Sciences, Paris, Série A, 274 (1972), pp. 163–165.

[24] R. Mifflin, *A quasi–second-order proximal bundle algorithm*, Math. Programming, 73 (1996), pp. 51–72.

[25] R. Mifflin, D. Sun, and L. Qi, *Quasi-Newton Bundle-Type Methods for Nondifferentiable Convex Optimization*, Technical report AMR 96/21, Dept. of Applied Math., University of New South Wales, Sydney, New South Wales, Australia, 1996.

[26] F. Mignot, *Control dan les inequations variationelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 130–185.

[27] G.J. Minty, *Monotone (nonlinear) operators in Hilbert space*, Duke Math. J., 29 (1962), pp. 341–346.

[28] J.J. Moreau, *Proximité et dualité dans un espace Hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.

[29] J.M. Ortega and W.G. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[30] J.-S. Pang and L. Qi, *Nonsmooth equations: Motivation and algorithms*, SIAM J. Optim., 3 (1993), pp. 443–465.

[31] J.-S. Pang and L. Qi, *A globally convergent Newton method for SC^1 problems*, J. Optim. Theory Appl., 85 (1995), pp. 633–648.

[32] G.B. Passty, *Weak convergence theorems for nonexpansive mappings in Banach spaces*, J. Math. Anal. Appl., 67 (1979), pp. 274–276.

[33] R.R. Phelps, *Convex Functions, Monotone Operators, and Differentiability*, Lecture Notes in Math., Springer-Verlag, New York, 1989.

[34] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.

[35] L. Qi, *Second-order analysis of the Moreau-Yosida regularization of a convex function*, Technical report AMR 94/20, Dept. of Applied Math., University of New South Wales, Sydney, New South Wales, Australia, 1994.

[36] L. Qi and X. Chen, *A preconditioning proximal Newton method for nondifferentiable convex optimization*, Math. Programming, 76 (1995), pp. 411–430.

[37] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 66 (1994), pp. 25–43.

[38] M. Qian, *The Variable Metric Proximal Point Algorithm: Theory and Application*, Ph.D. thesis, University of Washington, Seattle, WA, 1992.

[39] R.T. Rockafellar, *Conjugate Duality and Optimization*, SIAM, Philadelphia, 1974.

[40] R.T. Rockafellar, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.

[41] R.T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.

[42] R.T. Rockafellar, *Maximal monotone relations and the second derivatives of nonsmooth functions*, Ann. Inst. H. Poincaré Analyse Non Linéaire, 2 (1985), pp. 167–184.

[43] J.E. Spingarn, *Partial inverse of a monotone operator*, Appl. Math. Optim., 10 (1983), pp. 247–265.

[44] J.E. Spingarn, *Applications of the methods of partial inverses to convex programming: Decomposition*, Math. Programming, 32 (1985), pp. 199–223.

SIAM J. Control Optim.
Vol. 37, No. 2, pp. 376–387

OPTIMIZING THE RATE OF DECAY OF SOLUTIONS OF THE WAVE EQUATION USING GENETIC ALGORITHMS: A COUNTEREXAMPLE TO THE CONSTANT DAMPING CONJECTURE*

PEDRO FREITAS†

Abstract. We consider the problem of optimizing the rate of decay of solutions of the linear damped wave equation on a bounded interval. This corresponds to optimizing the spectral abscissa of the associated linear operator. By writing the damping term as a Fourier cosine series and obtaining some inequalities that the coefficients in this series have to satisfy in order that the spectral abscissa be larger than a real number α, we are then able to use a genetic algorithm to obtain values of the spectral abscissa which are better than those given by the constant damping case. This provides a counterexample to the conjecture that the best possible decay was obtained for constant damping.

Key words. damped wave equation, spectral abscissa, genetic algorithms

AMS subject classifications. 35B20, 35L05, 47A55, 68T05

PII. S0363012997329445

1. Introduction.

Let $\mathcal{H}$ be a Hilbert space and consider the system of differential equations

$$Au_{tt} + Bu_t + Cu = 0, \tag{1.1}$$

supplemented by initial conditions, and where A, B, and C are self-adjoint linear operators whose domain is dense in $\mathcal{H}$, with A and C uniformly positive. When B is the zero operator, and under suitable conditions on A and C, solutions of (1.1) oscillate in time. Otherwise, the system is said to be damped, and depending on the choice of B, solutions will converge to zero, become unbounded, or may still display an oscillatory behavior [4, 5].

When B is a positive operator, the system is dissipative and solutions will, in general, converge to zero. From the point of view of applications, the rate at which this happens is quite important, and one normally distinguishes between two cases, depending on the existence or not of real eigenvalues of the associated eigenvalue problem. When such eigenvalues are not present, solutions will oscillate to zero. If the operator B is large enough in some sense, then there will exist modes which are associated with real eigenvalues and whose corresponding solutions are of the form $e^{-\alpha t}v$, where α is some positive real number and v is an element of $\mathcal{H}$.

The effect caused by increasing the damping can be illustrated by considering the very simple case of the second-order ordinary differential equation

$$u'' + 2bu' + cu = 0, \tag{1.2}$$

where b is now a real parameter and c is a fixed positive number. In this case it is well known that while for both very small and very large values of b the decay of solutions

*Received by the editors November 3, 1997; accepted for publication (in revised form) March 26, 1998; published electronically November 23, 1998. This research was partially supported by FCT (Portugal), under projects PRAXIS/PCEX/P/MAT/36/96 and PRAXIS XXI Project 2/2.1/MAT/199/94.

http://www.siam.org/journals/sicon/37-2/32944.html

†Departamento de Matemática, Instituto Superior Técnico, Av. Rovisco Pais, 1096 Lisboa Codex, Portugal (pfreitas@math.ist.utl.pt).

is quite slow, there exists a critical value for which this rate of decay is optimal. This value is given by $b_o = c^{1/2}$ and corresponds exactly to a double real eigenvalue at the point of transition between complex and real eigenvalues. For values of b larger than b_o, the system is said to be overdamped.

This very simple situation suggests the following problem for the general case of (1.1):

> Consider (1.1) and assume that both A and C are given, together with a set $\mathcal{B}$ of operators. Maximize over $\mathcal{B}$ the values of ω for which all solutions of (1.1) satisfy
>
> $$\|u\|_{\mathcal{H}} \leq ke^{(\delta-\omega)t}$$
>
> for all positive values of t and all positive δ, where k is a positive constant. Describe the subset of $\mathcal{B}$ corresponding to the optimum.

The reason for including the positive number δ in this description of the problem is that, as can be seen in the simple example above, the optimal may correspond to a multiple eigenvalue for which the corresponding solutions are of the form $(c_0 + c_1 t + \cdots + c_m t^m)e^{-\omega_o t}v$, for some v in $\mathcal{H}$. If this is the case, then the inequality above might not hold when δ is zero.

The previous simple one-dimensional example also highlights the fact that although (1.1) is linear in u, the problem just described is nonlinear. Even in the general case of a finite-dimensional space $\mathcal{H}$, where A and C are real positive $n \times n$ matrices, the problem is already much more complex than is the case for (1.2). This situation has been addressed in [7], where it was shown that the optimal value ω is given by

$$\omega_o = \left[\frac{\det(C)}{\det(A)}\right]^{1/(2n)}$$

and that this optimum is attained. A full description of the class of symmetric matrices which are optimum is, however, unknown, except in the case where n is two.

As we have seen from the previous examples, this type of problem is related to the maximization of the minimum of the negative of the real parts of the eigenvalues of the quadratic pencil defined on $\mathcal{H}$ by

$$T(\lambda) = \lambda^2 A + \lambda B + C.$$

This value is usually referred to as the spectral abscissa of the pencil $T(\lambda)$ and, in general, it can only be guaranteed that it is greater than or equal to ω_o (see [2] and the references therein). Thus, in this version of the optimization problem the objective is, given the operators A and C, to maximize the spectral abscissa and describe the subset of operators B which are optimal.

When $\mathcal{H}$ is an infinite-dimensional Hilbert space, one of such optimization problems is related to the wave equation of the form

$$u_{tt} + 2b(x)u_t = u_{xx} \tag{1.3}$$

on the interval $J = (0, \pi)$, together with initial conditions and homogeneous Dirichlet boundary conditions, and with $b \in L^\infty(0, \pi)$. This problem and the related question of studying the distribution of the eigenvalues of the corresponding quadratic pencil have received much attention in the literature recently [1, 2, 3, 6, 8]. In [3], for instance, it has been proven that within the class of all damping terms which satisfy

$0 \leq b \leq 1$, the constant damping term $b(x) \equiv 1$ is optimal. Later, it was shown in [2] for the case of a general open bounded connected set D of $\mathbb{R}^n$ that the constant damping term is a critical point of the function

$$\Omega(b) = \inf_{\lambda \in \Sigma_T} \left[-\mathrm{Re}(\lambda)\right],$$

where Σ_T denotes the spectrum of the pencil T. More precisely, letting b_o denote the square root of the first eigenvalue of $-\Delta$ on D, and given b_1 in $L^\infty(D)$, they showed that the function $\epsilon \to \mu(\epsilon) = \Omega(b_o + \epsilon b_1)$ is strictly decreasing in $[0, \delta]$ for some positive δ. However, due to a lack of a uniform lower bound for $\delta = \delta(b_1)$, this does not necessarily imply that it is a local optimum. The authors then conjectured that this was in fact the optimal value of the spectral abscissa and that the optimal damping was constant. More recently, some numerical evidence was presented in [1] to support this conjecture.

The main purpose of this paper is to give numerical evidence, supported by some rigorous results, that this conjecture is in fact false and that there exist functions b which give a better decay than the constant damping.

In order to do this, we shall perform a spectral approximation of (1.3) based on a Fourier sine series of the unknown and then study the corresponding optimization problem in finite dimension. It should be remarked that this last step is not equivalent to solving the finite-dimensional problem mentioned above, as now the entries of the matrix obtained in the discretization process are not independent. The reason for opting for a spectral method lies in the fact that we are interested in a *global* property of the spectrum and not just in a small number of eigenvalues.

A crucial step in the whole process is the use of a Fourier cosine series for the damping coefficient b. This makes the equation resemble the well-known (complex) Hill equation, except that here we are dealing with different boundary conditions. The fundamental role of the first of the coefficients in the series—the average—in the asymptotic distribution of the high frequencies is well known (see, for instance, [3, 8]). Here we show that the other coefficients also play an important part and that an adequate way of looking at this problem in one dimension is to consider the effect of each of these coefficients on the spectrum. In particular, one of the main results in the paper shows that in order to increase the value of the spectral abscissa more and more terms in the series have to be considered.

One aspect of the optimization process that also has to be taken into consideration at this point is that the optimal value in such problems quite often corresponds to eigenvalues of high multiplicity. In the finite-dimensional problem, for instance, the optimum is achieved precisely when the spectrum consists of a single eigenvalue of maximum multiplicity. This implies that the function Ω described above will not be differentiable at such points. Furthermore, it may even happen that there will exist only one Jordan block associated with this eigenvalue, as can again be the case in the finite-dimensional problem considered in [7].

All this means that the choice of algorithms used for the optimization might be critical. Taking into account that the structure of this type of problem is not very well understood, and that in the case of (1.3) it is not even known if a finite optimum exists, the choice to use a genetic algorithm in the numerical optimization procedure arises naturally. These algorithms tend not to give the actual optimum, but, on the other hand, they quickly improve the starting value. Since the main purpose here is to provide a counterexample, this seems to be the right choice. Also for the same reason, we shall not go into much detail about the algorithm, although the main steps will be indicated. For a comprehensive treatment of this type of algorithms see, for instance, the books by Holland [10], Goldberg [9], or Michalewicz [12].

The paper is divided into three parts. In the first part (sections 2, 3, and 4), we state the problem and establish some rigorous results which in some sense show the direction that should be pursued in the numerical study. In the second part (sections 5 and 6), we briefly describe the algorithm and present the numerical results obtained. Finally, in section 7 we discuss the results obtained.

2. Statement of the problem and main results. Consider the self-adjoint quadratic pencil in $L^2(0,\pi)$

$$T(\lambda) = \lambda^2 I + 2\lambda b(x) - \frac{\partial^2}{\partial x^2}$$

with domain $\mathcal{H} = H^2(0,\pi) \cap H_0^1(0,\pi)$, and denote by Σ_T its spectrum, that is, the set of complex values of λ for which T does not have a bounded inverse with dense domain. In this instance, the spectrum is formed only by eigenvalues:

$$\Sigma_T = \{\lambda \in \mathbb{C} : T(\lambda)u = 0,\ u \neq 0\}.$$

We now define the function $\Omega : L^\infty \to \mathbb{R}$ by

$$\Omega(b) = \inf_{\lambda \in \Sigma_T} \left[-\mathrm{Re}(\lambda)\right].$$

For each function b in L^∞, the function above gives the corresponding spectral abscissa. In the one-dimensional case, this coincides with the uniform rate of decay of the solutions of (1.3) (see, for instance, [3]), and so by maximizing Ω we are optimizing the asymptotic rate of decay of solutions. The optimization problem then reduces to finding

$$\omega_o = \sup_{b \in L^\infty} \left[\Omega(b)\right]. \tag{2.1}$$

Although there are some results when further restrictions are imposed on the set where b is allowed to vary (see [3, 1]), for the full general problem described here, it is not even known whether ω_o is finite or not.

The procedure used here highlights the fact that the coefficients in the Fourier cosine series of the damping term are fundamental in the values that the spectral abscissa might take, and the main results of the paper along these lines are given in Theorem 4.2 and Corollaries 4.3 and 4.4. In particular, Corollary 4.3 gives intervals where each of the terms up to a certain order in the Fourier cosine series have to lie for the spectral abscissa to be larger than a certain value. This also establishes a *hierarchy* within these coefficients, showing their relative importance; note that this is not the case for, say, the Fourier sine coefficients of b.

Taking these results into account, we then obtain numerical evidence that the constant damping is in fact not optimal, by producing several examples with different sets of nonzero Fourier coefficients for which the corresponding spectral abscissa improves the value obtained for $b(x) \equiv 1$. These numerical results can be complemented with the asymptotics for the spectrum from [3, 8] to provide reliable evidence of this fact.

3. The discretized problem. We discretize (1.3) by means of a Fourier sine series on the interval $(0,\pi)$, in order to obtain a finite-dimensional approximation to the optimization problem (2.1). Writing

$$u(x,t) = \sum_{k=1}^{\infty} u_k(t)\sin(kx)$$

with

$$u_k(t) = \frac{2}{\pi}\int_0^{\pi} u(x,t)\sin(kx)dx,$$

we are thus led to the following infinite system of ordinary differential equations for the Fourier coefficients of u:

$$u_k''(t) + 2\sum_{j=1}^{\infty} b_{jk}u_k'(t) + k^2 u_k(t) = 0,$$

where

$$b_{jk} = b_{kj} = \frac{2}{\pi}\int_0^{\pi} b(x)\sin(jx)\sin(kx)dx.$$

So that we can obtain a manageable expression for these coefficients, we now expand the function b into a Fourier cosine series, that is,

$$b(x) = \frac{b_0}{2} + \sum_{k=1}^{\infty} b_k\cos(kx),$$

with

$$b_k = \frac{2}{\pi}\int_0^{\pi} b(x)\cos(kx)dx,\ k = 0, 1, \ldots.$$

Then

$$\begin{aligned} b_{jk} &= \frac{b_0}{2}\delta_{jk} + \frac{2}{\pi}\sum_{m=1}^{\infty} b_m \int_0^{\pi}\cos(mx)\sin(jx)\sin(kx)dx \\ &= \frac{b_0}{2}\delta_{jk} + \frac{1}{\pi}\sum_{m=1}^{\infty} b_m\left[\int_0^{\pi}\cos(mx)\left[\cos((k-j)x) - \cos((k+j)x)\right]\right]dx \\ &= \frac{b_{|k-j|} - b_{k+j}}{2}, \end{aligned}$$

where δ_{jk} denotes the Kronecker symbol. This can be summarized in the following proposition.

Proposition 3.1. *Let b_k, $k = 0, 1, \ldots,$ denote the Fourier cosine coefficients of the function b on the interval $(0,\pi)$. Then the nth-dimensional truncation of the discretized wave equation can be written as*

$$U_n''(t) + B_n U_n'(t) + C_n U_n(t) = 0,$$

where

$$B_n = \begin{bmatrix} b_0 - b_2 & b_1 - b_3 & b_2 - b_4 & \cdots & b_{n-1} - b_{n+1} \\ b_1 - b_3 & b_0 - b_4 & b_1 - b_5 & \cdots & b_{n-2} - b_{n+2} \\ b_2 - b_4 & b_1 - b_5 & b_0 - b_6 & \cdots & b_{n-3} - b_{n+3} \\ & & \vdots & & \\ b_{n-1} - b_{n+1} & b_{n-2} - b_{n+2} & b_{n-3} - b_{n+3} & \cdots & b_0 - b_{2n} \end{bmatrix} \tag{3.1}$$

and $C_n = \operatorname{diag}\{1, 4, \ldots, n^2\}$.

This is a known result, but no specific reference to it could be found in the literature.

The finite-dimensional problem obtained can now be written in the same manner as before, in terms of the optimization of the spectral abscissa of a quadratic pencil of matrices. To this end, define

$$L_n(\lambda) = \lambda^2 I + \lambda B + C$$

with B and C as in the previous proposition (in what follows we shall drop the indexes on the matrices) and let Σ_L^n denote its spectrum. The matrix B now belongs to a linear subspace of the space of symmetric $n \times n$ matrices. We shall denote this space by $\mathcal{B}_n$ and consider the function $\Omega_n : \mathcal{B}_n \to \mathbb{R}$ by

$$\Omega_n(B) = \min_{\lambda \in \Sigma_L^n} [-\operatorname{Re}(\lambda)] .$$

For each matrix B in $\mathcal{B}_n$, this function gives the corresponding spectral abscissa—note that C is fixed—and minimizing the rate at which solutions decay to zero then becomes equivalent to determining

$$\omega_o^n = \sup_{B \in \mathcal{B}_n} [\Omega_n(B)] .$$

4. Conditions on the coefficients b_j. We now present some results that emphasize the importance of the coefficients in the Fourier cosine series for b. The idea is that just like the average of b controls the real part of the high frequencies, the remaining coefficients in the Fourier cosine series play an important role in how large the spectral abscissa can be.

LEMMA 4.1. *If* $\Omega(b) > \alpha_0$ *for some* α_0, *then* $\langle T(-\alpha)u, u\rangle > 0$ *for all* $u \in D(T)$ *and all* $\alpha \leq \alpha_0$.

Proof. If $T(-\alpha)$ were not a positive operator, its smallest eigenvalue would clearly be nonpositive. This implies that the first eigencurve of $T(\lambda)$, that is, the curve described by the first eigenvalue of $T(\lambda)$ as λ varies over $\mathbb{R}$, would take a nonpositive value for λ equal to $-\alpha$. As this eigencurve is continuous and will be above the horizontal axis for large enough (real) λ, there would have to exist a number λ_0 larger than or equal to $-\alpha$ such that $T(\lambda_0)$ would have a zero eigenvalue. In other words, Σ_T would contain a point λ_0 satisfying $\lambda_0 \geq -\alpha \geq -\alpha_0$ and then $\Omega(b) \leq \alpha_0$. □

Using this result, we can now prove a relation between the first coefficients in the series for b and the maximum possible value for the spectral abscissa.

THEOREM 4.2. *If* $\Omega(b)$ *is larger than an integer* k, *then* $|b_0 - b_{2j}| < 2j$ *for* $j = 1, \ldots, k$.

Proof. If $\Omega(b) > k$, then the previous lemma gives that $\langle T(-\alpha)u, u\rangle > 0$ for all α smaller than or equal to k and all u in $\mathcal{H}$. In particular, for $u = \sin(jx)$ this yields that

$$\alpha^2 - (b_0 - b_{2j})\alpha + j^2 > 0,\ \alpha \leq k.$$

If $|b_0 - b_{2j}| \geq 2j$ for some j, then the polynomial above has a real root at

$$\alpha^- = \frac{b_0 - b_{2j} - \sqrt{(b_0 - b_{2j})^2 - 4j^2}}{2}.$$

If $b_0 - b_{2j} > 2j$, then the maximum possible value for α^- is j, and is attained when $b_0 - b_{2j} = 2j$. By the previous lemma this polynomial has to be positive for $\alpha \leq k$, and thus it follows that $b_0 - b_{2j}$ must be smaller than $2j$ for $j = 1, \ldots, k$. If $b_0 - b_{2j} < -2j$, then α^- will be negative and a similar argument applies. □

This result also highlights the importance of the coefficients corresponding to the even (around $\pi/2$) functions in the series. However, by itself it is not enough to show that the coefficients must be increased in order to increase the value of the spectral abscissa. In order to show that, we need the fact that the real part of the high frequencies clusters around minus the average of the function b, which, in this case, corresponds to $b_0/2$.

COROLLARY 4.3. *If b is in $BV(0, \pi)$ and $\Omega(b)$ is larger than a real number α, then b_0 is larger than 2α and the coefficients b_{2j} satisfy*

$$b_0 - 2j < b_{2j} < b_0 + 2j$$

for $j = 1, \ldots, k$, where k is the largest integer not exceeding α.

Proof. From [3, 8] we know that the real parts of the eigenvalues converge to $-b_0/2$ as the imaginary part of the eigenvalues grows to infinity. Thus if $\Omega(b)$ is larger than α, we must have b_0 larger than 2α. Since, by the previous theorem, $|b_0 - b_{2j}|$ is smaller than $2j$ for $j = 1, \ldots, k$, it follows that b_{2j} must satisfy the inequalities. □

We shall now consider the simple but important implications of this result. The eigenvalues of the operator $C = -\partial^2/\partial x^2$ on the interval $(0, \pi)$ and with homogeneous Dirichlet boundary conditions are $\gamma_j = j^2$, $j = 1, 2, \ldots$. When $b = \gamma_1^{1/2} = 1$, all eigenvalues have the same real parts which are equal to -1. If a damping term b exists for which $\Omega(b)$ is larger than $\gamma_1^{1/2}$, then b_0 must be larger than $2\gamma_1^{1/2}$ and it follows that b_2 must satisfy

$$0 < b_0 - 2 < b_2 < b_0 + 2.$$

In other words, if a damping term has a coefficient in $\cos(2x)$ which is smaller than or equal to zero, then the corresponding spectral abscissa has to be smaller than or equal to $\gamma_1^{1/2}$. This can be stated as follows.

COROLLARY 4.4. *If b is such that*

$$b_2 = \frac{2}{\pi} \int_0^\pi b(x) \cos(2x) dx \le 0,$$

then $\Omega(b)$ is smaller than or equal to 1.

In particular, if a function has a zero component in $\cos(2x)$, then Ω can be at most 1. Clearly similar results can be stated for the other Fourier coefficients. However, it should be kept in mind that the interest of this type of result for j larger than one will depend on whether Ω is unbounded or not.

5. Brief description of the algorithm. Genetic algorithms are a relatively new tool in optimization problems, with their first systematic presentation probably dating only as far back as 1975 with Holland's book [10]. Because of this, and also because of their nature, there are several possible choices to be made when using an algorithm of this type. Apart from the standard options, such as whether to use floating-point or binary representations, fixed of varying population sizes, etc., there are many others which normally depend on the problem in question. In this section we describe some of the choices which were made in this case, and also some of the main points related to the tuning of the algorithm.

The algorithm acts on a population with a fixed number p of vectors with m components, $\beta_j = (\beta_{j1}, \ldots, \beta_{jm})$, $j = 1, \ldots, p$, representing the Fourier cosine coefficients of the function b up to order m. In this respect, we have opted for a floating-point version of the algorithm instead of using a binary representation. These vectors are generated randomly at the beginning of each run, being weighted in order to impose a decay as k increases. In the cases presented in section 6, some of the components of each randomly generated vector were forced to be zero, while the nonzero components β_{jk} were uniformly distributed on intervals of the form $[-rs/(k+1), r(1-s)/(k+1)]$, where typically r and s where chosen to be, respectively, 10 and .3. This corresponds to considering functions whose Fourier coefficients roughly decay with $1/k$.

At each iteration the population is evaluated and a number n_r of vectors—those whose corresponding value of Ω_n is the smallest—are replaced by linear combinations of pairs of the members of the population—called arithmetical crossover. Once a vector has been in existence for a certain number of iterations, it will also be replaced, either by another linear combination or by a new randomly generated vector.

The main steps in each iteration can thus be described roughly as follows:

1. evaluation of the fitness of each vector over the total population,

2. generation of new vectors by means of arithmetical crossover to replace those with poorest performance in step 1,

3. replacement of the *older* vectors either by arithmetical crossover or by the introduction of new randomly generated vectors.

In the actual simulations, the size of the population p was taken to be 20 and that of n_r to be nine. The relation between these parameters turned out to be critical for the performance of the algorithm, in the sense that both smaller and larger values of n_r made the number of iterations necessary to attain a certain value much larger.

6. Numerical results. In this section we present a series of results obtained by running the algorithm for different choices of the set of nonzero Fourier cosine coefficients. The results obtained for eight trials are shown in Table 1. The best values for Ω obtained during each trial are given, together with the nonzero coefficients of the corresponding damping term. In general, the larger the number of nonzero coefficients used, the larger the corresponding value of Ω obtained. To keep numerical accuracy, the order of the coefficient corresponding to the highest frequency used was always kept much smaller than the order of the truncation. Typically, the value of n used varied between 15 and 30 during the runs, while the value of the best spectral abscissa obtained was then checked for n equal to 100. One of the features of the spectral method employed is that, with the exception of a few (one to four pairs in the cases tried) spurious eigenvalues in the highest frequencies, which, in any case, never deviated much from their real value, all the other eigenvalues are practically coincident for different values of n. This, together with the asymptotic behavior of the eigenvalues, ensured that the value obtained for the spectral abscissa remained constant when n was increased which makes it possible to compare different damping terms for relatively low values of n.

The first example shows that a damping term of the form $b_0/2 + b_2\cos(2x)$ is already sufficient to obtain a spectral abscissa greater than the best value obtained for the case of constant damping ($\Omega(1) = 1$). This case corresponds to a Mathieu equation with homogeneous Dirichlet boundary conditions and with complex coefficients which are not independent:

$$\begin{aligned}&u_{xx} + [w - z\cos(2x)]\,u = 0,\\ &w = -\lambda(b_0 + \lambda),\\ &z = 2b_2\lambda.\end{aligned}$$

TABLE 1
Values of Ω for different damping terms.

	i	ii	iii	iv	v	vi	vii	viii
b_0	3.1133	3.0083	3.4063	3.5366	3.5755	3.5191	3.5483	3.5692
b_1		0.2047						
b_2	1.4896	1.3644	1.8130	2.0046	2.0528	1.9950	2.0312	2.047
b_4			0.3930	0.5013	0.5450	0.4934	0.5314	0.5626
b_6				0.1843	0.2420	0.1654	0.2057	0.2050
b_8					-0.0365	-0.0571	-0.0897	-0.0513
b_{10}						-0.0939	-0.0645	-0.1105
b_{12}							-0.1343	-0.0316
b_{14}							-0.0420	-0.0783
b_{16}							0.0082	-0.0488
b_{18}							0.1382	0.0406
b_{20}								0.0916
b_{22}								0.2101
b_{24}								-0.0433
b_{26}								0.0553
b_{28}								0.0417
b_{30}								0.2643
b_{32}								-0.1077
b_{34}								0.1589
b_{36}								0.1110
b_{38}								0.2038
Ω	1.2869	1.3146	1.3984	1.5055	1.5353	1.5358	1.5675	1.5712

In the second example, we introduced an odd term, $b_1 \cos(x)$, which allowed for an improvement. However, in general, the introduction of the odd terms, although improving the value of Ω, did not do so in a significant manner when compared to the introduction of higher-order even terms. The other examples in Table 1 are those obtained by considering even higher-order coefficients in the series.

In Figures 6.1 and 6.2, we have plotted the different spectra corresponding to these damping terms. We considered n to be 100 in order to evaluate the spectral abscissa in each case, but plotted only 120 eigenvalues. Of the remaining 80, with the exception of about 8 (spurious) eigenvalues, they are all situated near the asymptotic line for the real part. In any case, the spurious eigenvalues are also quite close and do not affect the value of Ω.

The effect of introducing more terms in the series is quite visible in that this will increase the number of low-frequency eigenvalues which get further and further away from the vertical line $\mathrm{Re}(\lambda) = -b_0/2$.

In Figure 6.3 we show the plots of some of the graphs corresponding to these damping terms.

7. Discussion. By writing the damping term in the wave equation as a Fourier cosine series, we have been able to relate these coefficients to the value of the corresponding spectral abscissa. Based on this, and by successively adding more terms to the series, it has been possible to construct damping terms which give a better value of the spectral abscissa than the constant damping. Due to the fact that the real parts of the high frequencies are asymptotically close to $-b_0/2$, and that the numerical results give only a very small number of eigenvalues which are away from this vertical line and which converge very fast, we believe these results to be quite reliable.

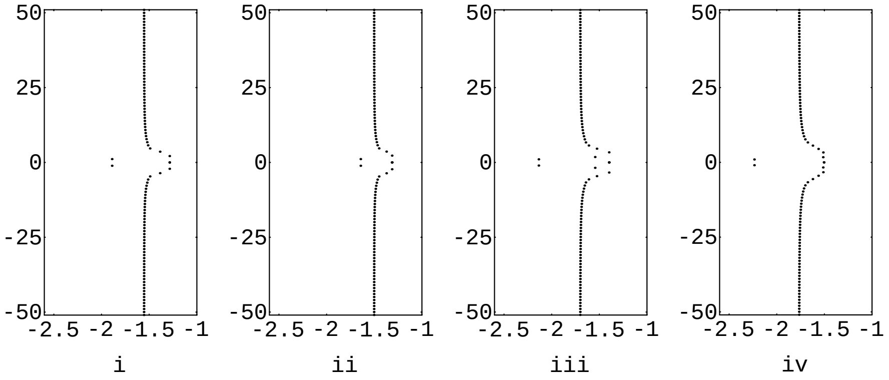


FIG. 6.1. *Spectra for cases* i–iv.

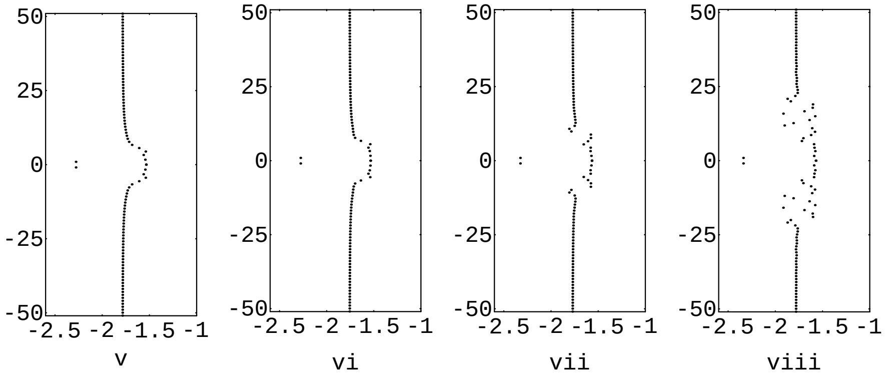


FIG. 6.2. *Spectra for cases* v–viii.

To prove in a rigorous way that this is in fact the case, one would have to proceed in two steps. First consider a ball B containing a number $2p$ of eigenvalues in such a way that the estimates for the eigenvalues outside B imply that their real parts are sufficiently close to $-b_0/2$. Then show that the eigenvalues given by the spectral approximation method considered converge to the eigenvalues of the original problem and obtain estimates for this convergence. After determining an order of the approximation which is sufficient to ensure that the approximate eigenvalues are close enough to the eigenvalues inside B, the desired result could be obtained. The second step indicated is fairly standard and methods for this type of problem can be found, for instance, in [11] (see also [13]). In particular, it is possible to obtain that $\Omega_n(B)$ converges to $\Omega(b)$. However, the estimates given in [8] for a general damping term and which are needed in order to determine the order p above which the eigenvalues outside B are close enough to the vertical line $\mathrm{Re}(\lambda) = -b_0/2$ are, from this point of view, quite poor. This makes the dimensions of the matrices that would have to be considered too high, and so either better estimates can be obtained for these particular cases or another line of approach has to be considered.

From a certain point onward it can be seen that the introduction of more terms in the Fourier series produces only a negligible increase in the best value of Ω obtained. There are several distinct reasons why this might be so. In the first place, it could

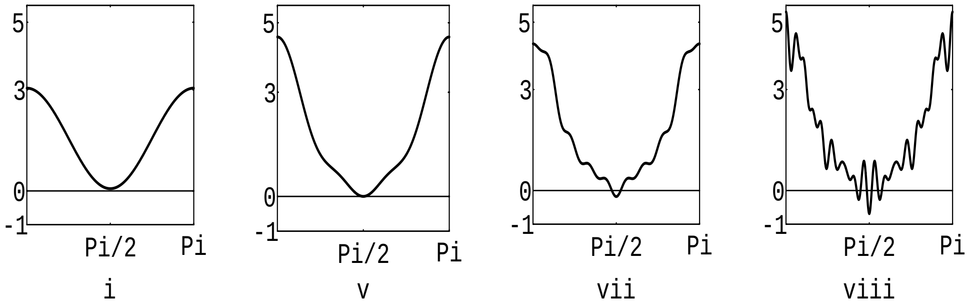


FIG. 6.3. *Graphs of* b *for cases* i, v, vii, *and* viii.

happen that the function Ω is in fact bounded from above, and that the values obtained are actually quite close to the optimum. Second, it might be possible that this algorithm ceases to be adequate at some stage due to a lack of capacity to negotiate its way past critical points other than the optimum; note that in all the runs the vector found always gave rise to a pair of eigenvalues which were very close to a double real eigenvalue with geometric multiplicity one. Finally, and related to this, it may also happen that the algorithm will have problems in getting the coefficients to lie in the desired intervals given by the results in section 4.

Regarding the shape of the damping terms given in section 6, there are several points worth noticing. To begin with, in all examples the way used to improve the spectral abscissa is by increasing the damping near the boundary and decreasing it in the middle of the interval. This becomes more noticeable as the number of nonzero terms grows. It also has the effect of making the minimum smaller and smaller, until it actually becomes negative. In the case of examples vii and viii, where, respectively, even terms up to b_{18} and b_{38} were included, and which gave a better value than those of the trials with less nonzero terms ($\Omega(b) = 1.5675$ and $\Omega(b) = 1.5712$), the corresponding damping terms obtained take negative values (see Figure 6.3). A similar phenomenon is already present in the finite-dimensional case treated in [7], where for some given matrices C the solution to the optimization problem is obtained for damping matrices which are indefinite (and in some cases only for these). From the results in [6], we know that if b is a damping term which changes sign, then the trivial solution of the equation

$$u_{tt} + 2pb(x)u_t = \Delta u,$$

where p is a real parameter, will become unstable if p is made large enough. This means that for this specific damping term, increasing the parameter p will first have the effect of moving all of the spectrum to the left side past the vertical line $\mathrm{Re}(\lambda) = -1.5$ for p close to one and then bringing some of the eigenvalues to the right-hand side of the complex plane for p large enough.

As is mentioned in the introduction, the main idea behind the numerical results presented here is to give a counterexample to the constant damping conjecture. In this sense, these should only be seen as a beginning to a more thorough study of the problem, both from the analytical and numerical points of view, but now from the perspective that the optimum is in fact not attained at the constant damping. In some sense, this is a characteristic use of genetic algorithms. As has been pointed out in [10], for instance, such algorithms can be quite useful in performing an initial search, selecting a subset of the whole possible space which is likely to contain the optimum. After that, the optimization process should be handed over to a procedure that takes advantage of the specific structure of the problem.

REFERENCES

[1] S. J. Cox, *Designing for optimal energy absorption* III*: Numerical minimization of the spectral abscissa*, Structural Optimization, 13 (1997), pp. 17–22.

[2] S. J. Cox and M. L. Overton, *Perturbing the critically damped wave equation*, SIAM J. Appl. Math., 56 (1996), pp. 1353–1362.

[3] S. Cox and E. Zuazua, *The rate at which energy decays in a damped string*, Comm. Partial Differential Equations, 19 (1994), pp. 213–243.

[4] P. Freitas, *The linear damped wave equation, Hamiltonian symmetry, and the importance of being odd*, Discrete Contin. Dyn. Syst., 4 (1998), pp. 635–640.

[5] P. Freitas, *Quadratic matrix polynomials with Hamiltonian spectrum and oscillatory damped systems*, Z. Angew. Math. Phys., 50 (1999), pp. 1–18.

[6] P. Freitas, *On some eigenvalue problems related to the wave equation with indefinite damping*, J. Differential Equations, 127 (1996), pp. 320–335.

[7] P. Freitas and P. Lancaster, *On the Optimal Value of the Spectral Abscissa for a System of Linear Oscillators*, I.S.T., Lisbon, preprint.

[8] P. Freitas and E. Zuazua, *Stability results for the wave equation with indefinite damping*, J. Differential Equations, 132 (1996), pp. 338–352.

[9] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison–Wesley, Reading, MA, 1989.

[10] J. H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI, 1975.

[11] M. A. Krasnosel'skii, G. M Vainikko, P. P. Zabreiko, Ya. B. Rutitskii, and V. Ya. Stetsenko, *Approximate Solution of Operator Equations*, Wolters–Noordhoff Publishing, Groningen, 1972.

[12] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, 3rd ed., Springer, Berlin, New York, 1996.

[13] J. E. Osborn, *Spectral approximation for compact operators*, Math. Comp., 29 (1975), pp. 712–725.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 2, pp. 388–408

A BEHAVIORAL APPROACH TO CONTROL OF DISTRIBUTED SYSTEMS*

HARISH K. PILLAI† AND SHIVA SHANKAR†

Abstract. This paper develops a theory of control for distributed systems (i.e., those defined by systems of constant coefficient partial differential operators) via the behavioral approach of Willems. The study here is algebraic in the sense that it relates behaviors of distributed systems to submodules of free modules over the polynomial ring in several indeterminates. As in the lumped case, behaviors of distributed ARMA systems can be reduced to AR behaviors. This paper first studies the notion of AR controllable distributed systems following the corresponding definition for lumped systems due to Willems. It shows that, as in the lumped case, the class of controllable AR systems is precisely the class of MA systems. It then shows that controllable 2-D distributed systems are necessarily given by free submodules, whereas this is not the case for n-D distributed systems, $n \geq 3$. This therefore points out an important difference between these two cases. This paper then defines two notions of autonomous distributed systems which mimic different properties of lumped autonomous systems.

Control is the process of restricting a behavior to a specific desirable autonomous subbehavior. A notion of stability generalizing bounded input–bounded output stability of lumped systems is proposed and the pole placement problem is defined for distributed systems. This paper then solves this problem for a class of distributed behaviors.

Key words. distributed systems, systems of partial differential equations, controllability, stability, pole placement

AMS subject classifications. 93C20, 93C35, 35B37, 35E20

PII. S0363012997321784

1. Introduction.

In this paper we develop a theory of control of distributed systems patterned after the behavioral approach for lumped systems in Willems [10, 11]. Thus we study the control of behaviors of systems of linear, constant coefficient partial differential operators. In this paper, we demonstrate that, while behaviors of distributed systems are similar to the behaviors of lumped systems in some respects, there are nonetheless many points of departure between the two cases, especially in the techniques employed to arrive at the results. This is essentially due to the fact that lumped systems are defined over a principal ideal domain (PID), whereas distributed systems are not.

In [10], Willems initiates his approach to the study of systems by first considering ARMA systems. (We adopt the terminology there to define various systems like ARMA, AR, and MA systems, which are again formally defined in section 2 below.) He establishes an "elimination theorem" for ARMA systems; i.e., he proves that every ARMA system is equivalent to an AR system. This follows from the fact that every submodule of a free module over the principal ideal domain $\mathbb{R}[\frac{d}{dx}]$ is free. On the other hand, the elimination theorem for distributed systems requires the celebrated Ehrenpreis–Palamodov theorem (see Oberst [5, Corollary 38] for a constructive proof).

Our study of distributed systems is algebraic in the sense that we set up a correspondence between smooth behaviors and submodules of free modules over polynomial rings (in several indeterminates). That this correspondence is one to one is the content of a hard theorem of Oberst [5] and is in fact a central result of his seminal paper.

*Received by the editors May 21, 1997; accepted for publication (in revised form) March 26, 1998; published electronically November 23, 1998.

http://www.siam.org/journals/sicon/37-2/32178.html

†Department of Electrical Engineering, Indian Institute of Technology, Powai, Bombay 400076, India (sshankar@ee.iitb.ernet.in).

The main body of this paper starts with a study of controllable systems. Our definition of controllability is patterned after Willems's definition for 1-D systems in [10, 9, 11] and that for discrete 2-D systems in [6]. (Following the spirit of Willems's work, our notion of controllability does not rely on any input-output structure.) We show that every MA system is controllable. We obtain a necessary and sufficient condition for an AR system to admit an MA representation. This condition allows us to conclude that every controllable AR system is MA. *Thus the class of distributed controllable AR systems is precisely the class of MA systems.* We also show that a submodule corresponding to such a system is maximal amongst the class of submodules with the same compactly supported behavior.

While every lumped behavior is of course given by a free submodule (as observed above), Rocha and Willems [6] show that every *controllable* 2-D discrete system is also given by a free submodule, which here corresponds to a system given by a left prime matrix. This is an important result, for free submodules over polynomial rings occupy a central position and geometrically correspond, by Serre, Quillen, and Suslin, to the class of vector bundles over affine space. In what we consider an important counterexample, we show that controllable AR distributed n-D systems for $n \geq 3$ are *not necessarily given by free submodules.* The same example also shows that systems described by left prime matrices need not be controllable. That such an example exists, we explain, is due to the fact that the global dimension of the ring $\mathbb{R}[\partial_1, \partial_2, \partial_3]$ over which 3-D systems are defined is 3. We show that such examples cannot occur if the global dimension of the ring over which the system is defined is less than or equal to 2. This therefore extends the result of Rocha and Willems on 2-D discrete systems alluded to earlier to 2-D distributed systems as well. These results follow from a necessary and sufficient condition which describes when a controllable system is given by a free submodule.

Given a behavior $\mathbf{B}$, corresponding to a submodule, say $\mathbf{R}$ of $(\mathbb{R}[\partial_1, \ldots, \partial_n])^k$, we wish to study subbehaviors of $\mathbf{B}$. These subbehaviors come from submodules of $(\mathbb{R}[\partial_1, \partial_2, \ldots, \partial_n])^k$ containing $\mathbf{R}$. By extending some set of generators of $\mathbf{R}$ to a set of generators of a submodule of $(\mathbb{R}[\partial_1, \partial_2, \ldots, \partial_n])^k$ containing $\mathbf{R}$, it is possible to restrict $\mathbf{B}$ to any subbehavior contained in it. Of all the subbehaviors contained in $\mathbf{B}$, we are interested primarily in a special class of behaviors which are analogous to autonomous behaviors of lumped systems. Restriction to such behaviors is the process of control. The above description explains the process of control for AR systems. We also translate this procedure to the case of MA systems.

Recollect from Willems [10, 9, 11] that an autonomous lumped behavior is one given by a submodule of $(\mathbb{R}[\frac{d}{dx}])^k$ of rank k. Such a behavior is a finite-dimensional $\mathbb{R}$-vector space. In the case of distributed systems, as shown in this paper, these two properties of a behavior need not coincide. This is essentially due to the fact that a subvariety of $\mathbb{C}$ (if not all of $\mathbb{C}$) is a finite set of points, whereas this is certainly not the case in $\mathbb{C}^n$, $n \geq 2$. As a result we introduce two kinds of autonomous behaviors to capture the above properties of a lumped autonomous behavior. By the first notion, the behavior of a single partial differential operator is autonomous, whereas the second notion of autonomy (namely, what we call a strongly autonomous system) implies a finite-dimensional behavior. Thus a strongly autonomous behavior resembles the behavior of a lumped autonomous system.

We generalize the notion of bounded input–bounded output stability of lumped behaviors by replacing the positive time axis by a (proper) cone S in $\mathbb{R}^n$ whose apex is the origin. A behavior is considered stable if the elements in it tend to zero along all directions in S. We show that a stable behavior must necessarily be autonomous. A stabilizing controller is one which restricts the behavior to a (nonzero) subbehavior stable with respect to S. We show that a controllable behavior can be stabilized with respect to any cone S. We also define the more general pole placement problem for AR systems. We solve this problem for a subclass of AR systems.

Our paper also contains what we feel are illuminating examples which point to the complexities of the behaviors of distributed systems. Moreover, since our approach to the study of distributed systems is algebraic, it makes available computational techniques from commutative algebra (Gröbner basis, etc.), although we do not directly address such issues here.

2. Preliminaries. In the notation of [10, 11], we consider systems of the form $(\mathbb{R}^n, \mathbb{R}^k, \mathbf{B})$, where $\mathbf{B}$ is a subspace of $(\mathcal{D}')^k$, the space of $\mathbb{R}^k$-valued distributions on $\mathbb{R}^n$. (In its stead we sometimes consider subspaces of $(\mathcal{C}^\infty)^k$, $\mathcal{D}^k$, etc.) These subspaces $\mathbf{B}$ are the behaviors of distributed ARMA systems which we now define.

Let

$$R : (\mathcal{D}')^k \rightarrow (\mathcal{D}')^l \tag{1}$$

and

$$M : (\mathcal{D}')^m \rightarrow (\mathcal{D}')^l \tag{2}$$

be *continuous local* $\mathcal{E}'$-module morphisms, where $\mathcal{E}'$ is the space of compactly supported distributions on $\mathbb{R}^n$. Note that $(\mathcal{D}')^p$ is an $\mathcal{E}'$-module for all p, where the action of $\mathcal{E}'$ on $(\mathcal{D}')^p$ is given by componentwise convolution, i.e.,

$$f * (v_1, \ldots, v_p) = (f * v_1, \ldots, f * v_p)$$

for f in $\mathcal{E}'$ and $(v_1, \ldots, v_p)$ in $(\mathcal{D}')^p$. We therefore require that $R(f * (u_1, \ldots, u_k)) = f * R(u_1, \ldots, u_k)$, and similarly for M. (In this paper we follow the notation in Hörmander [2], except that we use $\mathcal{D}$ to denote the space of compactly supported (complex valued) smooth functions instead of $\mathcal{C}_0^\infty$.)

It follows then that the kernel of R as well as the image of M are $\mathcal{E}'$-submodules of $(\mathcal{D}')^k$ and $(\mathcal{D}')^l$, respectively. Hence both the kernel of R as well as the image of M are subspaces that are shift invariant, i.e., closed under translations (take f above to be δ_x, the Dirac measure supported at x in $\mathbb{R}^n$). By local we mean that the support of $R(v)$ is contained in the support of v (and similarly for M), where for $v = (v_1, \ldots, v_p)$ in $(\mathcal{D}')^p$, the support of v is the union of the supports of v_i, $i = 1, \ldots, p$.

The behavior $\mathbf{B}$ of the ARMA system (R, M) defined by the maps (1) and (2) is the subspace of $(\mathcal{D}')^k$ consisting of those distributions that are mapped by R into the image of M, that is,

$$\mathbf{B} = \{u \in (\mathcal{D}')^k \mid Ru \in Im(M)\}.$$

As R is local, we have the restriction map

$$R_1 : (\mathcal{E}')^k \rightarrow (\mathcal{E}')^l$$

which is also continuous. We write a matrix for R_1, say (r_{ij}), with respect to the standard bases $e_1, \ldots, e_k$ and $f_1, \ldots, f_l$ on $(\mathcal{E}')^k$ and $(\mathcal{E}')^l$, respectively, where $e_i = (0, \ldots, \delta, 0, \ldots, 0)$ (δ in the ith slot) and similarly for the f_j's. As R_1 is local, the $R_1 e_i$'s are all supported at the origin, and hence each r_{ij} is also supported at the origin. Thus each r_{ij} is an $\mathbb{R}$-linear combination of δ and its derivatives (see Hörmander [2, Theorem 2.3.4]). We can therefore consider each r_{ij} as the corresponding constant coefficient partial differential operator. Then the action of R_1 on $(\mathcal{E}')^k$ is given by

$$R_1(u_1, \ldots, u_k) \;=\; \left(\sum_{i=1}^{k} r_{1i}(u_i), \ldots, \sum_{i=1}^{k} r_{li}(u_i) \right).$$

By density of $(\mathcal{E}')^k$ in $(\mathcal{D}')^k$, R is the unique extension of R_1 and thus admits the same matrix representation (r_{ij}). This same matrix also represents the restriction of R to the submodules $(\mathcal{C}^\infty)^k$, $\mathcal{D}^k$, etc. These considerations also hold for the map M yielding a matrix representation (m_{ij}) for it. We therefore, in this paper, study behaviors of distributed ARMA systems, that is, behaviors defined by *systems of constant coefficient partial differential operators.*

It is also of interest to consider, as in [10, 11], special cases of ARMA systems, namely MA and AR systems. MA systems are those systems for which $k = l$ and R is the identity morphism in (1). Thus an MA behavior is the image of an $\mathcal{E}'$-module morphism M. On the other hand, those behaviors that are kernels of $\mathcal{E}'$-module morphisms are called AR behaviors and correspond to the case when M in (2) is the zero map. In [10], Willems proves an elimination theorem for lumped systems that states that the $\mathcal{C}^\infty$ behavior of an ARMA system is equivalent to the $\mathcal{C}^\infty$ behavior of some AR system. An elimination theorem for distributed systems follows from the Ehrenpreis–Palamodov theorem, which we also use elsewhere in this paper. We now state this theorem in the notation that we employ.

Let (R, M) be an ARMA system and let (r_{ij}) and (m_{ij}) be the matrix representations for R and M, respectively, described above. Let $\mathcal{A}$ denote the commutative ring $\mathbb{R}[\partial_1, \ldots, \partial_n]$, that is, the polynomial ring in the variables $\partial_1, \ldots, \partial_n$, where ∂_i is the partial derivative in the ith direction x_i. Thus the entries of the above matrices are from the ring $\mathcal{A}$. Consider the set $\mathbf{Q}$ of all relations of the rows of (m_{ij}), that is, the set of all l-tuples $(q_1, \ldots, q_l)$ in $\mathcal{A}^l$ such that

$$\sum_{i=1}^{l} q_i m_{ij} \;=\; 0, \qquad j = 1, \ldots, m. \tag{3}$$

This set $\mathbf{Q}$ is clearly a submodule of $\mathcal{A}^l$. We now quote the theorem of Ehrenpreis–Palamodov from Hörmander [3, Theorem 7.6.13].

THEOREM (Ehrenpreis–Palamodov). *Let f be in $(\mathcal{C}^\infty)^l$. Then there exists a u in $(\mathcal{C}^\infty)^m$ such that $M(u) = f$ iff $q(f) = 0$ for all q in $\mathbf{Q}$, the module of relations of the rows of M, i.e., the rows of (m_{ij}).*

Remark. The Ehrenpreis–Palamodov theorem is valid in many other distribution spaces; see chapter 15 of Hörmander [2]. That it is also valid for $\mathcal{D}'$ is a consequence of a result of Oberst [5] which states that $\mathcal{D}'$ is an injective cogenerator.

COROLLARY 1 (the elimination theorem). *The $\mathcal{C}^\infty$ (or $\mathcal{D}'$) behavior of an ARMA system is the same as the $\mathcal{C}^\infty$ (or $\mathcal{D}'$) behavior of some AR system.*

This elimination theorem appears in Oberst [5], where there is in fact a constructive proof for it. Thus it suffices to study behaviors of AR systems.

Consider now the AR system defined by the map R in (1) represented by the matrix (r_{ij}). Consider each row of R as an element of the free $\mathcal{A}$-module $\mathcal{A}^k$. In what follows we consider each element $p = (p_1, \ldots, p_k)$ in $\mathcal{A}^k$ as an $\mathcal{E}'$-module morphism

$$p \;:\; (\mathcal{D}')^k \to \mathcal{D}',$$

$$(u_1, \dots, u_k) \to \sum_i p_i(u_i).$$

We also consider restrictions of the above morphism

$$p \ : \ (W)^k \to W,$$

where W is any $\mathcal{E}'$-submodule of $\mathcal{D}'$, for instance $\mathcal{C}^\infty$, $\mathcal{D}$, or $\mathcal{E}'$ itself. Correspondingly, we obtain the restrictions of R to the submodules $(W)^k$. Let $u = (u_1, \dots, u_k)$ be in the behavior $\mathbf{B}$ which is the kernel of R or, more generally, the kernel of its restriction to $(W)^k$. It then follows that this u is in the kernel of every p in the submodule $\mathbf{R}$ of $\mathcal{A}^k$ generated by the rows of R. Thus associated with a behavior in $(W)^k$ is a submodule of $\mathcal{A}^k$. Conversely, given a submodule $\mathbf{R}$ of $\mathcal{A}^k$, we obtain a behavior consisting of those u in $(W)^k$ that lie in the kernel of every p in $\mathbf{R}$. As $\mathcal{A}^k$ is Noetherian, $\mathbf{R}$ is finitely generated, say by $p_1, \dots, p_l$. The behavior corresponding to $\mathbf{R}$ can then be thought of as the kernel of a map $R : (W)^k \to (W)^l$ given by a matrix whose rows are $p_1, \dots, p_l$. Thus given an $\mathcal{E}'$-submodule W of $\mathcal{D}'$, this establishes a correspondence between behaviors in W^k and submodules of $\mathcal{A}^k$.

The principal idea in this paper is to exploit this correspondence to deduce properties of behaviors in terms of submodules of $\mathcal{A}^k$. This makes our study of behaviors of distributed systems tractable; note that $\mathcal{A}^k$ is a Noetherian $\mathcal{A}$-module whereas $(W)^k$ is not finitely generated as an $\mathcal{E}'$-module when W is $\mathcal{D}$, $\mathcal{D}'$, or $\mathcal{C}^\infty$.

We now elaborate on this correspondence between behaviors and submodules of $\mathcal{A}^k$. Thus given a behavior $\mathbf{B} \subset (W)^k$, denote by $\mathcal{M}(\mathbf{B})$ the submodule of $\mathcal{A}^k$ consisting of all the elements p which map to zero every element in $\mathbf{B}$, i.e., those p for which $p(u) = 0$ for all u in $\mathbf{B}$. Similarly, given a submodule $\mathbf{R}$ of $\mathcal{A}^k$, denote by $\mathcal{B}_W(\mathbf{R})$ the behavior in W corresponding to $\mathbf{R}$, i.e., those u in $(W)^k$ with $p(u) = 0$ for all p in $\mathbf{R}$. Clearly $\mathcal{B}_W(\mathbf{R})$ is an $\mathcal{E}'$-submodule of $(W)^k$ (note that $p(f * u) = f * p(u)$). Observe that a behavior $\mathbf{B}$ in W is the kernel of an $\mathcal{E}'$-module morphism R restricted to $(W)^k$ (equation (1)), and so in this notation, $\mathbf{B}$ is actually $\mathcal{B}_W(\mathbf{R})$, where $\mathbf{R}$ is the submodule generated by the rows of R.

LEMMA 1. *$\mathcal{B}_W \circ \mathcal{M}$ is the identity map on AR behaviors for any $\mathcal{E}'$-submodule W, i.e., $\mathcal{B}_W \circ \mathcal{M}(\mathbf{B}) = \mathbf{B}$.*

Proof. Clearly $\mathbf{B} \subset \mathcal{B}_W(\mathcal{M}(\mathbf{B}))$. As mentioned above, since $\mathbf{B}$ is an AR behavior, it is equal to $\mathcal{B}_W(\mathbf{R})$ for some submodule $\mathbf{R}$ of $\mathcal{A}^k$. Clearly, $\mathbf{R}$ is then contained in $\mathcal{M}(\mathbf{B}) = \mathcal{M}(\mathcal{B}_W(\mathbf{R}))$. Thus $\mathcal{B}_W(\mathcal{M}(\mathbf{B})) \subset \mathcal{B}_W(\mathbf{R}) = \mathbf{B}$. □

COROLLARY 2. *The correspondence $\mathbf{B} \to \mathcal{M}(\mathbf{B})$ is injective.*

Proof. Suppose $\mathcal{M}(\mathbf{B_1}) = \mathcal{M}(\mathbf{B_2})$. Then $\mathbf{B_1} = \mathcal{B}_W(\mathcal{M}(\mathbf{B_1})) = \mathcal{B}_W(\mathcal{M}(\mathbf{B_2})) = \mathbf{B_2}$. □

This raises the question of whether the correspondence $\mathbf{R} \to \mathcal{B}_W(\mathbf{R})$ is bijective, i.e., whether $\mathcal{B}_W$ and $\mathcal{M}$ are inverses of one another. The answer to this question of course depends upon W. Prompted by this question, we make the following definition.

DEFINITION 1. *A submodule $\mathbf{R}$ of $\mathcal{A}^k$ is called a Willems submodule with respect to W if $\mathcal{M}(\mathcal{B}_W(\mathbf{R})) = \mathbf{R}$.*

Thus the correspondence $\mathbf{R} \to \mathcal{B}_W(\mathbf{R})$ is bijective when restricted to the class of Willems submodules with respect to W. Given a submodule $\mathbf{R}$ of $\mathcal{A}^k$, $\mathcal{M}(\mathcal{B}_W(\mathbf{R}))$ is clearly the smallest Willems submodule that contains $\mathbf{R}$. We will call $\mathcal{M}(\mathcal{B}_W(\mathbf{R}))$ the Willems submodule with respect to W generated by $\mathbf{R}$. Clearly, $\mathcal{M}(\mathcal{B}_W(\mathbf{R}))$ is the largest submodule in $\mathcal{A}^k$ determining the same behavior in W^k as that of $\mathbf{R}$.

In terms of this definition, we now state the theorem of Oberst referred to in the introduction.

THEOREM (Oberst). *Every submodule of $\mathcal{A}^k$ is Willems with respect to $\mathcal{C}^\infty$ (or $\mathcal{D}'$).*

Later in this paper we determine *the class of submodules which are Willems with respect to $\mathcal{D}$ or $\mathcal{E}'$ to be precisely the class of MA systems.*

In this paper we often look at the subclass of AR systems given by MA systems which display several "nice" properties such as controllability and stabilizability. We therefore now discuss a parametrization of such systems.

PROPOSITION 1. *Every MA system is the image of a right prime morphism.*

Remark. By a right prime morphism, we mean a morphism given by a factor right prime matrix. For various notions of right and left primeness, see Bose [1].

Proof. Let an MA behavior be given as the image of a morphism whose matrix representation is given by M as in (2). We first show that this same MA behavior is also given by a submatrix M_0 of M which has full column rank.

Let the columns of M be given by elements $c_1, \dots, c_m$. Let the rank of M be m_0 and assume, without loss of generality, that $c_1, \dots, c_{m_0}$ are $\mathcal{A}$-independent. Let c_i, $i \neq 1, \dots, m_0$, be any other column of the matrix M. By assumption, there is a relation between the columns $c_1, \dots, c_{m_0}$ and c_i, say $ac_i = \sum_{j=1}^{m_0} \alpha_j c_j$, where a and the α_j's belong to $\mathcal{A}$. We will show that the image of c_i in $(\mathcal{C}^\infty)^l$ (or $(\mathcal{D}')^l$ as the case may be) is contained in the span of the images of $c_1, \dots, c_{m_0}$. For this consider any f in $\mathcal{C}^\infty$ (or $\mathcal{D}'$). By a standard result (see for instance Hörmander [2, Corollary 10.6.8]), there is a g in $\mathcal{C}^\infty$ (or $\mathcal{D}'$) such that $a(g) = f$. Then $c_i(f) = c_i(ag) = ac_i(g) = (\sum_{j=1}^{m_0} \alpha_j c_j)(g)$. This element is in the image of the submatrix M_0 given by the columns $c_1, \dots, c_{m_0}$.

Suppose now that M_0 is not right prime. Then $M_0 = M_1 T$, where M_1 is right prime and where T is a square $m_0 \times m_0$ matrix whose determinant is a nonconstant polynomial. Note that the module of relations of the rows of T is the 0 submodule of $\mathcal{A}^{m_0}$. Using the Ehrenpreis–Palamodov theorem or the injectivity result of Oberst in the remark following it, we conclude that T is surjective on $(\mathcal{C}^\infty)^{m_0}$ or $(\mathcal{D}')^{m_0}$, respectively.

Thus the image of M_0 equals the image of M_1, which is to say that the given MA behavior is also the image of a right prime morphism, namely M_1 above. □

In view of this proposition, we will henceforth describe MA behaviors using only right prime morphisms.

We conclude this section with the following definition and approximation theorem of Malgrange which we quote from Hörmander [2].

DEFINITION. *A solution f of the constant coefficient partial differential equation $p(f) = 0$ is called an exponential solution if it is of the form*

$$f(x) = q(x)e^{\langle x, \xi \rangle},$$

where $q(x)$ is a polynomial and ξ is in $\mathbb{C}^n$.

Remark. If $x = (x_1, \dots, x_n)$ and $\xi = (\xi_1, \dots, \xi_n)$, then by $\langle x, \xi \rangle$ we mean $\sum x_i \xi_i$. This is not the standard Hermitian inner product on $\mathbb{C}^n$. However, as in this paper x is always in $\mathbb{R}^n$, the above choice of the bilinear form $\langle , \rangle$ suffices.

It is easy to check that if $f = q(x)e^{\langle x, \xi \rangle}$ is a solution of $p(f) = 0$, then $p(\xi) = 0$, i.e., ξ lies in the variety of p (in $\mathbb{C}^n$).

THEOREM (see Theorem 7.3.6 in [2]). *The closed linear hull in $\mathcal{C}^\infty(\mathbb{R}^n)$ of the exponential solutions of the equation $p(f) = 0$ consists of all its solutions in $\mathcal{C}^\infty(\mathbb{R}^n)$.*

Remark (see remark following Theorem 10.5.1 in [2]). In fact Malgrange has proved that it is sufficient to use polynomial solutions in the above theorem iff every nonconstant factor of p vanishes at the origin. Also, it suffices to use solutions of the form $e^{\langle x,\xi \rangle}$ iff p has no multiple factors.

3. Controllable systems. We now discuss controllability properties of AR systems. Motivated by the definition of controllability of lumped systems in [10, 11] and of 2-D discrete systems in [6], we adopt a similar definition for distributed systems below.

DEFINITION 2. *Let* $(\mathbb{R}^n, \mathbb{R}^k, \mathbf{B})$ *be an AR system. Then* $\mathbf{B}$ *is said to be controllable if for* w_1 *and* w_2*, any two elements in* $\mathbf{B}$*, and for* U_1 *and* U_2 *any two open subsets of* $\mathbb{R}^n$ *such that their closures are disjoint (i.e.,* $\overline{U}_1 \cap \overline{U}_2 = \emptyset$*), there exists an element* w *in* $\mathbf{B}$ *which coincides with* w_1 *on* U_1 *and with* w_2 *on* U_2*.*

The above definition means that the action of w coincides with that of w_1 on test functions whose supports lie in U_1 and with the action of w_2 on test functions whose supports lie in U_2. If w_1 and w_2 are smooth functions, then the above implies that w coincides pointwise with w_1 on U_1 and with w_2 on U_2. Intuitively, w has patched up w_1 and w_2.

Note that our definition is symmetric with respect to all the variables $x_1, \dots, x_n$. This is motivated by the theory of 2-D systems applied to image processing where neither variable plays a special role. It is possible however that in some systems a variable plays a special role, for instance, that of time. In such systems it might be that it is necessary only to patch up elements in the behavior along this special variable. Then U_1 and U_2 in the above defintion need only be "strips," that is, open sets of the form $I \times \mathbb{R}^{n-1}$, where I is an open interval in $\mathbb{R}$. Note that such open sets are of course included in the definition above.

DEFINITION 3. *Let* U *be an open subset of* $\mathbb{R}^n$*, and let* V *be any closed subset whose interior contains the closure of* U*. Let* w *be an element in* $(\mathcal{D}')^k$*. An element* w' *in* $(\mathcal{D}')^k$ *is a cutoff of* w *with respect to* U *and* V *if* w' *coincides with* w *on* U *and with* 0 *on* V^C*, the complement of* V*.*

Recollect that given any U and V as above, there is a smooth function f which is identically 1 on U and 0 on V^C (i.e., a "bump" function). Thus in the above definition fw is such a w'. Note that if w is smooth, then such a cutoff is also smooth.

LEMMA 2. *Let* $\mathbf{B}$ *be an AR behavior in* $(W)^k$*, where* W *is any* $\mathcal{E}'$*-submodule of* $\mathcal{D}'$*. Then* $\mathbf{B}$ *is controllable iff for every* w *in* $\mathbf{B}$ *and for every* U *and* V *as in the definition above, some cutoff of* w *with respect to* U *and* V *is also in* $\mathbf{B}$*.*

Proof. Let $\mathbf{B}$ be controllable and let U and V be as in the definition above. Then U and V^C, the complement of V, are open sets whose closures are disjoint. Let w be in $\mathbf{B}$. As $\mathbf{B}$ is a linear subspace, the 0 distribution is in $\mathbf{B}$. By the definition of controllability, there is a w' in $\mathbf{B}$ which coincides with w on U and with 0 on V^C. This w' is a cutoff of w with respect to U and V.

Conversely, let w_1 and w_2 be any two elements in $\mathbf{B}$, and let U_1 and U_2 be open subsets whose closures are disjoint. As $\mathbb{R}^n$ is normal, we can find closed disjoint subsets V_1 and V_2 containing U_1 and U_2 in their interiors. There is a cutoff w_1' of w_1 with respect to U_1 and V_1 in $\mathbf{B}$ and, similarly, a cutoff w_2' of w_2 with respect to U_2 and V_2 in $\mathbf{B}$. By linearity of $\mathbf{B}$, $w = w_1' + w_2'$ is also in $\mathbf{B}$. This w obviously coincides with w_1 on U_1 and with w_2 on U_2. □

In [10, 11], Willems shows that all lumped controllable AR systems are MA systems and conversely. Rocha and Willems prove this same fact for 2-D discrete systems. They also show that controllable 2-D discrete systems are always given by free submodules or equivalently given by kernels of left prime matrices. The situation in the distributed case in three or more variables is more involved. We explain all this in terms of the global dimension of the ring $\mathcal{A}$.

In what follows we rely on the Ehrenpreis–Palamodov theorem (or its variant in the case of $\mathcal{D}'$ due to Oberst) in the proofs of many statements. As such a theorem is not valid for an arbitrary $\mathcal{E}'$-submodule W of $\mathcal{D}'$, *we henceforth restrict ourselves to behaviors with respect to* $\mathcal{C}^\infty$ *or* $\mathcal{D}'$.

We start this development with the following proposition.

PROPOSITION 2. *Every MA system is controllable.*

Proof. Let the MA system be given by the $\mathcal{E}'$-module morphism $M : (\mathcal{D}')^m \to (\mathcal{D}')^l$; i.e., let the behavior $\mathbf{B}$ be the image of M. Let w be any element in $\mathbf{B}$. Then $w = M(v)$ for some v in $(\mathcal{D}')^m$. Let U and V be as in Definition 3. Let v' be any cutoff of v with respect to U and V. Consider $w' = M(v')$. We claim that w' is a cutoff of w also with respect to U and V.

Let f be any test function in $(\mathcal{D})^l$ whose support is contained in U. Then $w'(f) = M(v')(f) = v'(M'(f))$, where $M' : (\mathcal{D})^l \to (\mathcal{D})^m$, the adjoint of M, is also a local continuous $\mathcal{E}'$-module morphism. Thus the support of $M'(f)$ is also contained in U. Then as v' is a cutoff of v with respect to U and V, $v'(M'(f)) = v(M'(f)) = M(v)(f) = w(f)$. Hence $w'(f) = w(f)$ for all f in $(\mathcal{D})^l$ with support in U. Similarly, w' coincides with 0 in V^C. Thus w' is indeed a cutoff of w with respect to U and V.

If we restrict the morphism M to the subspace $(\mathcal{C}^\infty)^m$, then there is a cutoff w' of w in $(\mathcal{C}^\infty)^l$ which is also smooth and is obtained by choosing a smooth cutoff v' of v. Thus MA systems are also controllable in the $\mathcal{C}^\infty$ sense. □

The question now arises as to which AR systems are controllable. We show below that the class of distributed controllable AR systems coincides with the class of MA systems. Toward this we first characterize those AR systems which are MA systems and hence controllable by the above proposition. For the sake of convenience, *we restrict ourselves further to the* $\mathcal{C}^\infty$ *category.* Note, however, that all the results that follow are equally valid for behaviors with respect to $\mathcal{D}'$.

Given the behavior $\mathbf{B}$ of an AR system $R : (\mathcal{C}^\infty)^k \to (\mathcal{C}^\infty)^l$, consider the subbehaviors of $\mathbf{B}$ which are MA systems, i.e., images of morphisms $M : (\mathcal{C}^\infty)^m \to (\mathcal{C}^\infty)^k$ that lie in the kernel of R. (Recollect from the discussion in section 2 that MA behaviors are in one-to-one correspondence with right prime morphisms M.) Representing R and such an M by matrices (r_{ij}) and (m_{ij}) as in section 2, it follows that

$$\sum_{j=1}^{k} r_{ij} m_{jh} = 0 \quad \text{for all } i = 1, \dots, l, \quad h = 1, \dots, m. \tag{4}$$

We can therefore consider the columns of M as relations between the columns of R. Conversely, relations between the columns of R determine a morphism whose image lies in the kernel of R. Consider now the module of all relations between the columns of R. Generators of this module, say g in number, determine a morphism $M_0 : (\mathcal{C}^\infty)^g \to (\mathcal{C}^\infty)^k$, whose image is clearly the largest subbehavior of $\mathbf{B}$ which admits an MA representation. Denote by $\mathbf{M_0}$ the submodule of $\mathcal{A}^g$ generated by the rows of the matrix representation of M_0. Consider next the submodule $\mathbf{R_0}$ of relations between the rows of the matrix representation of M_0. By (4) it follows that the rows of R lie in this module of relations. By Ehrenpreis–Palamodov, the image of M_0 is precisely the kernel of R_0, the morphism determined by $\mathbf{R_0}$. Thus we have the following.

THEOREM 1. *Let* $R : (\mathcal{C}^\infty)^k \to (\mathcal{C}^\infty)^l$ *be an AR system and let* $\mathbf{R}$ *be the submodule of* $\mathcal{A}^k$ *generated by the rows of* R*. Then the system admits an MA representation iff the module* $\mathbf{R_0}$ *defined above equals* $\mathbf{R}$.

Proof. By the theorem of Oberst, $\mathbf{R}$ and $\mathbf{R_0}$ define the same behavior iff they are equal. By construction the behavior of $\mathbf{R_0}$ is an MA behavior. □

Remark. We supplement the discussion preceding the above theorem with a few elementary remarks that we use elsewhere. Given a submodule of $\mathcal{A}^k$, define its rank to be the rank of the largest free submodule contained in it. It is easy to see that this rank is equal to the rank of any matrix whose rows generate the submodule and whose entries are now considered as belonging to the quotient field of the domain $\mathcal{A}$. It is equally elementary that this rank is also the dimension of the vector space obtained by tensoring the submodule with this quotient field. From this it follows that if the rank of a submodule $\mathbf{R}$ of $\mathcal{A}^k$ is i, then the rank of the submodule generated by all the relations between the columns of any matrix representation R of $\mathbf{R}$, namely the submodule $\mathbf{M_0}$ defined above, is $(k - i)$. We therefore conclude, similarly, that the rank of the submodule $\mathbf{R_0}$ in Theorem 1 equals the rank of $\mathbf{R}$. As $\mathbf{R}$ is contained in $\mathbf{R_0}$, it follows that for any p in $\mathbf{R_0}$, there is a nonzero a in $\mathcal{A}$ such that ap is in $\mathbf{R}$. In other words the quotient $\mathbf{R_0}/\mathbf{R}$ is a torsion module. We finally remark that from the construction of $\mathbf{M_0}$ and $\mathbf{R_0}$ above, it follows that if a p in $\mathcal{A}^k$ has the property that some ap is in $\mathbf{R}$ for a nonzero a in $\mathcal{A}$, then this p must be in $\mathbf{R_0}$. This is because if ap is in $\mathbf{R}$, then it lies in the module of relations of the rows of the matrix M_0, which by definition is $\mathbf{R_0}$. Thus the set of torsion elements of the module $\mathcal{A}^k/\mathbf{R}$ is precisely the submodule $\mathbf{R_0}/\mathbf{R}$. We can therefore reformulate the above theorem as follows.

THEOREM 2. *The AR system defined by a submodule* $\mathbf{R}$ *is MA iff* $\mathcal{A}^k/\mathbf{R}$ *is torsion free.*

An important case where the above theorem is applicable is when $\mathbf{R}$ can be "decoupled," namely, the proposition below.

PROPOSITION 3. *Let* $\mathbf{R}$ *be a submodule of* $\mathcal{A}^k$ *which is a direct summand. Then the AR behavior determined by it admits an MA representation.*

Proof. As $\mathbf{R}$ is a direct summand, it follows that $\mathcal{A}^k = \mathbf{R} \oplus \mathcal{A}^k/\mathbf{R}$. Thus $\mathcal{A}^k/\mathbf{R}$ is a projective module and hence free by Serre, Quillen, and Suslin. The result then follows from Theorem 2. □

The importance of this special case follows from the fact that this is *the* case for lumped systems.

COROLLARY 3 (1-D systems). *Let* $\mathcal{A} = \mathbb{R}[\frac{d}{dx}]$. *Then the behavior given by a submodule* $\mathbf{R}$ *of* $\mathcal{A}^k$ *is MA iff* $\mathbf{R}$ *is a direct summand.*

Proof. Suppose that the behavior of the submodule $\mathbf{R}$ is MA. Then $\mathbf{R}$ equals $\mathbf{R_0}$. By the above theorem $\mathcal{A}^k/\mathbf{R}$ is torsion-free. As $\mathcal{A}$ here is a PID; this implies that $\mathcal{A}^k/\mathbf{R}$ is free. It then follows that $0 \to \mathbf{R} \to \mathcal{A}^k \to \mathcal{A}^k/\mathbf{R} \to 0$ splits, i.e., that $\mathbf{R}$ is a direct summand. □

Remark. It is a result of Willems that every controllable lumped system is MA. Thus the above proposition actually provides a necessary and sufficient condition for the controllability of lumped systems.

We conclude this development with the following classical example.

Example 1 (the deRham complex on $\mathbb{R}^3$). Consider the behavior given by the kernel of the map $R : (\mathcal{C}^\infty)^3 \to (\mathcal{C}^\infty)^3$ whose matrix representation is given by

$$\begin{pmatrix} 0 & \frac{-\partial}{\partial z} & \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} & 0 & \frac{-\partial}{\partial x} \\ \frac{-\partial}{\partial y} & \frac{\partial}{\partial x} & 0 \end{pmatrix}.$$

Then the relations between the columns of R determine a map $M_0 : \mathcal{C}^\infty \to (\mathcal{C}^\infty)^3$ which is given by the matrix $(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z})^T$. As explained above, M_0 determines the largest MA subbehavior contained in the kernel of R. The module of relations $\mathbf{R_0}$ of the rows of M_0 is a submodule of $\mathcal{A}^3$ generated by $(0, \frac{-\partial}{\partial z}, \frac{\partial}{\partial y})$, $(\frac{\partial}{\partial z}, 0, \frac{-\partial}{\partial x})$, $(\frac{-\partial}{\partial y}, \frac{\partial}{\partial x}, 0)$, which coincides with $\mathbf{R}$, the module generated by the rows of R above. Thus by the above theorem, the kernel of R is an MA behavior, given by the image of M_0. Similarly, if one considers the map $R_1 : (\mathcal{C}^\infty)^3 \to \mathcal{C}^\infty$, given by $(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z})$, then the module of relations of its columns is generated by the columns of R. Moreover, the module of relations between the rows of R is a cyclic module, again generated by $(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z})$. Thus the kernel of R_1 is equal to the image of R. Note that this is just the familiar fact that the image of the gradient equals the kernel of curl and that the image of curl equals the kernel of divergence. □

Willems, and Rocha and Willems, prove that controllable AR systems are MA for 1-D and discrete 2-D systems by showing that MA behaviors are the kernels of left prime morphisms. We show below that while this is also true for distributed 2-D systems, n-D behaviors given by left prime morphisms are not necessarily controllable for $n \geq 3$ (Example 2). We also show that controllable systems, although MA, are not necessarily given by free submodules.

We first characterize the class of controllable AR systems.

Theorem 3. *Every distributed controllable AR system is an MA system.*

Proof. We show that if an AR system, say defined by the morphism $R : (\mathcal{C}^\infty)^k \to (\mathcal{C}^\infty)^l$, is not MA, then it is not controllable.

So assume that the behavior given by $\ker(R)$, the kernel of R, is not MA. By Theorem 1 above, the submodule $\mathbf{R}$ corresponding to R is strictly contained in the submodule $\mathbf{R_0}$. Let $p = (p_1, \ldots, p_k)$ be any element in $\mathbf{R_0}$ which is not in $\mathbf{R}$. It follows from the remark following Theorem 1 that ap is in $\mathbf{R}$ for some nonzero a in $\mathcal{A}$. Consider the maps

$$\begin{aligned} P : \ker(R) &\to \mathcal{C}^\infty, \\ (f_1, \ldots, f_k) &\mapsto \sum p_i(f_i), \end{aligned}$$

and

$$\begin{aligned} A : \mathcal{C}^\infty &\to \mathcal{C}^\infty, \\ f &\mapsto a(f), \end{aligned}$$

where p and a are as above.

By the theorem of Oberst quoted in the introduction, as p is not in $\mathbf{R}$, it does not kill every element in $\ker(R)$, i.e., the map P defined above is not the zero map. However, the composition $A \circ P : \ker(R) \to \mathcal{C}^\infty$, which maps $(f_1, \ldots, f_k)$ in $\ker(R)$ to $\sum ap_i(f_i)$, is the zero map as ap is in $\mathbf{R}$.

Now let u be any element in $\ker(R)$ which is not in the kernel of P. Let U be any bounded open subset of $\mathbb{R}^n$, where $P(u)$ is nonzero, and let V be any compact set whose interior contains the closure of U. Let u_c be any cutoff with respect to U and V. Then u_c has compact support and is clearly not in the kernel of P. If $\ker(R)$ were controllable, then some cutoff of u with respect to this U and V must also be in the kernel of R. This implies that the image under P of this cutoff must be in the kernel of A. It follows then that the PDE a admits a nonzero compactly supported solution. This is a contradiction to the Paley–Wiener theorem. □

If we were to consider the behavior of R in $(\mathcal{D}')^k$, then the same proof as above shows that every distributed controllable AR system is an MA system. A variant of the above proof yields another characterization of controllable systems.

THEOREM 4. *A submodule* $\mathbf{R}$ *of* $\mathcal{A}^k$ *defines a controllable AR system iff* $\mathbf{R}$ *is Willems with respect to* $\mathcal{D}$ *(or* $\mathcal{E}'$*).*

Proof. We prove the theorem for the case of $\mathcal{D}$, the proof for $\mathcal{E}'$ being similar. Thus we have to show that $\mathbf{R}$ is controllable iff $\mathcal{M}(\mathcal{B}_{\mathcal{D}}(\mathbf{R})) = \mathbf{R}$, i.e., $\mathbf{R}$ is Willems with respect to $\mathcal{D}$.

Suppose first that $\mathcal{M}(\mathcal{B}_{\mathcal{D}}(\mathbf{R})) = \mathbf{R_1}$ strictly contains $\mathbf{R}$. Then by the theorem of Oberst, $\mathcal{B}_{\mathcal{C}^\infty}(\mathbf{R_1})$ is strictly contained in $\mathcal{B}_{\mathcal{C}^\infty}(\mathbf{R})$. As $\mathcal{B}_{\mathcal{D}} \circ \mathcal{M}$ is the identity (Lemma 1), it follows that $\mathcal{B}_{\mathcal{D}}(\mathbf{R}) = \mathcal{B}_{\mathcal{D}}(\mathbf{R_1})$ which is in turn contained in $\mathcal{B}_{\mathcal{C}^\infty}(\mathbf{R_1})$. Thus the closure of $\mathcal{B}_{\mathcal{D}}(\mathbf{R})$ in $\mathcal{C}^\infty(\mathbb{R}^n)$ is contained in $\mathcal{B}_{\mathcal{C}^\infty}(\mathbf{R_1})$. This implies that the closure of $\mathcal{B}_{\mathcal{D}}(\mathbf{R})$ is strictly contained in $\mathcal{B}_{\mathcal{C}^\infty}(\mathbf{R})$. Thus $\mathbf{R}$ cannot be controllable, since the compactly supported behavior of any controllable system is dense in its $\mathcal{C}^\infty$-behavior.

Conversely, suppose the $\mathcal{C}^\infty$-behavior given by $\mathbf{R}$ is not controllable. We have to show that $\mathbf{R}$ is not Willems with respect to $\mathcal{D}$; i.e., $\mathbf{R}$ is strictly contained in $\mathcal{M}(\mathcal{B}_{\mathcal{D}}(\mathbf{R}))$. Since $\mathbf{R}$ is not controllable, $\mathbf{R}$ is not MA by the above theorem. Hence by Theorem 1, $\mathbf{R}$ is strictly contained in $\mathbf{R_0}$. This implies that $\mathcal{B}_{\mathcal{D}}(\mathbf{R_0}) \subset \mathcal{B}_{\mathcal{D}}(\mathbf{R})$. In fact we claim that these two compactly supported behaviors are identical, i.e., every p in $\mathbf{R_0}$ kills every element in $\mathcal{B}_{\mathcal{D}}(\mathbf{R})$. For suppose that this was not true for some p in $\mathbf{R_0}$. By the remark following Theorem 1, there exists some a in $\mathcal{A}$ such that ap is in $\mathbf{R}$. Thus ap kills every element in $\mathcal{B}_{\mathcal{D}}(\mathbf{R})$. So if p does not kill some element in $\mathcal{B}_{\mathcal{D}}(\mathbf{R})$, it follows as in the proof of the above theorem that the Paley–Wiener theorem is contradicted. Hence $\mathbf{R}$ is strictly contained in $\mathbf{R_0} \subset \mathcal{M}(\mathcal{B}_{\mathcal{D}}(\mathbf{R_0})) = \mathcal{M}(\mathcal{B}_{\mathcal{D}}(\mathbf{R}))$. Thus $\mathbf{R}$ is not Willems with respect to $\mathcal{D}$. □

We have thus established the following equivalences: $\mathbf{R}$ is controllable $\Leftrightarrow$ $\mathbf{R}$ is MA $\Leftrightarrow$ $\mathbf{R}$ is Willems with respect to $\mathcal{D}$ (or $\mathcal{E}'$) $\Leftrightarrow$ $\mathcal{A}^k/\mathbf{R}$ is torsion-free.

Rocha and Willems in [6] show that every controllable 2-D discrete system is given by a factor left prime matrix (whose rows necessarily generate a free submodule). We have already pointed out the importance of this result in the introduction. The question then arises as to whether this is so for distributed systems as well. This is not in general true as the following counterexample demonstrates.

Example 2. Let $R : (\mathcal{C}^\infty)^3 \to (\mathcal{C}^\infty)^2$ be the AR system determined by the following matrix:

$$R = \begin{pmatrix} 0 & -\frac{\partial}{\partial z} & \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} & 0 & -\frac{\partial}{\partial x} \end{pmatrix}.$$

Clearly R is factor left prime. Nonetheless the AR behavior determined by R is *not* controllable, as an easy calculation shows that R_0 is the matrix that appears in Example 1, namely curl. The system defined by curl is of course controllable, since it is the image of the gradient map. Note, however, that the matrix representation of curl is not factor left prime; in fact the submodule generated by the rows of this matrix is not even free. □

Thus the question now arises as to which controllable systems are given by free submodules. As pointed out in the introduction, this is an important question. We provide a necessary and sufficient condition below in the notation preceding Theorem 1.

THEOREM 5. *Let* $\mathbf{R}$ *be a submodule of* $\mathcal{A}^k$ *which determines a controllable behavior. Let* $\mathbf{R}$ *determine the submodule* $\mathbf{M_0}$ *of* $\mathcal{A}^g$ *as above. Then* $\mathbf{R}$ *is a free submodule iff the projective dimension of* $\mathcal{A}^g/\mathbf{M_0}$ *is less than or equal to* 2.

Proof. Consider the following resolution of $\mathcal{A}^g/\mathbf{M_0}$:

$$0 \to \mathbf{R_0} \xrightarrow{\phi_3} \mathcal{A}^k \xrightarrow{\phi_2} \mathcal{A}^g \xrightarrow{\phi_1} \mathcal{A}^g/\mathbf{M_0} \to 0,$$

where ϕ_1 and ϕ_3 are canonically defined and ϕ_2 takes a basis of $\mathcal{A}^k$ to the rows of the matrix M_0. By the discussion preceding Theorem 1, the kernel of ϕ_2 is the module $\mathbf{R_0}$, and hence the sequence above is indeed exact. As the behavior of $\mathbf{R}$ is controllable, $\mathbf{R_0}$ equals $\mathbf{R}$ by Theorem 3. A standard argument shows that $\mathrm{Ext}^1(\mathbf{R}, -) \approx \mathrm{Ext}^3(\mathcal{A}^g/\mathbf{M_0}, -)$, which equals 0 iff the projective dimension of $\mathcal{A}^g/\mathbf{M_0}$ is at most 2. This implies that $\mathbf{R}$ is projective. Now by Serre, Quillen, and Suslin, $\mathbf{R}$ must be free. $\square$

COROLLARY 4 (2-D distributed systems). *Every distributed 2-D controllable behavior is given by a free submodule.*

Proof. Observe that here $\mathcal{A} = \mathbb{R}[\partial_1, \partial_2]$. Thus the global dimension of $\mathcal{A}$ is 2 and hence the projective dimension of any $\mathcal{A}$-module is less than or equal to 2. This then satisfies the conditions of the above proposition. $\square$

Remark. From the proof of Theorem 3, it can be easily seen that if a controllable AR system is given by a free submodule $\mathbf{R}$, then $\mathbf{R}$ is maximal among all submodules with the same number of generators as $\mathbf{R}$. This would imply that the matrix representation of $\mathbf{R}$ is factor left prime. From the above corollary we can therefore further conclude that every distributed 2-D controllable behavior is given by a factor left prime matrix.

Thus the phenomenon exhibited in Example 2 above occurs only in dimension 3 or more.

Remark. It is now well known that there is an important difference between 1-D and n-D, $n \geq 2$, discrete systems, namely that while every 1-D system is feedback stabilizable, this is not, however, the case for 2-D systems (see Shankar and Sule [8]). The counterexample and corollary above point out, likewise, a difference between 2-D systems and n-D, $n \geq 3$, systems.

4. Autonomous systems. In [10, 11] Willems defines a lumped autonomous AR behavior to be one given by the kernel of a map R in $\mathbb{R}^{k\times k}[\frac{d}{dt}]$ whose characteristic polynomial, i.e., its determinant, is a nonzero polynomial. Intuitively, this definition singles out those AR behaviors that are far from being controllable in the sense that every element in the behavior of an autonomous system is determined by its values on any open interval of $\mathbb{R}$, i.e., by its germ at any point in $\mathbb{R}$. Thus if we specify the germ of an element in such a behavior, that element is specified in its entirety. We therefore have no "control" over it; i.e., it is "autonomous." In fact the elements in the behavior of a lumped autonomous system are entire functions which are linear combinations of exponential functions that arise from the zeros of the characteristic polynomial. As these zeros are finite in number, the exponential solutions span a finite-dimensional $\mathbb{R}$-vector subspace in the space of entire as well as $\mathcal{C}^\infty$ functions. Being finite dimensional, this subspace is closed, and hence by the Malgrange approximation theorem, the $\mathcal{C}^\infty$-behavior of an autonomous lumped system R also consists only of these entire functions. (In fact this finite dimensionality of the behavior is equivalent to R having full rank.) This finite dimensionality of the $\mathcal{C}^\infty$-behavior implies that if the derivatives up to some finite order of an element all vanish at a point, then this element must be the zero element. Hence an element in such a behavior can be specified by a "finite amount of data." This is considerably stronger than saying that the $\mathcal{C}^\infty$-behavior of a system consists only of entire functions. Note finally that the behavior of a single ordinary differential equation is autonomous.

We now wish to extend this definition of an autonomous behavior to distributed systems. So consider a distributed AR behavior given by the kernel of a morphism R. This behavior corresponds to a submodule of $\mathcal{A}^k$ generated by the rows of R. Observe that submodules of $\mathcal{A}^k$ can have larger than k number of generators when $\mathcal{A} = \mathbb{R}[\partial_1, \ldots, \partial_n]$, $n > 1$. (This does not happen in the lumped case, i.e., when $n = 1$, as then every submodule is free and thus requires fewer than or equal to k generators.) Thus in the distributed case, the matrix representation of R might necessarily need more than k rows. Like the behavior of an autonomous lumped system, we would expect the behavior of a distributed autonomous system R to be dependent on the various $k \times k$ submatrices of R. To describe this we use the following notation from Lang [4]. Let M be an $l \times m$ matrix with coefficients in $\mathcal{A}$. For $r \leq \min(l, m)$, the rth determinant ideal of M, denoted $I_r(M)$, is the ideal of $\mathcal{A}$ generated by the determinants of all the $r \times r$ minors of M. If $r > \min(l, m)$, $I_r(M)$ is defined to be the zero ideal. Define also the rth determinant variety of M to be the variety in $\mathbb{C}^n$ of $I_r(M)$.

In terms of this, a lumped autonomous behavior is one given by an $R \in \mathbb{R}^{k \times k}[\frac{d}{dt}]$ whose kth determinant ideal is not the zero ideal. It is this property that we generalize to define a distributed autonomous behavior. As behaviors correspond to submodules of $\mathcal{A}^k$, we first state the following lemma whose proof easily follows from the Cauchy–Binet formula.

Lemma 3. *Let* $\mathbf{R}$ *be a submodule of* $\mathcal{A}^k$. *Let* R_1 *and* R_2 *be two matrices whose rows generate* $\mathbf{R}$. *Then for every* i, *the* i*th determinant ideals of* R_1 *and* R_2 *are identical, and therefore their* i*th determinant varieties coincide.*

We can now make the following definition.

Definition 4. *Let* $\mathbf{R}$ *be any submodule of* $\mathcal{A}^k$. *Define the characteristic ideal of* $\mathbf{R}$ *to be the* k*th determinant ideal of any matrix representation of* $\mathbf{R}$. *The variety of this ideal* (*in* $\mathbb{C}^n$) *is called the characteristic variety of* $\mathbf{R}$. *We denote the characteristic ideal (characteristic variety) of* $\mathbf{R}$ *by* $I(\mathbf{R})(V(\mathbf{R}))$.

As pointed out above, a lumped autonomous behavior corresponds to an $\mathbf{R}$ whose characteristic variety is not all of $\mathbb{C}$. Hence we make the following definition.

Definition 5. *The behavior given by a submodule* $\mathbf{R}$ *of* $\mathcal{A}^k$ *is said to be* autonomous *if the characteristic variety of* $\mathbf{R}$ *is not all of* $\mathbb{C}^n$ *(i.e., its* k*th determinant ideal is not the zero ideal).*

We show below that if a behavior is not autonomous, then it contains nontrivial controllable subbehaviors. We wish to remark first that behaviors of autonomous systems whose characteristic varieties are nonempty contain exponentials corresponding to points on this variety. This is because if R is any matrix representation of $\mathbf{R}$, then substituting any point ξ of $V(\mathbf{R})$ into the entries of R will result in a matrix $R(\xi)$ with entries in $\mathbb{C}$ whose column rank is less than k. This implies that there is a nonzero element in the kernel of $R(\xi)$, say $(c_1, \ldots, c_k)$. Then an easy check shows that $(c_1 e^{\langle x, \xi \rangle}, \ldots, c_k e^{\langle x, \xi \rangle})$ is in the behavior of $\mathbf{R}$. Thus there are exponential solutions in the behavior corresponding to every point in $V(\mathbf{R})$.

Clearly any matrix representation of a distributed autonomous system has full column rank, i.e., there are no polynomial relations between the columns of R. As pointed out above, a submodule of $\mathcal{A}^k$ might have more than k generators, and thus every matrix representation of it will have more than k rows. Therefore a distributed autonomous system need not have a square matrix representation, for which reason we specify its column rank. This therefore is the right generalization of Willems's definition of lumped autonomous systems as those whose matrix representations have full rank.

Note also that as in the lumped case, the behavior of a single (nonzero) partial differential equation is autonomous. We further substantiate this definition below.

Proposition 4. *A behavior is autonomous iff it does not contain any (nontrivial) controllable subbehaviors.*

Proof. Suppose that the behavior of a submodule $\mathbf{R}$ of $\mathcal{A}^k$ is autonomous. Let R be any matrix representation of $\mathbf{R}$. Let R_1 be a $k \times k$ submatrix of R whose determinant Δ is nonzero. Clearly the behavior of R is contained in that of R_1. In turn the behavior of R_1 is contained in the behavior of $R_1^* R_1 = R_1 R_1^* = \operatorname{diag}(\Delta)$, where R_1^* is the adjoint matrix of R_1. Thus each entry in an element of the behavior of R lies in the kernel of Δ. As Δ is nonzero, none of these can have compact support (by Paley–Wiener). By Lemma 2, a controllable behavior necessarily has elements with compact support. Hence an autonomous behavior does not contain any nontrivial controllable subbehaviors.

Conversely, suppose that the behavior of R is not autonomous. This means that R does not have full column rank, and thus that there are nontrivial relations between the columns of R. Let $p = (p_1, \ldots, p_k)$ be such a relation. Then Rp^T is equal to zero. This implies that the image of the map $p^T : \mathcal{C}^\infty \to (\mathcal{C}^\infty)^k$ is contained in the behavior of R. Thus the behavior of R contains MA subbehaviors, which are controllable by Proposition 2. □

Corollary 5. *No nonzero element in an autonomous behavior can vanish outside a compact subset of* $\mathbb{R}^n$.

Proof. The proof is clear from the proof of Proposition 4. □

For lumped systems, the characteristic variety of an autonomous system is discrete, as it is not all of $\mathbb{C}$. However for distributed autonomous systems, this need not be the case. As a result, many of the properties of an autonomous lumped system that are consequences of the discreteness of the characteristic variety, do not carry over to distributed systems. To capture these properties, we present another definition.

Definition 6. *The behavior given by a submodule* $\mathbf{R}$ *of* $\mathcal{A}^k$ *is said to be* strongly autonomous *if the characteristic variety* $V(\mathbf{R})$ *of* $\mathbf{R}$ *is discrete* (*i.e., if* $\mathcal{A}/I(\mathbf{R})$ *has* (*Krull*) *dimension* 0).

Note that as $\mathcal{A}$ is Noetherian, discreteness of $V(\mathbf{R})$ implies that it is finite.

We show in the following theorem that a strongly autonomous distributed system satisfies the property of finite dimensionality of lumped autonomous behaviors. We first prove the following preliminary results.

Lemma 4. *Let I be a maximal ideal in $\mathcal{A} = \mathbb{R}[\partial_1, \ldots, \partial_n]$. Then $\mathcal{B}(I)$, the $\mathcal{C}^\infty$-behavior given by I, is finite dimensional.*

Proof. Observe first that as I is maximal, its variety in $\mathbb{C}^n$ consists of a finite number of points (in fact at most 2^n in number). Assume first that this variety intersects $\mathbb{R}^n$, say at $\xi = (\xi_1, \ldots, \xi_n)$. Then the ideal I must equal $(\partial_1 - \xi_1, \ldots, \partial_n - \xi_n)$ and the variety of this ideal must consist of this point alone. To say that an exponential solution is in $\mathcal{B}(I)$ is to say that $(\partial_i - \xi_i)(q(x)e^{\langle x, \xi \rangle}) = 0$ for all i. This forces $q(x)$ to be a constant. As the $\mathcal{C}^\infty$-behavior is the closed linear hull of the exponential solutions (by Malgrange), $\mathcal{B}(I)$ is 1-dimensional (and is in fact spanned by $e^{\langle x, \xi \rangle}$).

Suppose now that the variety does not intersect $\mathbb{R}^n$. Then it is a finite set, say $\{\xi^1, \ldots, \xi^k\} \subset \mathbb{C}^n$. The point $\xi^j = (\xi_1^j, \ldots, \xi_n^j)$ corresponds to the maximal ideal $(\partial_1 - \xi_1^j, \ldots, \partial_n - \xi_n^j)$ in $\mathbb{C}[\partial_1, \ldots, \partial_n]$. As above, the space of exponential solutions corresponding to this point is 1-dimensional. Thus the space of the exponential solutions in $\mathcal{B}(I)$ is the k-dimensional space spanned by the exponential solutions corresponding to each of the k points in the variety of I. Being finite dimensional, the closure of this space in $\mathcal{C}^\infty$ is itself. □

LEMMA 5. *Let $I = (p_1, \ldots, p_i, \ldots, p_N)$ be an ideal in $\mathbb{R}[\partial_1, \ldots, \partial_n]$ whose $\mathcal{C}^\infty$ behavior is finite dimensional. Then the $\mathcal{C}^\infty$-behavior of the ideal $J = (p_1, \ldots, p_i^2, \ldots, p_N)$ is also finite dimensional.*

Proof. Let I' be the ideal $(p_1, \ldots, \hat{p_i}, \ldots, p_N)$, where $\hat{p_i}$ means that p_i has been omitted. Consider the following $\mathbb{R}$-linear map:

$$P_i : \mathcal{B}(I') \to \mathcal{C}^\infty(\mathbb{R}^n),$$
$$f \mapsto p_i(f).$$

Then the behavior $\mathcal{B}(I)$ of I is the kernel of P_i, which by assumption is finite dimensional. Hence $P_i^{-1}(\mathcal{B}(I))$ is also finite dimensional. But $P_i^{-1}(\mathcal{B}(I))$ is precisely the behavior of J. □

COROLLARY 6. *Let I be an ideal in $\mathbb{R}[\partial_1, \ldots, \partial_n]$ such that the behavior of its radical, $\sqrt{I}$, is finite dimensional. Then the behavior of I is also finite dimensional.*

Proof. Some power of $\sqrt{I}$, say the rth power, is contained in I (as $\mathbb{R}[\partial_1, \ldots, \partial_n]$ is Noetherian!). Let J be the ideal generated by the rth powers of the generators of $\sqrt{I}$. An easy induction using Lemma 5 shows that $\mathcal{B}(J)$ is finite dimensional. However, now J is contained in I. So $\mathcal{B}(I)$ is contained in $\mathcal{B}(J)$, and hence $\mathcal{B}(I)$ is also finite dimensional. □

THEOREM 6. *A behavior is strongly autonomous iff it is a finite-dimensional subspace of the $\mathbb{R}$-vector space of $\mathcal{C}^\infty$ functions.*

Proof. Let a behavior $\mathbf{B}$ corresponding to a submodule $\mathbf{R}$ be strongly autonomous. Let $I(\mathbf{R})$ be generated by $\Delta_1, \ldots, \Delta_l$ in $\mathcal{A}$. Then every component of $f = (f_1, \ldots, f_k)$ in $\mathbf{B}$ is also a solution of Δ_i, $i = 1, \ldots, l$.

Now consider the variety V of $I(\mathbf{R})$. By assumption it is discrete. Hence the ideal of this variety, which is the radical of $I(\mathbf{R})$, is a finite intersection of maximal ideals. Now the exponential solutions of $\sqrt{I(\mathbf{R})}$ is the union of the exponential solutions corresponding to each of these maximal ideals. By Lemma 4, this spans a finite-dimensional $\mathbb{R}$-vector space. Thus by Malgrange, this space is also the $\mathcal{C}^\infty$-behavior of $\sqrt{I(\mathbf{R})}$. By Corollary 6, the $\mathcal{C}^\infty$-behavior of I is finite dimensional as well.

Conversely, suppose that $\mathbf{B}$ is not strongly autonomous. Then as $V(\mathbf{R})$ is not finite, it must contain an infinite number of points in $\mathbb{C}^n$. By the discussion following Definition 5, each of these points contribute nonzero exponential elements in $\mathbf{B}$. Clearly these exponential elements corresponding to different points in $V(\mathbf{R})$ are linearly independent. So $\mathbf{B}$ is not finite dimensional. □

COROLLARY 7. *Every $\mathcal{C}^\infty$ solution of a strongly autonomous system is entire. Thus no nonzero element in a strongly autonomous behavior can vanish on any open subset of $\mathbb{R}^n$.*

Proof. The proof is clear from the first part of the proof of Theorem 6. □

As a strongly autonomous system is clearly autonomous as well, Corollary 5 is valid for such systems. In fact even more is true, namely, the following.

COROLLARY 8. *No cutoff of an element in a strongly autonomous behavior $\mathbf{B}$ is in $\mathbf{B}$.*

COROLLARY 9. *An element in a strongly autonomous behavior $\mathbf{B}$ is determined by its values on any open subset of $\mathbb{R}^n$.*

Proof. If w_1 and w_2 are two elements in $\mathbf{B}$ that agree on some open subset of U of $\mathbb{R}^n$, then $w_1 - w_2$ is an element of $\mathbf{B}$ that vanishes on U. Then by Corollary 7, $w_1 - w_2$ equals zero. □

Remark. In fact much more is true. By the finite dimensionality of a strongly autonomous behavior $\mathbf{B}$, every element in it can be determined by its derivatives up to some finite order. It is in order to mimic this property of a lumped autonomous behavior that we made the above definition. In the case of a lumped system, the $\mathcal{C}^\infty$-behavior either contains elements that are not entire or is finite dimensional. For a distributed system, the behavior could consist of only entire functions but may not be finite dimensional. This will happen, for instance, if the behavior is specified by a single partial differential operator which is elliptic. As a result, such behaviors will also satisfy the above corollary but will not be strongly autonomous.

5. Control. Let $\mathbf{B}$ be a $\mathcal{C}^\infty$ AR behavior and let $\mathbf{R}$ be the submodule of $\mathcal{A}^k$ corresponding to it. Let $\{\mathbf{R}_\alpha\}$ be the class of submodules of $\mathcal{A}^k$ containing $\mathbf{R}$. The $\mathcal{C}^\infty$ AR behaviors given by any of these $\mathbf{R}_\alpha$'s is clearly contained in $\mathbf{B}$. This, by definition, is the class of AR subbehaviors of $\mathbf{B}$. (In fact, by Oberst, there is a one-to-one correspondence between the subbehaviors of $\mathbf{B}$ and the collection $\{\mathbf{R}_\alpha\}$.)

Let R be a matrix whose rows generate $\mathbf{R}$ so that $\mathbf{B}$ is the kernel of R. Let $\mathbf{B}_\alpha$ be a subbehavior of $\mathbf{B}$, and let $\mathbf{R}_\alpha$ be a submodule of $\mathcal{A}^k$ that generates $\mathbf{B}_\alpha$. As $\mathbf{R}_\alpha$ contains $\mathbf{R}$, we can obtain a matrix representation R_α by appending rows to R.

Suppose now that the behavior $\mathbf{B}$ is controllable. Then by Theorem 3, $\mathbf{B}$ is an MA behavior, i.e., suppose that $\mathbf{B}$ is the image of a morphism $M : (\mathcal{C}^\infty)^m \to (\mathcal{C}^\infty)^k$. We wish to explain how to obtain the subbehaviors $\{\mathbf{B}_\alpha\}$ of $\mathbf{B}$ in this setting.

Observe that the behavior $\mathbf{B}$ is the image under M of $(\mathcal{C}^\infty)^m$, which in turn can be thought of as the AR behavior given by the 0 submodule of $\mathcal{A}^m$. We claim that the subbehaviors of $\mathbf{B}$ can all be obtained as the images under M of the various AR subbehaviors of $(\mathcal{C}^\infty)^m$. Clearly, without loss of generality, it suffices to consider only those AR subbehaviors of $(\mathcal{C}^\infty)^m$ that contain the kernel of M.

Let $\mathbf{M}$ be the submodule of $\mathcal{A}^m$ generated by the rows of the matrix representation of the morphism M. Clearly any AR subbehavior of $(\mathcal{C}^\infty)^m$ that contains the kernel of M corresponds to a submodule $\mathbf{K}$ of $\mathbf{M}$. Let K be the morphism $(\mathcal{C}^\infty)^m \to (\mathcal{C}^\infty)^r$ corresponding to the submodule $\mathbf{K}$ (where r is the number of generators for $\mathbf{K}$). Clearly $K = TM$ for some morphism $T : (\mathcal{C}^\infty)^k \to (\mathcal{C}^\infty)^r$ (see the diagram below).

$$\begin{tikzcd}
(\mathcal{C}^\infty)^m \arrow[r, "M"] \arrow[d, "K"'] & (\mathcal{C}^\infty)^k \arrow[r, "R"] \arrow[dl, "T"] & (\mathcal{C}^\infty)^l \\
(\mathcal{C}^\infty)^r. & &
\end{tikzcd}$$

Equally clearly, the image under M of the kernel of K equals the intersection of the image of M with the kernel of T. By assumption the image of M is the AR behavior given by the morphism R. Thus the image under M of the kernel of K equals the intersection of the kernels of R and T. Hence we obtain a subbehavior $\mathbf{B}_\alpha$ (namely, by appending the rows of the matrix representation of the morphism T to the rows of R) as the image under M of a subbehavior of $(\mathcal{C}^\infty)^m$.

Control problems are always accompanied by criteria which single out certain subbehaviors as desirable. Such criteria are usually in the nature of stability or optimality requirements. In this paper we are concerned with stability requirements, which we model after the lumped situation. There, a very fruitful notion of stability has been the notion of bounded input–bounded output stability, where the growth of the signals in the system (i.e., inputs and outputs) is specified along the half-line $\mathbb{R}_+$, i.e., as the independent variable (time) tends to $+\infty$. We wish to generalize this notion of stability to distributed systems.

Definition 7. *The directions of stability is a closed convex cone S in $\mathbb{R}^n$ (with vertex at the origin). A $\mathcal{C}^\infty$-behavior* $\mathbf{B}$ *is stable with respect to S if every element in* $\mathbf{B}$ *tends to zero along every half-line in S.*

Remark. By a cone in $\mathbb{R}^n$ we mean a subset S, such that if x is in S, then tx is also in S, for every $t > 0$. Given a closed cone S, we also define the subset $S_<$ of $\mathbb{R}^n$ consisting of those points y in $\mathbb{R}^n$ such that $\langle y, x \rangle < 0$ for every nonzero x in S. Clearly $S_<$ is also a cone in $\mathbb{R}^n$. It is easy to check that $S_<$ is nonempty if S is a proper closed cone, i.e., S does not contain a full line. In fact $S_<$, in this case, has nonempty interior.

Proposition 5. *Let* $\mathbf{R}$ *be a submodule of $\mathcal{A}^k$ and* $\mathbf{B}$ *its behavior. If the projection of the characteristic variety of* $\mathbf{R}$*, $V(\mathbf{R})$, to $\mathbb{R}^n$ does not lie in $S_<$, then* $\mathbf{B}$ *is not stable with respect to S.*

Proof. If there is a point in the projection of $V(\mathbf{R})$ which does not lie in $S_<$, then the exponential solution corresponding to this point will not tend to zero along at least one half-line in S. □

It is not clear whether the assumption in the above proposition is also necessary. However, an additional hypothesis, which is generically satisfied, does guarantee stability.

Proposition 6. *Suppose that the characteristic ideal $I(\mathbf{R})$ of* $\mathbf{R}$ *contains a polynomial without repeated factors. Then if the projection of the characteristic variety is contained in $S_<$, and if its distance from the boundary of $S_<$ is strictly positive, then the behavior of* $\mathbf{R}$ *is stable with respect to S.*

Proof. As observed before, every component of every element in the behavior of $\mathbf{R}$ is also a homogeneous solution of every polynomial in the characteristic ideal $I(\mathbf{R})$ of $\mathbf{R}$. By assumption there is a polynomial p in $I(\mathbf{R})$ without repeated factors. Then by a remark in [2] quoted earlier at the end of section 2, it follows that the linear hull of exponential solutions of the form $e^{\langle x, \xi \rangle}$ is dense in the behavior of p. As the behavior of $I(\mathbf{R})$ is a closed subspace of the behavior of p, it follows that the linear hull of such exponentials (i.e., those without polynomial factors) is also dense in the behavior of $I(\mathbf{R})$.

We now claim that given any $\epsilon > 0$, there exists a ball, say B_ϵ, such that all the above exponential solutions (i.e., those of the form $e^{\langle x, \xi \rangle}$) are less in absolute value than ϵ at every point in $B_\epsilon^C \cap S$. This is because, by assumption, the projection of the characteristic variety to $\mathbb{R}^n$ is at a strictly positive distance from the boundary of $S_<$. Therefore every element in the behavior of $I(\mathbf{R})$ tends to zero along every half-line in S. This is therefore also true of every element in the behavior of $\mathbf{R}$. □

Observe that as the characteristic variety of a nonautonomous behavior $\mathbf{B}$ is all of $\mathbb{C}^n$, it contains exponentials corresponding to every $p \in \mathbb{C}^n$. Thus $\mathbf{B}$ cannot be stable with respect to any cone S in $\mathbb{R}^n$. Stability with respect to a cone S is therefore a property of autonomous behaviors. Hence by Proposition 4, a behavior that contains a controllable subbehavior is not stable with respect to any cone S. Restricting to an autonomous subbehavior stable with respect to a cone S is the process of *control* in stability problems. More generally we define control as the process of restricting $\mathbf{B}$ to some autonomous subbehavior.

Assume therefore that we are given an AR behavior $\mathbf{B}$ defined by a morphism $\mathbf{R}$. Let R be an $l \times k$ matrix representation of $\mathbf{R}$. Let K be a $j \times k$ matrix whose rows, when appended to the rows of R, define an autonomous *nontrivial* subbehavior $\mathbf{B}'$ of $\mathbf{B}$. The system defined by the matrix K we call the *controller.* If this subbehavior $\mathbf{B}'$, defined by $\left[\begin{smallmatrix} R \\ K \end{smallmatrix} \right]$, is stable with respect to a cone S, then we call K a stabilizing controller for R with respect to the cone S. We have already explained above how to obtain a controller if $\mathbf{B}$ is also a controllable behavior.

THEOREM 7. *A controllable behavior can be stabilized with respect to any proper cone S. In fact there is a controller that restricts this behavior to a strongly autonomous behavior stable with respect to S.*

Proof. As S is proper, $S_<$ is nonempty. Given any ξ in $\mathbb{C}^n$ which projects into $S_<$, the function $e^{\langle x,\xi\rangle}$, $x \in \mathbb{R}^n$, is stable with respect to S. The set of all such ξ has nonempty interior.

Observe now that if $(e^{\langle x,\xi\rangle}, \dots, e^{\langle x,\xi\rangle})$ lies in the kernel of a morphism $M : (\mathcal{C}^\infty)^m \to (\mathcal{C}^\infty)^k$ (the image of which defines the given behavior; for being controllable it is MA), then ξ must lie in the characteristic variety of M. We can assume that this characteristic variety is not all of $\mathbb{C}^n$ by Proposition 1 in section 2. Thus this variety has empty interior. Therefore for almost all ξ which project into $S_<$, the function $(e^{\langle x,\xi\rangle}, \dots, e^{\langle x,\xi\rangle})$ does not lie in the kernel of M.

Let $\xi_1, \dots, \xi_r$ be any finite set of points that project into $S_<$ and such that the corresponding exponentials do not lie in the kernel of M. Assume further that $\xi_1, \dots, \xi_r$ is closed under conjugation. Then $\xi_1, \dots, \xi_r$ is an affine variety in $\mathbb{C}^n$ given by an ideal say, $(p_1, \dots, p_t)$ in $\mathcal{A}$; i.e., each p_i has real coefficients. Now consider the matrix

$$\begin{pmatrix} 1 & & & & \\ & 1 & & 0 & \\ & & \ddots & & \\ & & & 1 & \\ & & & & p_1 \\ & 0 & & & \vdots \\ & & & & p_t \end{pmatrix}$$

in $\mathcal{A}^{(m+t-1)\times m}$. Clearly its determinantal ideal is $(p_1, \dots, p_t)$. The AR subbehavior of $(\mathcal{C}^\infty)^m$ defined by the kernel of this matrix is strongly autonomous and hence spanned by a finite number of exponential solutions that are all stable with respect to S. Thus the subbehavior itself is stable with respect to S. Clearly the image under M of such a subbehavior is also strongly autonomous and stable with respect to S. □

We now define the more general pole-placement problem for distributed systems which we pattern after the corresponding problem for lumped systems.

The pole-placement problem. Given a behavior $\mathbf{B}$ defined by R, and any ideal I in $\mathcal{A}$, is there a controller K such that the characteristic ideal of the augmented system $[\begin{smallmatrix} R \\ K \end{smallmatrix}]$ is I?

Observe that if the pole-placement problem can be solved for a behavior $\mathbf{B}$, then it contains nontrivial autonomous subbehaviors stable with respect to any proper cone S in $\mathbb{R}^n$.

A necessary condition for the solution of the above problem is the following.

LEMMA 6. *If the pole-placement problem can be solved for a behavior $\mathbf{B}$, then $\mathbf{B}$ is nonautonomous.*

Proof. Suppose $\mathbf{B}$ defined by R is autonomous. Then its characteristic ideal I is nonzero, and hence the characteristic ideal of any augmented system $[\begin{smallmatrix} R \\ K \end{smallmatrix}]$ contains I. □

The question therefore arises of whether the above is also sufficient. While we have not been able to solve this problem in all generality, we provide partial answers in what follows.

In this paper we treat the pole-placement problem only for free submodules $\mathbf{R}$ of $\mathcal{A}^k$. While this is not the most general situation for distributed systems, it is nonetheless the first generalization of Willems's results on pole placement for lumped systems [9], where everything is free. (Recall also our remarks on free submodules in the introduction.)

We begin by first considering a special class of AR systems introduced in Proposition 3, defined by submodules $\mathbf{R}$ which are direct summands of $\mathcal{A}^k$. By Proposition 3, a behavior defined by such an $\mathbf{R}$ is an MA behavior and hence is controllable. We have already shown that controllable behaviors can be stabilized with respect to any cone S. We now solve the pole-placement problem for such AR systems.

THEOREM 8. *The pole-placement problem can be solved for a behavior* $\mathbf{B}$ *defined by a submodule* $\mathbf{R}$ *which is a direct summand of* $\mathcal{A}^k$.

Proof. Since $\mathbf{R}$ is a direct summand, a basis for it consisting of l elements, say, can be extended to a basis for $\mathcal{A}^k$. Let K be the matrix whose rows are given by the $(k-l)$ elements that extend the basis for $\mathbf{R}$ above. Then the controller defined by the matrix K is such that the characteristic ideal of $\left[\begin{smallmatrix} R \\ K \end{smallmatrix}\right]$ is (u), where u is some unit in $\mathcal{A}$. Hence by replacing the last row of K by $u^{-1}r$ times that row (for some $r \in \mathcal{A}$), we can obtain a controller yielding the characteristic ideal (r).

Now let I be any ideal of $\mathcal{A}$, generated by $p_1, \dots, p_t$, say. Then we construct a controller K' from K as follows : K' has $k-l+t-1$ rows, where the first $k-l-1$ rows are the same as the first $k-l-1$ rows of K. The remaining t rows of K' are obtained by multiplying the last row of K by $u^{-1}p_1, \dots, u^{-1}p_t$, respectively.

The characteristic ideal of the augmented system $\left[\begin{smallmatrix} R \\ K' \end{smallmatrix}\right]$ is generated by the determinants of its $k \times k$ submatrices. If, in any $k \times k$ submatrix, more than $k-l$ rows of K' appear, then by construction the determinant is zero. Thus every nonzero determinant is obtained by choosing all the l rows of R and some $k-l$ rows of K'. Thus the characteristic ideal of the augmented system $\left[\begin{smallmatrix} R \\ K' \end{smallmatrix}\right]$ is precisely I. □

Observe from the proof of the above theorem that every principal ideal in $\mathcal{A}$ can be obtained as a characteristic ideal of a system augmented by a *free* controller. In [9], Willems shows that given a controllable system R in $\mathbb{R}^{l\times k}[\frac{d}{dx}]$ and any polynomial r in $\mathbb{R}[\frac{d}{dx}]$, there is a controller K in $\mathbb{R}^{(k-l)\times k}[\frac{d}{dx}]$ such that the characteristic polynomial of the augmented system $\left[\begin{smallmatrix} R \\ K \end{smallmatrix}\right]$ is (r). The theorem above is therefore the generalization of this result to distributed systems. For we have already shown that a lumped controllable system is given by a left prime matrix, i.e., by a submodule of $(\mathbb{R}[\frac{d}{dx}])^k$ which is a direct summand. See Proposition 3 and Corollary 3.

We therefore have solved the pole-placement problem for this special class of AR systems. We wish to extend these results to more general systems given by free submodules of $\mathcal{A}^k$.

We now address the question as to which principal ideals can be obtained using a free controller for systems defined by free submodules.

Let $\mathbf{R}$ be a free submodule of $\mathcal{A}^k$ of rank l and let R be any matrix representation for $\mathbf{R}$. Let $I_l(\mathbf{R})$ be its lth determinant ideal. Let $e_1, \dots, e_k$ be the (standard) basis of $\mathcal{A}^k$. Let $f_1, \dots, f_l$ in $\mathcal{A}^k$ freely generate the submodule $\mathbf{R}$. Consider the following map:

$$\phi : \Lambda^{k-l}\mathcal{A}^k \to \Lambda^k\mathcal{A}^k,$$

$$\sum g_1 \wedge \cdots \wedge g_{k-l} \mapsto \sum f_1 \wedge \cdots \wedge f_l \wedge g_1 \wedge \cdots \wedge g_{k-l}.$$

This is clearly an $\mathcal{A}$-module morphism. As $\Lambda^k \mathcal{A}^k$ is isomorphic to $\mathcal{A}$, the image of ϕ is therefore an ideal in $\mathcal{A}$. This image is of course spanned by the images of any basis for $\Lambda^{k-l}\mathcal{A}^k$.

Consider the following basis for $\Lambda^{k-l}\mathcal{A}^k$ obtained from the basis $e_1, \ldots, e_k$ for $\mathcal{A}^k$ by choosing a $(k-l)$-tuple from it, namely, elements of the form $e_{i_1} \wedge e_{i_2} \wedge \cdots \wedge e_{i_{k-l}}$ for $1 \leq i_1 < i_2 < \cdots < i_{k-l} \leq k$. The image of $e_{i_1} \wedge e_{i_2} \wedge \cdots \wedge e_{i_{k-l}}$ under ϕ is, up to sign, the determinant of the $l \times l$ submatrix of R obtained by choosing l columns of R which are *not* $i_1, i_2, \ldots, i_{k-l}$. However, this determinant is one of the generators of $I_l(\mathbf{R})$. Thus the image of the morphism ϕ is precisely $I_l(\mathbf{R})$. Hence if the controller is to be defined by a free submodule of rank $(k-l)$, then the characteristic ideal of the augmented system is a principal ideal that must lie in $I_l(\mathbf{R})$.

The question now is whether every principal ideal contained in $I_l(\mathbf{R})$ can be so obtained. To answer this observe that if K is a free controller of rank $(k-l)$, then the characteristic ideal of the augmented system $[\begin{smallmatrix} R \\ K \end{smallmatrix}]$ is a principal ideal generated by $f_1 \wedge \cdots \wedge f_l \wedge g_1 \wedge \cdots \wedge g_{k-l}$, where $g_1, \ldots, g_{k-l}$ are the rows of K. Thus these ideals come from the image under ϕ of homogeneous elements, denoted by $\mathcal{H}^{k-l}$, in $\Lambda^{k-l}\mathcal{A}^k$. Thus $\phi(\mathcal{H}^{k-l})$ correspond to the principal ideals in $I_l(\mathbf{R})$ which can be obtained by a free controller of rank $(k-l)$. It is also easy to see that $\phi(\mathcal{H}^{k-l})$ is in correspondence with free submodules of $\mathcal{A}^k$ of rank k, which contain $\mathbf{R}$ as a direct summand.

COROLLARY 10. *Let* $\mathbf{R}$ *be a free submodule of* $\mathcal{A}^k$ *of rank* $k-1$. *Then every principal ideal in* $I_{k-1}(\mathbf{R})$ *can be obtained by a free controller of rank* 1.

Proof. From the above discussion, observe that if l equals $k-1$, then $\mathcal{H}^{k-l} = \mathcal{H}^1$ is equal to all of $\Lambda^1\mathcal{A}^k$. Thus every element in $\Lambda^1\mathcal{A}^k$ is in $\mathcal{H}^1$ and so $\phi(\mathcal{H}^1)$ is all of $I_{k-1}(\mathbf{R})$. □

A constructive proof for the above corollary can also be given. Let R be a matrix representation of $\mathbf{R}$. Let (r) be any principal ideal contained in $I_{k-1}(\mathbf{R})$. Then $r = \sum a_i p_i$, where p_i is the determinant of the $(k-1) \times (k-1)$ submatrix of R obtained by dropping the ith column of R and $a_i \in \mathcal{A}$ for $i = 1, \ldots, k$. Then the controller given by $K = (a_1, -a_2, \ldots, (-1)^{i-1}a_i, \ldots, (-1)^{k-1}a_k)$ would give an augmented system that has (r) as its characteristic ideal.

COROLLARY 11. *Let* $\mathbf{R}$ *be a submodule of* $\mathcal{A}^k$ *and let* R *in* $\mathcal{A}^{l \times k}$ *be a matrix representation for it. Let* K *be a controller given by a submodule generated by* $g_1, \ldots, g_N$. *Suppose that every subset of* $\{g_1, \ldots, g_N\}$ *containing* $k-l+1$ *elements is* $\mathcal{A}$*-dependent. Then the characteristic ideal of the augmented system* $[\begin{smallmatrix} R \\ K \end{smallmatrix}]$ *is contained in* $I_l(\mathbf{R})$.

Proof. The proof is clear from the proof of Theorem 8 and the discussion following it. □

Remark. More generally in the above corollary, if every subset of $\{g_1, \ldots, g_N\}$ containing $k-l+t$ elements were $\mathcal{A}$-dependent, then the characteristic ideal of the augmented system would be contained in $I_{l-t+1}(\mathbf{R})$.

We have shown that a principal reason for the complexities of distributed behaviors as compared to lumped behaviors is due to the existence of submodules that are not free. We have, however, shown that in low projective dimensions (i.e., less than or equal to 2), the imposition of the property of controllability on a behavior forces the corresponding submodule to be free. This explains why controllable 2-D systems, even though not defined over a principal ideal domain, exhibit properties similar to controllable lumped systems. Our treatment therefore not only extends the results of Willems, and Rocha and Willems, on 1-D and discrete 2-D systems, but also provides a better understanding of this phenomena. Our treatment also shows why n-D systems for $n \geq 3$ are essentially different from 2-D systems. Our treatment of the pole-placement problem was essentially confined to behaviors given by free submodules. The pole-placement problem for general submodules will be considered elsewhere. Applications of this theory to specific distributed systems will also appear elsewhere [7].

Acknowledgments. We would like to thank the referees of an earlier version of this paper for their comments which led to significant improvements.

REFERENCES

[1] N. K. Bose, *Applied Multidimensional Systems Theory*, Van Nostrand Reinhold, New York, 1982.

[2] L. Hörmander, *The Analysis of Linear Partial Differential Operators,* I *&* II, Grundlehren Math. Wiss. 256 & 257, Springer-Verlag, Berlin, 1983.

[3] L. Hörmander, *An Introduction to Complex Analysis in Several Variables*, 3rd ed., North-Holland, Amsterdam, 1990.

[4] S. Lang, *Algebra*, 3rd ed., Addison-Wesley, Reading, MA, 1993.

[5] U. Oberst, *Multidimensional constant linear systems*, Acta Appl. Math., 20 (1990), pp. 1–175.

[6] P. Rocha and J. C. Willems, *Controllability of* 2*-D systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 413–423.

[7] S. Shankar, *Can One Control the Vibrations of a Drum?*, preprint, Indian Institute of Technology, Bombay, 1998.

[8] S. Shankar and V. Sule, *Algebraic geometric aspects of feedback stabilization*, SIAM J. Control Optim., 30 (1992), pp. 11–30.

[9] J. C. Willems, *Control as interconnection*, in Feedback Control, Nonlinear Systems and Complexity, Lecture Notes in Control and Inform. Sci. 202, B. A. Francis and A. R. Tannenbaum, eds., Springer-Verlag, 1995, pp. 261–275.

[10] J. C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.

[11] J. C. Willems, *The Behavioural Approach to Dynamical Systems and Control*, The Distinguished Lecture Series at IIT Bombay, 1995.

SIAM J. Control Optim.
Vol. 37, No. 2, pp. 409–428

$\mathcal{H}_\infty$ CONTROL OF DIFFUSION SYSTEMS BY USING A FINITE-DIMENSIONAL CONTROLLER[*]

HIDEKI SANO[†] AND YOSHIYUKI SAKAWA[‡]

Abstract. In this paper, we consider $\mathcal{H}_\infty$ control of linear and semilinear diffusion systems by using a finite-dimensional controller. The main aim is to construct a finite-dimensional stabilizing controller for the linear diffusion system that makes the $\mathcal{H}_\infty$ norm of the closed-loop transfer function less than a given positive number δ. For that purpose, we first derive a finite-dimensional reduced-order system for the linear diffusion system. Then, a stabilizing controller that makes the $\mathcal{H}_\infty$ norm of the closed-loop transfer function less than another positive number is constructed for the reduced-order model. It is proved that the finite-dimensional controller, together with a residual mode filter, plays a role of a finite-dimensional stabilizing controller that makes the $\mathcal{H}_\infty$ norm of the closed-loop transfer function less than δ for the original linear diffusion system if the order of residual mode filter is chosen sufficiently large. Moreover, it is shown that the finite-dimensional $\mathcal{H}_\infty$ controller constructed for the linear diffusion system also works as a finite-dimensional $\mathcal{H}_\infty$ controller for a semilinear diffusion system with sufficiently small nonlinear term.

Key words. $\mathcal{H}_\infty$ control, diffusion systems, residual mode filters

AMS subject classifications. 93B36, 93C25

PII. S0363012996302901

1. Introduction.

In recent years, $\mathcal{H}_\infty$ control theory for distributed parameter systems has been developed by extending the existing results for lumped parameter systems. In particular, $\mathcal{H}_\infty$ control with measurement feedback, which is of special importance from the practical point of view, has been investigated by Van Keulen [4]. In [4] the results given by Doyle et al. [3] are extended to the infinite-dimensional case by using the infinite-dimensional version of Redheffer's lemma developed by Van Keulen, and $\mathcal{H}_\infty$ controllers are constructed by using the solutions to two kinds of Riccati equations in an infinite-dimensional space. However, it seems difficult to apply Van Keulen's algorithm directly to actual infinite-dimensional systems, because the controllers obtained there are of infinite-dimension. In this paper, we study a design method of finite-dimensional $\mathcal{H}_\infty$ controllers for linear and semilinear diffusion systems. In our design method, we need not solve Riccati equations numerically in an infinite-dimensional space nor to reduce infinite-dimensional controllers to finite-dimensional ones by using approximation techniques. Instead, we use residual mode filters, which were originally adopted for the problem of stabilizing distributed parameter systems of modal type by using finite-dimensional dynamic compensators (see [6], [1], [7]).

In this paper, in order to construct a finite-dimensional controller that internally stabilizes the linear diffusion system and makes the $\mathcal{H}_\infty$ norm of the closed-loop transfer function from the disturbance input w to the controlled output z less than $\delta > 0$, we first derive a finite-dimensional model for the linear diffusion system. Next, we construct a stabilizing controller for the finite-dimensional model such that the $\mathcal{H}_\infty$

[*] Received by the editors May 1, 1996; accepted for publication (in revised form) February 16, 1998; published electronically November 23, 1998.

http://www.siam.org/journals/sicon/37-2/30290.html

[†] Department of Mathematics and Computer Science, Kagoshima University, 1-21-35 Korimoto, Kagoshima 890-0065, Japan (sano@sci.kagoshima-u.ac.jp).

[‡] Department of Intelligent Mechanics, Kinki University, 930 Nishimitani, Uchita-cho, Naga-gun, Wakayama 649-6493, Japan (sakawa@mec.waka.kindai.ac.jp).

norm of the closed-loop transfer function from w to z is less than γ $(0 < \gamma < \delta)$, by using the algorithm given by Doyle et al. [3]. However, the finite-dimensional controller constructed in this way is not necessarily a stabilizing controller for the original infinite-dimensional system that makes the $\mathcal{H}_\infty$ norm of the closed-loop transfer function from w to z less than δ. Therefore, we consider a controller consisting of the above controller and a residual mode filter. One of our main results in this paper is to prove that a controller, which consists of a residual mode filter and the finite-dimensional controller constructed for the finite-dimensional model, yields a finite-dimensional stabilizing controller such that the closed-loop transfer function from w to z has $\mathcal{H}_\infty$ norm less than δ for the given original infinite-dimensional system if the order of the residual mode filter is chosen sufficiently large. Moreover, it is shown that the finite-dimensional $\mathcal{H}_\infty$ controller constructed for the linear diffusion system also works as a finite-dimensional $\mathcal{H}_\infty$ controller for a semilinear diffusion system with sufficiently small Lipschitz constant.

2. $\mathcal{H}_\infty$ control of linear diffusion systems.

2.1. System description. Let H be a real Hilbert space with inner product $\langle\,\cdot\,,\,\cdot\,\rangle$ and associated norm $\|\cdot\|$. We shall consider the linear system of the form

$$(2.1)\qquad \begin{cases} \dfrac{dx(t)}{dt} = Ax(t) + B_1 w(t) + B_2 u(t) \quad (t > 0), \quad x(0) = x_0, \\ z(t) = C_1 x(t) + D_{12} u(t), \\ y(t) = C_2 x(t) + D_{21} w(t), \end{cases}$$

where $x(t) \in H$ is the state, $w(t) \in \mathbf{R}^{m_1}$ is the disturbance input, $u(t) \in \mathbf{R}^{m_2}$ is the control input, $z(t) \in \mathbf{R}^{p_1}$ is the controlled output, and $y(t) \in \mathbf{R}^{p_2}$ is the measured output. A is the infinitesimal generator of a C_0-semigroup e^{tA} on H. Moreover, we assume that A is self-adjoint and that A has compact resolvent. From these assumptions for the operator A, by using the Hilbert–Schmidt theory for compact self-adjoint operators, it follows that there exists a set of eigenpairs $(\lambda_i,\, \phi_{ij})$ with finite multiplicity μ_i, such that

(1) $\sigma(A) = \{\, \lambda_1,\, \lambda_2,\, \dots \}$, with $\lambda_1 > \lambda_2 > \cdots > \lambda_i > \cdots$, $\lim_{i\to\infty} \lambda_i = -\infty$,

(2) $A\phi_{ij} = \lambda_i \phi_{ij} \quad (i \geq 1,\, 1 \leq j \leq \mu_i)$,

(3) $\{\, \phi_{ij}\,;\, i \geq 1,\, 1 \leq j \leq \mu_i \,\}$ forms a complete orthonormal system in H,

where $\sigma(A)$ denotes the spectrum of A. Therefore, $x \in H$ has a unique expression

$$x = \sum_{i=1}^{\infty} \sum_{j=1}^{\mu_i} \langle x, \phi_{ij} \rangle \phi_{ij},$$

and the operator A is expressed as

$$Ax = \sum_{i=1}^{\infty} \sum_{j=1}^{\mu_i} \lambda_i \langle x, \phi_{ij} \rangle \phi_{ij} \quad \text{for } x \in D(A),$$

where

$$D(A) = \left\{ x \in H\,;\, \sum_{i=1}^{\infty} \sum_{j=1}^{\mu_i} \lambda_i^2 \langle x, \phi_{ij} \rangle^2 < \infty \right\}.$$

The C_0-semigroup e^{tA} generated by A is analytic in $t > 0$, and it is given as

$$e^{tA}x = \sum_{i=1}^{\infty} \sum_{j=1}^{\mu_i} e^{\lambda_i t} \langle x, \phi_{ij} \rangle \phi_{ij} \quad (t \geq 0,\, x \in H).$$

In (2.1), $B_1 : \mathbf{R}^{m_1} \to H$, $B_2 : \mathbf{R}^{m_2} \to H$, $C_1 : H \to \mathbf{R}^{p_1}$, $C_2 : H \to \mathbf{R}^{p_2}$ are bounded linear operators, and D_{12} is a $p_1 \times m_2$ matrix and D_{21} is a $p_2 \times m_1$ matrix.

In this paper, we impose the following assumptions for the operators and the matrices appearing in (2.1):

(A1) The pair (A, B_1) is stabilizable and the pair (C_1, A) is detectable, where the former means that there exists a bounded linear operator $F_1 : H \to \mathbf{R}^{m_1}$ such that the analytic semigroup generated by $A + B_1F_1$ is exponentially stable, and the latter means that there exists a bounded linear operator $K_1 : \mathbf{R}^{p_1} \to H$ such that the analytic semigroup generated by $A + K_1C_1$ is exponentially stable.

(A2) The pair (A, B_2) is stabilizable and the pair (C_2, A) is detectable.

(A3) $D_{12}^T C_1 = 0$, $D_{12}^T D_{12} = I$, where D_{12}^T denotes the transpose of D_{12}.

(A4) $B_1 D_{21}^T = 0$, $D_{21} D_{21}^T = I$.

Moreover, it is assumed that $x_0 \in H$. As is well known, the system (2.1) represents a linear diffusion system subjected to both the disturbance input and the control input.

2.2. $\mathcal{H}_\infty$control problem. In the system (2.1), if a proper control law $u = K(s)y$ is given, the closed-loop transfer function from w to z can be calculated, which will be denoted by $G_{zw}(s)$.

Given a positive number $\delta > 0$, the problem is to find a finite-dimensional control law $u = K(s)y$ such that the controller is internally stabilizing the system (2.1), and that $\|G_{zw}(\cdot)\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1},\mathbf{C}^{p_1}))} < \delta$, where

$$\|G_{zw}(\cdot)\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1},\mathbf{C}^{p_1}))} := \sup_{s \in \mathbf{C}^+} \|G_{zw}(s)\|_{\mathcal{L}(\mathbf{C}^{m_1},\mathbf{C}^{p_1})},$$
$$\mathbf{C}^+ := \{\, s \in \mathbf{C}\,;\, \mathrm{Re}(s) > 0 \,\}.$$

Any finite-dimensional controller that satisfies the above two conditions will be called a finite-dimensional $\mathcal{H}_\infty$ controller.

Remarks. (i) In this paper, we do not discuss how to find a finite-dimensional stabilizing controller that minimizes $\|G_{zw}(\cdot)\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1},\mathbf{C}^{p_1}))}$ under the order constraint. This problem is very important from the practical point of view, but it is difficult to solve.

(ii) In [4] Van Keulen has given a design method of infinite-dimensional controllers satisfying the above two properties for a large class of infinite-dimensional systems.

3. Finite-dimensional $\mathcal{H}_\infty$controllers for linear diffusion systems.

3.1. Partitioned system. First of all, let us define the positive number β as

$$\begin{aligned}\beta := \max\Big\{ &1,\ \|C_1\|\,\|B_2\|\,\|D_{12}^T\|,\ \|C_1\|\,\|B_1\|,\ \|D_{21}^T\|\,\|C_2\|\,\|B_1\|,\\ &\|C_1\|\,\|B_2\|\,\|D_{12}^T\|\,\|D_{21}^T\|\,\|C_2\|\,\|B_1\| \Big\}.\end{aligned} \tag{3.1}$$

Let $\delta > 0$ and $\theta \in (0, 1)$ be given positive numbers. To derive a finite-dimensional model for the system (2.1), we define the orthogonal projection P_K by

$$P_K x = \sum_{i=1}^{K} \sum_{j=1}^{\mu_i} \langle x, \phi_{ij} \rangle \phi_{ij}.$$

Using the operator P_K, we decompose the system (2.1) according to the following steps.

Step 1. Given a positive number $\epsilon \in (0, \frac{\theta\delta}{\beta})$, we choose an integer l (≥ 1) such that

$$0 < \frac{1}{-\lambda_{l+1}} < \epsilon.$$

Step 2. We take another integer n such that $n > l$.

Step 3. Using the operators P_l and P_n corresponding to the integers l and n, we decompose the state variable $x(t)$ as follows:

$$x(t) = x_1(t) + x_2(t) + x_3(t),$$

where $x_1(t) := P_l x(t)$, $x_2(t) := (P_n - P_l)x(t)$, $x_3(t) := (I - P_n)x(t)$. The state space H can also be decomposed as

$$H = H_1 \oplus H_2 \oplus H_3,$$

where $H_1 := P_l H$, $H_2 := (P_n - P_l)H$, $H_3 := (I - P_n)H$, and their dimensions are given as $\dim H_1 = \bar{l}$, $\dim H_2 = \bar{n} - \bar{l}$, $\dim H_3 = \infty$. In the above $\bar{l} := \mu_1 + \mu_2 + \cdots + \mu_l$, $\bar{n} := \mu_1 + \mu_2 + \cdots + \mu_n$. Therefore, we see that the system (2.1) can be decomposed as follows:

$$\begin{cases} \dfrac{dx_1(t)}{dt} = A_1 x_1(t) + B_{11} w(t) + B_{21} u(t), \quad x_1(0) = x_{01}, \\ \dfrac{dx_2(t)}{dt} = A_2 x_2(t) + B_{12} w(t) + B_{22} u(t), \quad x_2(0) = x_{02}, \\ \dfrac{dx_3(t)}{dt} = A_3 x_3(t) + B_{13} w(t) + B_{23} u(t), \quad x_3(0) = x_{03}, \\ z(t) = C_{11} x_1(t) + C_{12} x_2(t) + C_{13} x_3(t) + D_{12} u(t), \\ y(t) = C_{21} x_1(t) + C_{22} x_2(t) + C_{23} x_3(t) + D_{21} w(t), \end{cases}$$

where

$$\begin{cases} A_1 := P_l A P_l, \\ B_{11} := P_l B_1, \\ B_{21} := P_l B_2, \\ C_{11} := C_1 P_l, \\ C_{21} := C_2 P_l, \\ x_{01} := P_l x_0, \end{cases} \quad \begin{cases} A_2 := (P_n - P_l) A (P_n - P_l), \\ B_{12} := (P_n - P_l) B_1, \\ B_{22} := (P_n - P_l) B_2, \\ C_{12} := C_1 (P_n - P_l), \\ C_{22} := C_2 (P_n - P_l), \\ x_{02} := (P_n - P_l) x_0, \end{cases} \quad \begin{cases} A_3 := (I - P_n) A (I - P_n), \\ B_{13} := (I - P_n) B_1, \\ B_{23} := (I - P_n) B_2, \\ C_{13} := C_1 (I - P_n), \\ C_{23} := C_2 (I - P_n), \\ x_{03} := (I - P_n) x_0. \end{cases}$$

In the above, only the operator A_3 is unbounded, whereas the other operators are bounded.

Hereafter, the Hilbert space H_1 will be identified with the Euclidean space $\mathbf{R}^{\bar{l}}$ with respect to the basis $(\phi_{11}, \ldots, \phi_{1\mu_1}, \ldots, \phi_{l,\mu_l})$. Thus, $x_1 \in H_1$ will be identified with the $\bar{l}$-dimensional vector, and the operators A_1, B_{11}, B_{21}, C_{11}, C_{21} are identified with the matrices, respectively. In the same way, $x_2 \in H_2$ will be identified with the $(\bar{n} - \bar{l})$-dimensional vector, and the operators A_2, B_{12}, B_{22}, C_{12}, C_{22} are identified with the matrices, respectively.

3.2. Design of finite-dimensional $\mathcal{H}_\infty$controllers using residual mode filters. In connection with (2.1), let us consider a finite-dimensional model given by

$$\begin{cases} \dfrac{dx_1(t)}{dt} = A_1x_1(t) + B_{11}w(t) + B_{21}u(t), \\ z(t) = C_{11}x_1(t) + D_{12}u(t), \\ y(t) = C_{21}x_1(t) + D_{21}w(t). \end{cases} \tag{3.2}$$

Then, by [9, Proposition 4.12 and Proposition 4.13], it is easy to see that the assumptions (A1) and (A2) for the original infinite-dimensional system (2.1) imply that

(A1$'$) the pair (A_1, B_{11}) is stabilizable and the pair (C_{11}, A_1) is detectable, and

(A2$'$) the pair (A_1, B_{21}) is stabilizable and the pair (C_{21}, A_1) is detectable.

Moreover, noting that $C_{11} = C_1P_l$ and $B_{11} = P_lB_1$, we see that the assumptions (A3) and (A4) imply

(A3$'$) $D_{12}^TC_{11} = 0$, $D_{12}^TD_{12} = I$, and

(A4$'$) $B_{11}D_{21}^T = 0$, $D_{21}D_{21}^T = I$.

Given a proper control law $u = \tilde{K}(s)y$, we denote by $T_{zw}(s)$ the closed-loop transfer function from w to z for the finite-dimensional model (3.2). Here, for the given positive numbers $\delta > 0$, $\theta \in (0, 1)$, and $\epsilon \in (0, \frac{\theta\delta}{\beta})$, let us define γ by

$$\gamma := \frac{1}{(1+\epsilon)^2}\left(\frac{\theta\delta}{\beta} - \epsilon\right) \in (0, \delta). \tag{3.3}$$

We seek a controller $u = \tilde{K}(s)y$ that satisfies the following two conditions.

Condition 1. The controller is internally stabilizing the finite-dimensional system (3.2).

Condition 2. $\|T_{zw}(\cdot)\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1},\mathbf{C}^{p_1}))} < \gamma$.

From the results given by Doyle et al. [3], under the conditions (A1$'$)–(A4$'$) for the finite-dimensional system (3.2), there exists a controller $u = \tilde{K}(s)y$ that satisfies the above two conditions 1 and 2 if and only if, for the matrices H_∞ and J_∞ defined by

$$H_\infty := \begin{bmatrix} A_1 & \gamma^{-2}B_{11}B_{11}^T - B_{21}B_{21}^T \\ -C_{11}^TC_{11} & -A_1^T \end{bmatrix},$$

$$J_\infty := \begin{bmatrix} A_1^T & \gamma^{-2}C_{11}^TC_{11} - C_{21}^TC_{21} \\ -B_{11}B_{11}^T & -A_1 \end{bmatrix},$$

the following conditions hold:

(B1) $H_\infty \in \mathrm{dom(Ric)}$, $X_\infty := \mathrm{Ric}(H_\infty) \geq 0$,

(B2) $J_\infty \in \mathrm{dom(Ric)}$, $Y_\infty := \mathrm{Ric}(J_\infty) \geq 0$,

(B3) $\rho(X_\infty Y_\infty) < \gamma^2$,

where the notations dom(Ric) and Ric(H_∞) are as defined in [3], and $\rho(X_\infty Y_\infty)$ denotes the spectral radius of the matrix $X_\infty Y_\infty$.

Hereafter, it is supposed that the conditions (B1)–(B3) are satisfied. Then, by [3], all the stabilizing controllers that satisfy $\|T_{zw}(\cdot)\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1},\mathbf{C}^{p_1}))} < \gamma$ for the finite-dimensional model (3.2) are given as follows:

$$\begin{cases} \dfrac{dq(t)}{dt} = \hat{A}_\infty q(t) - Z_\infty L_\infty y(t) + Z_\infty B_{21}v(t), \quad q(0) = q_0, \\ u(t) = F_\infty q(t) + v(t), \\ r(t) = -C_{21}q(t) + y(t), \end{cases} \tag{3.4}$$

$$\text{(3.5)}\qquad \begin{cases} \dfrac{d\lambda(t)}{dt} = A_\Lambda \lambda(t) + B_\Lambda r(t), \quad \lambda(0) = \lambda_0, \\ v(t) = C_\Lambda \lambda(t) + D_\Lambda r(t), \end{cases}$$

where

$$\hat{A}_\infty := A_1 + \gamma^{-2} B_{11} B_{11}^T X_\infty + B_{21} F_\infty + Z_\infty L_\infty C_{21},$$
$$F_\infty := -B_{21}^T X_\infty, \quad L_\infty := -Y_\infty C_{21}^T, \quad Z_\infty := (I - \gamma^{-2} Y_\infty X_\infty)^{-1},$$

and the linear system $(A_\Lambda, B_\Lambda, C_\Lambda, D_\Lambda)$ indicates a free parameter such that

$$\begin{cases} \operatorname{Re}\sigma(A_\Lambda) < 0, \\ \|C_\Lambda((\cdot)I - A_\Lambda)^{-1} B_\Lambda + D_\Lambda\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{p_2}, \mathbf{C}^{m_2}))} < \gamma, \end{cases}$$

$\sigma(A_\Lambda)$ being the spectrum of A_Λ.

For simplicity, let us assume $D_\Lambda = 0$ in (3.5). Then, the closed-loop system consisting of the finite-dimensional model (3.2) and the controller (3.4), (3.5) is written as follows:

$$\begin{cases} \dfrac{d\eta(t)}{dt} = \mathcal{A}_1 \eta(t) + \mathcal{B}_1 w(t), \quad \eta(0) = \eta_0, \\ z(t) = \mathcal{C}_1 \eta(t), \end{cases}$$

where

$$\eta(t) := \begin{bmatrix} x_1(t) \\ p(t) \end{bmatrix} \quad \left(p(t) := \begin{bmatrix} q(t) \\ \lambda(t) \end{bmatrix} \right)$$

is in $\mathbf{R}^{\bar{l}} \times (\mathbf{R}^{\bar{l}} \times \mathbf{R}^S)$ ($\mathbf{R}^S$ denotes the state space of the free parameter (3.5)), and the matrices $\mathcal{A}_1$, $\mathcal{B}_1$, and $\mathcal{C}_1$ are defined by

$$\mathcal{A}_1 := \begin{bmatrix} A_1 & B_{21} L \\ N C_{21} & M \end{bmatrix}, \quad \mathcal{B}_1 := \begin{bmatrix} B_{11} \\ N D_{21} \end{bmatrix}, \quad \mathcal{C}_1 := \begin{bmatrix} C_{11} & D_{12} L \end{bmatrix},$$

$$M := \begin{bmatrix} \hat{A}_\infty & Z_\infty B_{21} C_\Lambda \\ -B_\Lambda C_{21} & A_\Lambda \end{bmatrix}, \quad N := \begin{bmatrix} -Z_\infty L_\infty \\ B_\Lambda \end{bmatrix}, \quad L := \begin{bmatrix} F_\infty & C_\Lambda \end{bmatrix}.$$

The condition 1 that the controller $\tilde{K}(s)$ is internally stabilizing the system means that the matrix $e^{t\mathcal{A}_1}$ is exponentially stable, i.e., there exist positive constants $m\,(\geq 1)$ and α such that

$$\text{(3.6)}\qquad \|e^{t\mathcal{A}_1}\| \leq m e^{-\alpha t} \quad (t \geq 0).$$

It should be noted that both m and α do not depend on the integer n. Also, noting that $T_{zw}(s) = \mathcal{C}_1 (sI - \mathcal{A}_1)^{-1} \mathcal{B}_1$, we see that Condition 2 is equivalent to the inequality

$$\text{(3.7)}\qquad \|\mathcal{C}_1((\cdot)I - \mathcal{A}_1)^{-1} \mathcal{B}_1\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1}, \mathbf{C}^{p_1}))} < \gamma.$$

However, it cannot be ensured that the controller (3.4), (3.5) works as a finite-dimensional stabilizing controller that satisfies $\|G_{zw}(\cdot)\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1}, \mathbf{C}^{p_1}))} < \delta$ for the given original infinite-dimensional system (2.1).

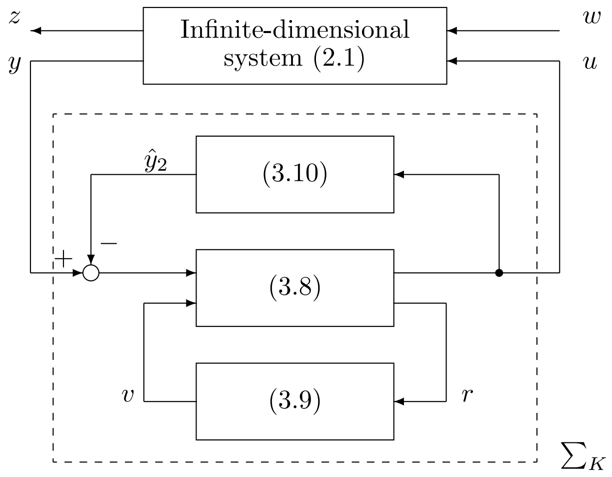


FIG. 1. *Finite-dimensional $\mathcal{H}_\infty$ controller Σ_K.*

Let us consider a controller, which consists of the controller (3.4), (3.5) with $D_\Lambda = 0$ constructed for (3.2) as well as a residual mode filter

$$\begin{cases} \dfrac{d\hat{x}_2(t)}{dt} = A_2\hat{x}_2(t) + B_{22}u(t), \quad \hat{x}_2(0) = \hat{x}_{20}, \\ \hat{y}_2(t) = C_{22}\hat{x}_2(t). \end{cases}$$

Then, the whole controller Σ_K (see Fig. 1) is described by

$$\text{(3.8)} \quad \begin{cases} \dfrac{dq(t)}{dt} = \hat{A}_\infty q(t) - Z_\infty L_\infty(y(t) - \hat{y}_2(t)) + Z_\infty B_{21}v(t), \quad q(0) = q_0, \\ u(t) = F_\infty q(t) + v(t), \\ r(t) = -C_{21}q(t) + y(t) - \hat{y}_2(t), \end{cases}$$

$$\text{(3.9)} \quad \begin{cases} \dfrac{d\lambda(t)}{dt} = A_\Lambda \lambda(t) + B_\Lambda r(t), \quad \lambda(0) = \lambda_0, \\ v(t) = C_\Lambda \lambda(t), \end{cases}$$

$$\text{(3.10)} \quad \begin{cases} \dfrac{d\hat{x}_2(t)}{dt} = A_2\hat{x}_2(t) + B_{22}u(t), \quad \hat{x}_2(0) = \hat{x}_{20}, \\ \hat{y}_2(t) = C_{22}\hat{x}_2(t). \end{cases}$$

The following theorem is one of our main results.

THEOREM 3.1. *Suppose that the assumptions* (A1)–(A4) *are satisfied. Let $\delta > 0$, $\theta \in (0,1)$, and $\epsilon \in (0, \frac{\theta\delta}{\beta})$ be given constants, where the constant β is defined by* (3.1), *and let the integers l and n be chosen such that $0 < \frac{1}{-\lambda_{l+1}} < \epsilon$ and $n > l$ hold, where λ_{l+1} is the $(l+1)$th eigenvalue of the operator A. Moreover, let the conditions* (B1)–(B3) *be satisfied with $\gamma = \frac{1}{(1+\epsilon)^2}(\frac{\theta\delta}{\beta} - \epsilon) \in (0, \delta)$. Then, the controller Σ_K, which consists of* (3.8), (3.9), *and* (3.10), *gives a finite-dimensional stabilizing controller that satisfies $\|G_{zw}(\cdot)\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1},\mathbf{C}^{p_1}))} < \delta$ for the linear infinite-dimensional system* (2.1) *if the integer n is taken sufficiently large.*

4. Proof of Theorem 3.1. First of all, let us state the lemma which will be frequently used in this section.

LEMMA 4.1 (Lemma 4.1 in [8]). *Let A_{11} and A_{22} be the generators of C_0-semigroups $S_1(t)$, $S_2(t)$ with $\|S_1(t)\|_{\mathcal{L}(X_1)} \leq M_1 e^{\omega_1 t}$, $\|S_2(t)\|_{\mathcal{L}(X_2)} \leq M_2 e^{\omega_2 t}$ $(\omega_1 \neq \omega_2)$ for $t \geq 0$ on Hilbert spaces X_1, X_2, respectively, and let $A_{12} : X_2 \to X_1$ and $A_{21} : X_1 \to X_2$ be bounded linear operators. Then, the C_0-semigroup $\bar{S}_{21}(t)$ on $X_1 \times X_2$ generated by the operator $\left[\begin{smallmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{smallmatrix}\right]$ and the C_0-semigroup $\bar{S}_{12}(t)$ on $X_1 \times X_2$ generated by the operator $\left[\begin{smallmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{smallmatrix}\right]$ have the following operator norm bounds:*

$$\|\bar{S}_{21}(t)\|_{\mathcal{L}(X_1 \times X_2)} \leq \max(M_1, M_2)\left(1 + \frac{\max(M_1, M_2)\|A_{21}\|_{\mathcal{L}(X_1, X_2)}}{|\omega_1 - \omega_2|}\right) e^{\max(\omega_1, \omega_2)t} \quad \textit{for} \quad t \geq 0,$$

$$\|\bar{S}_{12}(t)\|_{\mathcal{L}(X_1 \times X_2)} \leq \max(M_1, M_2)\left(1 + \frac{\max(M_1, M_2)\|A_{12}\|_{\mathcal{L}(X_2, X_1)}}{|\omega_1 - \omega_2|}\right) e^{\max(\omega_1, \omega_2)t} \quad \textit{for} \quad t \geq 0.$$

Let us introduce the variable $e_2(t) := x_2(t) - \hat{x}_2(t)$. Then, the closed-loop system of state equations from w to z in Fig. 1 is written as follows:

$$\begin{cases} \dfrac{d\xi(t)}{dt} = (\mathcal{A} + \Delta\mathcal{A})\xi(t) + (\mathcal{B} + \Delta\mathcal{B})w(t), \quad \xi(0) = \xi_0, \\ z(t) = (\mathcal{C} + \Delta\mathcal{C})\xi(t), \end{cases}$$

where the state vector

$$\xi(t) := \begin{bmatrix} x_1(t) \\ p(t) \\ x_2(t) \\ x_3(t) \\ e_2(t) \end{bmatrix} \quad \left(p(t) := \begin{bmatrix} q(t) \\ \lambda(t) \end{bmatrix}\right)$$

is in a real Hilbert space $X := \mathbf{R}^{\bar{l}} \times (\mathbf{R}^{\bar{l}} \times \mathbf{R}^{S}) \times \mathbf{R}^{\bar{n}-\bar{l}} \times H_3 \times \mathbf{R}^{\bar{n}-\bar{l}}$ with the inner product $\langle \cdot, \cdot \rangle_X$ defined by

$$\langle \xi, \tilde{\xi} \rangle_X := x_1^T \tilde{x}_1 + p^T \tilde{p} + x_2^T \tilde{x}_2 + \langle x_3, \tilde{x}_3 \rangle + e_2^T \tilde{e}_2$$
$$\text{for} \quad \xi = \begin{bmatrix} x_1 \\ p \\ x_2 \\ x_3 \\ e_2 \end{bmatrix}, \ \tilde{\xi} = \begin{bmatrix} \tilde{x}_1 \\ \tilde{p} \\ \tilde{x}_2 \\ \tilde{x}_3 \\ \tilde{e}_2 \end{bmatrix} \in X,$$

and the operators $\mathcal{A}$, $\Delta\mathcal{A}$, $\mathcal{B}$, $\Delta\mathcal{B}$, $\mathcal{C}$, and $\Delta\mathcal{C}$ are defined by

$$\mathcal{A} := \begin{bmatrix} A_1 & B_{21}L & 0 & 0 & 0 \\ NC_{21} & M & 0 & 0 & NC_{22} \\ 0 & B_{22}L & A_2 & 0 & 0 \\ 0 & 0 & 0 & A_3 & 0 \\ 0 & 0 & 0 & 0 & A_2 \end{bmatrix}, \quad \Delta\mathcal{A} := \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & NC_{23} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & B_{23}L & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\mathcal{B} := \begin{bmatrix} B_{11} \\ ND_{21} \\ B_{12} \\ 0 \\ B_{12} \end{bmatrix}, \quad \Delta\mathcal{B} := \begin{bmatrix} 0 \\ 0 \\ 0 \\ B_{13} \\ 0 \end{bmatrix},$$

$$\mathcal{C} := \begin{bmatrix} C_{11} & D_{12}L & C_{12} & 0 & 0 \end{bmatrix}, \quad \Delta\mathcal{C} := \begin{bmatrix} 0 & 0 & 0 & C_{13} & 0 \end{bmatrix},$$

$$M := \begin{bmatrix} \hat{A}_\infty & Z_\infty B_{21} C_\Lambda \\ -B_\Lambda C_{21} & A_\Lambda \end{bmatrix}, \quad N := \begin{bmatrix} -Z_\infty L_\infty \\ B_\Lambda \end{bmatrix}, \quad L := \begin{bmatrix} F_\infty & C_\Lambda \end{bmatrix}.$$

In the above, we remark that the operator $\mathcal{A}$ is unbounded because A_3 is unbounded, whereas the operators $\Delta\mathcal{A}$, $\mathcal{B}$, $\Delta\mathcal{B}$, $\mathcal{C}$, and $\Delta\mathcal{C}$ are bounded.

4.1. Exponential stability. Define the matrix $\mathcal{A}_2$ by

$$\mathcal{A}_2 := \begin{bmatrix} \mathcal{A}_1 & 0 \\ \begin{bmatrix} 0 & B_{22}L \end{bmatrix} & A_2 \end{bmatrix}.$$

Then, the operator $\mathcal{A}$ can be written as

$$\mathcal{A} = \begin{bmatrix} \mathcal{A}_2 & \begin{bmatrix} 0 & 0 \\ 0 & NC_{22} \\ 0 & 0 \end{bmatrix} \\ 0 & \begin{bmatrix} A_3 & 0 \\ 0 & A_2 \end{bmatrix} \end{bmatrix}.$$

First, we shall estimate the norm of the matrix $e^{t\mathcal{A}_2}$. For the matrix $e^{t\mathcal{A}_1}$, we know that under the assumptions of the theorem, inequality (3.6) holds. For a matrix e^{tA_2}, the following estimate holds:

$$\|e^{tA_2}\| = e^{\lambda_{l+1}t} \leq e^{-\omega t} \quad (t \geq 0), \tag{4.1}$$

where $\omega > 0$ is independent of n and is chosen such that $\lambda_{l+1} \leq -\omega$, $\alpha \neq \omega$. Also, we can easily derive the inequality

$$\left\| \begin{bmatrix} 0 & B_{22}L \end{bmatrix} \right\| \leq N_1, \tag{4.2}$$

where $N_1 := \|B_2\|\|L\|$ is independent of n. Therefore, by using Lemma 4.1 for $\mathcal{A}_2$ and by using (3.6), (4.1), and (4.2), it follows that

$$\|e^{t\mathcal{A}_2}\| \leq m' e^{-\sigma t} \quad (t \geq 0), \tag{4.3}$$

where

$$m' := m\left(1 + \frac{mN_1}{|\alpha - \omega|}\right) \ (\geq 1), \quad \sigma := \min(\alpha, \omega) \ (> 0)$$

are independent of n by Lemma 4.1.

Next, based on the above estimate (4.3), we shall find the operator norm bound of an analytic semigroup $e^{t\mathcal{A}}$ generated by the operator $\mathcal{A}$. For an analytic semigroup $\begin{bmatrix} e^{tA_3} & 0 \\ 0 & e^{tA_2} \end{bmatrix}$ generated by the operator $\begin{bmatrix} A_3 & 0 \\ 0 & A_2 \end{bmatrix}$, it is clear that

$$\left\| \begin{bmatrix} e^{tA_3} & 0 \\ 0 & e^{tA_2} \end{bmatrix} \right\|_{\mathcal{L}(H_3 \times \mathbf{R}^{\bar{n}-\bar{l}})} \leq e^{\lambda_{l+1}t} \quad (t \geq 0). \tag{4.4}$$

Also, it is easy to see that

$$\left\| \begin{bmatrix} 0 & 0 \\ 0 & NC_{22} \\ 0 & 0 \end{bmatrix} \right\|_{\mathcal{L}(H_3 \times \mathbf{R}^{\bar{n}-\bar{l}}, \mathbf{R}^{\bar{l}} \times (\mathbf{R}^{\bar{l}} \times \mathbf{R}^{S}) \times \mathbf{R}^{\bar{n}-\bar{l}})} \leq N_2, \tag{4.5}$$

where $N_2 := \|N\| \|C_2\|$ is independent of n. Therefore, by using Lemma 4.1 and (4.3), (4.4), and (4.5), it follows that

$$\|e^{t\mathcal{A}}\|_{\mathcal{L}(X)} \leq m'' e^{-\sigma t} \quad (t \geq 0), \tag{4.6}$$

where

$$m'' := m' \left(1 + \frac{m' N_2}{|\sigma + \lambda_{l+1}|} \right) (\geq 1)$$

is independent of n.

By using the well-known result (e.g. [5, Theorem 3.1.1], [11, Theorem 3.4.1]), for an analytic semigroup $e^{t(\mathcal{A}+\Delta\mathcal{A})}$ generated by the perturbed operator $\mathcal{A} + \Delta\mathcal{A}$, where $\Delta\mathcal{A}$ is bounded, we get

$$\|e^{t(\mathcal{A}+\Delta\mathcal{A})}\|_{\mathcal{L}(X)} \leq m'' e^{-\mu t} \quad (t \geq 0), \tag{4.7}$$

where

$$\mu := \sigma - m'' \|\Delta\mathcal{A}\|_{\mathcal{L}(X)}.$$

Therefore, noting that $\|\Delta\mathcal{A}\|_{\mathcal{L}(X)} \leq \|B_{23}\| \|L\| + \|N\| \|C_{23}\| \to 0$ $(n \to \infty)$ holds since $\|B_{23}\| \to 0$ $(n \to \infty)$ and $\|C_{23}\| \to 0$ $(n \to \infty)$, it follows that there exists an integer n_1 such that

$$\mu = \sigma - m'' \|\Delta\mathcal{A}\|_{\mathcal{L}(X)} > 0 \quad (\forall\, n \geq n_1),$$

i.e., $e^{t(\mathcal{A}+\Delta\mathcal{A})}$ is exponentially stable when n is chosen such that $n \geq n_1$. □

4.2. Norm condition. First, let us estimate the $\mathcal{H}_\infty$ norm of $G(s) := \mathcal{C}(sI - \mathcal{A})^{-1}\mathcal{B}$. Using the assumptions (A3) and (A4) (i.e., (A3′) and (A4′)), $G(s)$ is calculated as follows (see appendix):

$$\begin{aligned} G(s) = {} & T_{zw}(s) + C_{12}(sI - A_2)^{-1} B_{22} D_{12}^T T_{zw}(s) + C_{12}(sI - A_2)^{-1} B_{12} \\ & + T_{zw}(s) D_{21}^T C_{22}(sI - A_2)^{-1} B_{12} \\ & + C_{12}(sI - A_2)^{-1} B_{22} D_{12}^T T_{zw}(s) D_{21}^T C_{22} (sI - A_2)^{-1} B_{12}, \end{aligned} \tag{4.8}$$

where

$$T_{zw}(s) = \mathcal{C}_1 (sI - \mathcal{A}_1)^{-1} \mathcal{B}_1 = \begin{bmatrix} C_{11} & D_{12}L \end{bmatrix} \begin{bmatrix} sI - A_1 & -B_{21}L \\ -NC_{21} & sI - M \end{bmatrix}^{-1} \begin{bmatrix} B_{11} \\ ND_{21} \end{bmatrix}.$$

Under the assumptions of the theorem, we see that

$$\|T_{zw}(\cdot)\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1}, \mathbf{C}^{p_1}))} < \gamma.$$

Moreover, noting the fact that

$$\|((\cdot)I - A_2)^{-1}\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{\bar{n}-\bar{l}}))} = \frac{1}{-\lambda_{l+1}} < \epsilon,$$

we can give the $\mathcal{H}_\infty$ norm bound of $G(s)$ as

$$\|G(\cdot)\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1},\mathbf{C}^{p_1}))} \leq \beta \left[(1+\epsilon)^2 \|T_{zw}(\cdot)\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1},\mathbf{C}^{p_1}))} + \epsilon\right] < \theta\delta.$$

Based on this fact, we shall find the $\mathcal{H}_\infty$ norm bound of

$$G_{zw}(s) = (\mathcal{C} + \varDelta\mathcal{C})\big(sI - (\mathcal{A} + \varDelta\mathcal{A})\big)^{-1}(\mathcal{B} + \varDelta\mathcal{B}).$$

Let us denote by X^c the complexification of X. First, noting that

$$\|((\cdot)I - \mathcal{A})^{-1}\|_{\mathcal{H}_\infty(\mathcal{L}(X^c))} \leq m''/\sigma$$

(by [5, Theorem 1.5.3 and Remark 1.5.4], together with the definition of $\mathcal{H}_\infty$ norm) and that $\|\varDelta\mathcal{A}\|_{\mathcal{L}(X^c)} = \|\varDelta\mathcal{A}\|_{\mathcal{L}(X)} \to 0$ $(n \to \infty)$, it follows that

$$\|((\cdot)I - \mathcal{A})^{-1}\varDelta\mathcal{A}\|_{\mathcal{H}_\infty(\mathcal{L}(X^c))} \to 0 \quad (n \to \infty),$$

which implies that there exists an integer n_2 such that

$$\|((\cdot)I - \mathcal{A})^{-1}\varDelta\mathcal{A}\|_{\mathcal{H}_\infty(\mathcal{L}(X^c))} < 1 \quad (\forall\, n \geq n_2).$$

Therefore, when n is chosen such that $n \geq n_2$, $(I - ((\cdot)I - \mathcal{A})^{-1}\varDelta\mathcal{A})^{-1} \in \mathcal{H}_\infty(\mathcal{L}(X^c))$ exists and the $\mathcal{H}_\infty$ norm can be estimated as follows:

$$\|(I - ((\cdot)I - \mathcal{A})^{-1}\varDelta\mathcal{A})^{-1}\|_{\mathcal{H}_\infty(\mathcal{L}(X^c))} \leq \frac{1}{1 - \|((\cdot)I - \mathcal{A})^{-1}\varDelta\mathcal{A}\|_{\mathcal{H}_\infty(\mathcal{L}(X^c))}}.$$

Here, setting $n_3 := \max(n_1, n_2)$, we can calculate

$$G_{zw}(s) = (\mathcal{C}+\varDelta\mathcal{C}) \Big[(sI - \mathcal{A})^{-1} + (sI - \mathcal{A})^{-1}\varDelta\mathcal{A}\big(I - (sI - \mathcal{A})^{-1}\varDelta\mathcal{A}\big)^{-1}(sI - \mathcal{A})^{-1}\Big] \times(\mathcal{B} + \varDelta\mathcal{B}), \quad s \in \mathbf{C}^+,$$

if n is chosen such that $n \geq n_3$. Moreover, noting that $\|\varDelta\mathcal{B}\|_{\mathcal{L}(\mathbf{C}^{m_1},X^c)} = \|\varDelta\mathcal{B}\|_{\mathcal{L}(\mathbf{R}^{m_1},X)} = \|B_{13}\| \to 0$ $(n \to \infty)$ and $\|\varDelta\mathcal{C}\|_{\mathcal{L}(X^c,\mathbf{C}^{p_1})} = \|\varDelta\mathcal{C}\|_{\mathcal{L}(X,\mathbf{R}^{p_1})} = \|C_{13}\| \to 0$ $(n \to \infty)$, we have

$$\|G_{zw}(\cdot) - G(\cdot)\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1},\mathbf{C}^{p_1}))} \to 0 \quad (n \to \infty), \tag{4.9}$$

which implies that there exists an integer n_4 $(\geq n_3)$ such that

$$\|G_{zw}(\cdot) - G(\cdot)\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1},\mathbf{C}^{p_1}))} < (1-\theta)\delta \quad (\forall\, n \geq n_4).$$

Hence, we finally obtain

$$\begin{aligned} \|G_{zw}(\cdot)\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1},\mathbf{C}^{p_1}))} &\leq \|G_{zw}(\cdot) - G(\cdot)\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1},\mathbf{C}^{p_1}))} + \|G(\cdot)\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1},\mathbf{C}^{p_1}))} \\ &< (1-\theta)\delta + \theta\delta = \delta \quad (\forall\, n \geq n_4). \qquad \square \end{aligned} \tag{4.10}$$

The proof of Theorem 3.1 is thus complete.

Remarks. (i) We cannot a priori estimate the order of the resulting controller $\bar{n}+S$ in Theorem 3.1 as well as the order of finite-dimensional dynamic compensators (see [6], [1], [7]). This is an open problem.

(ii) In Theorem 3.1, we can replace the operator A by a modal operator [1]. An unbounded closed linear operator A with the domain $D(A)$ dense in H is called a modal operator if A has the form

$$Ax = \sum_{i=1}^{\infty} \lambda_i \langle x, \phi_i \rangle \phi_i, \quad x \in D(A),$$

where λ_i are eigenvalues of A satisfying

$$\mathrm{Re}(\lambda_1) \geq \mathrm{Re}(\lambda_2) \geq \cdots \geq \mathrm{Re}(\lambda_i) \geq \cdots, \lim_{i\to\infty} \mathrm{Re}(\lambda_i) = -\infty,$$

and ϕ_i denote eigenfunctions corresponding to λ_i, and $\{\phi_i\,;\, i \geq 1\}$ forms a complete orthonormal system in H.

5. Example. Let $\Omega = (0, \pi)$ be a domain in $\mathbf{R}^1$ with a boundary $\Gamma = \{0, \pi\}$. We consider the linear diffusion system described by

$$(5.1) \quad \begin{cases} \dfrac{\partial x}{\partial t}(t,\xi) = \triangle x(t,\xi) + b_1(\xi)w_1(t) + b_2(\xi)u(t), \quad (t,\xi) \in (0,\infty)\times\Omega, \\ \dfrac{\partial x}{\partial \nu}(t,\eta) = 0, \quad (t,\eta) \in (0,\infty)\times\Gamma, \\ x(0,\xi) = x_0(\xi), \quad \xi \in \Omega, \\ z_1(t) = \displaystyle\int_\Omega c_1(\xi)x(t,\xi)d\xi, \\ z_2(t) = u(t), \\ y(t) = \displaystyle\int_\Omega c_2(\xi)x(t,\xi)d\xi + w_2(t). \end{cases}$$

In the above, $\triangle$ denotes the Laplacian, and $\partial/\partial\nu$ means the outward normal differentiation at the points $\eta = 0, \pi$.

We will formulate the above system in a real Hilbert space $L^2(\Omega)$ with inner product $\langle\,\cdot\,,\,\cdot\,\rangle_{L^2(\Omega)}$ and corresponding norm $\|\cdot\|_{L^2(\Omega)}$. First, we define the operator A in $L^2(\Omega)$ as follows:

$$Ax = \triangle x, \quad x \in D(A) = \left\{ x \in H^2(\Omega)\,;\, \frac{\partial x}{\partial \nu} = 0 \text{ on } \Gamma \right\}.$$

Then, A is a self-adjoint operator with compact resolvent and the eigenpairs of A are completely specified by

$$\begin{cases} \lambda_i = -(i-1)^2 \quad (i = 1, 2, 3, \ldots), \\ \phi_1(\xi) = \dfrac{1}{\sqrt{\pi}}, \quad \phi_i(\xi) = \sqrt{\dfrac{2}{\pi}}\cos(i-1)\xi \quad (i = 2, 3, \ldots). \end{cases}$$

Here, assuming that $b_1, b_2, c_1, c_2 \in L^2(\Omega)$, and defining bounded linear operators $\tilde{B}_1 : \mathbf{R}^1 \to L^2(\Omega)$, $B_2 : \mathbf{R}^1 \to L^2(\Omega)$, $\tilde{C}_1 : L^2(\Omega) \to \mathbf{R}^1$, and $C_2 : L^2(\Omega) \to \mathbf{R}^1$ by

$$\tilde{B}_1 w_1(t) := b_1 w_1(t), \quad B_2 u(t) := b_2 u(t),$$

$$\tilde{C}_1 x(t) := \langle c_1, x(t) \rangle_{L^2(\Omega)}, \quad C_2 x(t) := \langle c_2, x(t) \rangle_{L^2(\Omega)},$$

we can express the above system (5.1) as

$$\text{(5.2)}\qquad \begin{cases} \dfrac{dx(t)}{dt} = Ax(t) + \tilde{B}_1 w_1(t) + B_2 u(t), \quad x(0) = x_0, \\ z_1(t) = \tilde{C}_1 x(t), \\ z_2(t) = u(t), \\ y(t) = C_2 x(t) + w_2(t). \end{cases}$$

Moreover, defining the operators $B_1 : \mathbf{R}^2 \to L^2(\Omega)$, $C_1 : L^2(\Omega) \to \mathbf{R}^2$, the 2×1 matrix D_{12}, the 1×2 matrix D_{21}, and the vectors $w(t) \in \mathbf{R}^2$, $z(t) \in \mathbf{R}^2$ by

$$B_1 := \begin{bmatrix} \tilde{B}_1 & 0 \end{bmatrix}, \quad C_1 := \begin{bmatrix} \tilde{C}_1 \\ 0 \end{bmatrix}, \quad D_{12} := \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad D_{21} := \begin{bmatrix} 0 & 1 \end{bmatrix},$$

$$w(t) := \begin{bmatrix} w_1(t) \\ w_2(t) \end{bmatrix}, \quad z(t) := \begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix},$$

the above system (5.2) can be expressed as in (2.1).

Now, we especially consider the case where $b_j \in L^2(\Omega)$ $(j = 1,2)$ and $c_j \in L^2(\Omega)$ $(j = 1,2)$ are given by

$$b_j(\xi) = \begin{cases} \dfrac{1}{\sqrt{(\beta_j - \alpha_j)\pi}}, \ \xi \in [\alpha_j\pi, \beta_j\pi], \\ 0, \quad \xi \in (0, \alpha_j\pi) \cup (\beta_j\pi, \pi), \end{cases} \qquad c_j(\xi) = \begin{cases} \dfrac{1}{\sqrt{(\delta_j - \gamma_j)\pi}}, \ \xi \in [\gamma_j\pi, \delta_j\pi], \\ 0, \quad \xi \in (0, \gamma_j\pi) \cup (\delta_j\pi, \pi), \end{cases}$$

where $\alpha_1 = 0.05$, $\beta_1 = 0.25$, $\alpha_2 = 0.5$, $\beta_2 = 0.75$, $\gamma_1 = 0.3$, $\delta_1 = 0.45$, $\gamma_2 = 0.8$, $\delta_2 = 0.95$. Then, noting that

$$\begin{aligned} &\|B_1\| = \|b_1\|_{L^2(\Omega)} = 1, \quad \|B_2\| = \|b_2\|_{L^2(\Omega)} = 1, \\ &\|C_1\| = \|c_1\|_{L^2(\Omega)} = 1, \quad \|C_2\| = \|c_2\|_{L^2(\Omega)} = 1, \\ &\|D_{12}^T\| = \|D_{12}\| = 1, \quad \|D_{21}^T\| = \|D_{21}\| = 1, \end{aligned}$$

it follows from (3.1) that $\beta = 1$.

Here, let us set $\delta = 1.85$ and $\theta = 0.96479$ (< 1). Then, we can choose ϵ as $\epsilon = 0.063$ $(< \frac{\theta\delta}{\beta} = 1.78486)$. Moreover, we can choose $l = 4$, and calculate the matrices of A_1, B_{11}, B_{21}, C_{11}, C_{21} with respect to the basis $(\phi_1, \phi_2, \phi_3, \phi_4)$ as follows:

$$A_1 = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{bmatrix}, \quad B_{11} = \begin{bmatrix} g_{1,1} & 0 \\ g_{1,2} & 0 \\ g_{1,3} & 0 \\ g_{1,4} & 0 \end{bmatrix}, \quad B_{21} = \begin{bmatrix} g_{2,1} \\ g_{2,2} \\ g_{2,3} \\ g_{2,4} \end{bmatrix},$$

$$C_{11} = \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} & h_{1,4} \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad C_{21} = \begin{bmatrix} h_{2,1} & h_{2,2} & h_{2,3} & h_{2,4} \end{bmatrix},$$

where

$$\begin{aligned} g_{j,i} &:= \langle b_j, \phi_i \rangle_{L^2(\Omega)} \quad (j = 1,2\,;\, i = 1,2,3,4), \\ h_{j,i} &:= \langle c_j, \phi_i \rangle_{L^2(\Omega)} \quad (j = 1,2\,;\, i = 1,2,3,4). \end{aligned}$$

In the above, A_1, B_{11}, B_{21}, C_{11}, and C_{21} are identified with their corresponding matrices. It is easily verified that the conditions (A3′) and (A4′) are satisfied. From

(3.3), we can calculate $\gamma = 1.52381$. Then, by using Robust Control TOOLBOX of MATLAB [2], we can verify that the conditions (A1′), (A2′), and (B1)–(B3) are satisfied.

Remark. In this example, we treated the system with $\beta = 1$. If β is larger, we see from (3.3) that the $\mathcal{H}_\infty$ norm bound γ for the finite-dimensional model becomes smaller. Consequently, if β is larger, conditions (B1)–(B3) may not be satisfied.

6. $\mathcal{H}_\infty$control of semilinear diffusion systems. Let H be a real Hilbert space with inner product $\langle\,\cdot\,,\,\cdot\,\rangle$ and associated norm $\|\cdot\|$. Consider the semilinear system of the form

$$\text{(6.1)}\qquad \begin{cases} \dfrac{dx(t)}{dt} = Ax(t) + F(x(t)) + B_1 w(t) + B_2 u(t) \quad (t > 0), \quad x(0) = x_0, \\ z(t) = C_1 x(t) + D_{12} u(t), \\ y(t) = C_2 x(t) + D_{21} w(t), \end{cases}$$

where the nonlinear term $F(\cdot) : H \to H$ satisfies the following condition:

$$\text{(6.2)}\qquad \begin{cases} \|F(x) - F(\tilde{x})\| \le k\|x - \tilde{x}\|, \quad x, \tilde{x} \in H, \\ F(0) = 0. \end{cases}$$

The other operators and matrices are the same ones as in section 2.

A function $x(\cdot) : [0,\infty) \to H$ is called a solution of the first equation of (6.1) if $x(\cdot) \in C([0,\infty);H) \cap C^1((0,\infty);H)$, $x(t) \in D(A)$ for $t > 0$, and $x(\cdot)$ satisfies the first equation of (6.1). For the first equation of (6.1), the following existence and uniqueness result is obtained by using the theory of semilinear evolution equations (see [5], [10]).

LEMMA 6.1. *Let $x_0 \in H$, $w(\cdot) \in C_\nu([0,\infty);\mathbf{R}^{m_1})$, and $u(\cdot) \in C_\nu([0,\infty);\mathbf{R}^{m_2})$ $(0 < \nu \le 1)$, where $C_\nu([0,\infty);\mathbf{R}^{m_1})$ denotes the Hölder space with the exponent ν, on $[0,\infty)$ taking values in $\mathbf{R}^{m_1}$. Then, the first equation of* (6.1) *has a unique solution.*

To the above system (6.1), we shall apply the linear controller (3.8)–(3.10) constructed for the linear diffusion system (2.1). By using the same notation as in section 4, the closed-loop system from w to z can be written as follows:

$$\text{(6.3)}\qquad \begin{cases} \dfrac{d\xi(t)}{dt} = (\mathcal{A} + \varDelta\mathcal{A})\xi(t) + \mathcal{F}(\xi(t)) + (\mathcal{B} + \varDelta\mathcal{B})w(t), \quad \xi(0) = \xi_0, \\ z(t) = (\mathcal{C} + \varDelta\mathcal{C})\xi(t), \end{cases}$$

where $\mathcal{F}(\cdot) : X \to X$ is defined as

$$\mathcal{F}(\xi(t)) := \begin{bmatrix} F_1(x_1(t), x_2(t), x_3(t)) \\ 0 \\ F_2(x_1(t), x_2(t), x_3(t)) \\ F_3(x_1(t), x_2(t), x_3(t)) \\ F_2(x_1(t), x_2(t), x_3(t)) \end{bmatrix},$$

$$F_1(\cdot) := P_l F(\cdot), \quad F_2(\cdot) := (P_n - P_l)F(\cdot), \quad F_3(\cdot) := (I - P_n)F(\cdot).$$

Here, we remark that $F_1(\cdot)$ and $F_2(\cdot)$ are identified with their corresponding vectors with respect to the bases $(\phi_{11}, \dots, \phi_{1\mu_1}, \dots, \phi_{l,\mu_l})$ of H_1 and $(\phi_{l+1,1}, \dots, \phi_{l+1,\mu_{l+1}}, \dots, \phi_{n,\mu_n})$ of H_2, respectively.

Then, we have the following theorem.

THEOREM 6.2. *Suppose that the assumptions of Theorem* 3.1 *are satisfied and that the integer n is chosen sufficiently large such that the inequality* (4.10) *holds. Moreover, assume that the Lipschitz constant k satisfies the inequality*

$$0 < k < \frac{\mu}{\sqrt{2}m''}, \tag{6.4}$$

where the constants m'' and μ are defined in (4.6) *and* (4.7)*, respectively. Then, the following results* (1) *and* (2) *hold:*

(1) *The solution of the equation*

$$\frac{d\xi(t)}{dt} = (\mathcal{A} + \varDelta\mathcal{A})\xi(t) + \mathcal{F}(\xi(t)), \quad \xi(0) = \xi_0 \tag{6.5}$$

is exponentially stable.

(2) *For the closed-loop system* (6.3) *with the initial condition $\xi(0) = 0$, the inequality*

$$\|z\|_{L^2(0,\infty;\mathbf{R}^{p_1})} \le (\delta + \delta_k)\|w\|_{L^2(0,\infty;\mathbf{R}^{m_1})}$$

holds for all $w \in L^2(0,\infty;\mathbf{R}^{m_1})$, where

$$\delta_k := \frac{\sqrt{2}km''^2\|\mathcal{B} + \varDelta\mathcal{B}\|_{\mathcal{L}(\mathbf{R}^{m_1},X)}\|\mathcal{C} + \varDelta\mathcal{C}\|_{\mathcal{L}(X,\mathbf{R}^{p_1})}}{\mu(\mu - \sqrt{2}km'')}. \tag{6.6}$$

Proof of (1). From (6.5), we have

$$\xi(t) = e^{t(\mathcal{A}+\varDelta\mathcal{A})}\xi_0 + \int_0^t e^{(t-s)(\mathcal{A}+\varDelta\mathcal{A})}\mathcal{F}(\xi(s))ds. \tag{6.7}$$

Since the condition (6.2) implies

$$\begin{cases} \|\mathcal{F}(\xi) - \mathcal{F}(\tilde{\xi})\|_X \le \sqrt{2}k\|\xi - \tilde{\xi}\|_X, & \xi, \tilde{\xi} \in X, \\ \mathcal{F}(0) = 0, \end{cases} \tag{6.8}$$

from (4.7) and (6.7), we obtain

$$e^{\mu t}\|\xi(t)\|_X \le m''\|\xi_0\|_X + \int_0^t \sqrt{2}km''e^{\mu s}\|\xi(s)\|_X ds.$$

Here, using Gronwall's inequality yields

$$\|\xi(t)\|_X \le m''\|\xi_0\|_X e^{-(\mu-\sqrt{2}km'')t}.$$

Hence, it follows under the assumption (6.4) that the solution of (6.5) is exponentially stable. □

Proof of (2). By setting $\xi(0) = 0$ in the closed-loop system (6.3), we have

$$\begin{cases} \xi(t) = \displaystyle\int_0^t e^{(t-s)(\mathcal{A}+\varDelta\mathcal{A})}\{\mathcal{F}(\xi(s)) + (\mathcal{B} + \varDelta\mathcal{B})w(s)\}ds, \\ z(t) = (\mathcal{C} + \varDelta\mathcal{C})\xi(t). \end{cases} \tag{6.9}$$

From (6.9), we get

$$z(t)=\int_0^t(\mathcal{C}+\varDelta\mathcal{C})e^{(t-s)(\mathcal{A}+\varDelta\mathcal{A})}(\mathcal{B}+\varDelta\mathcal{B})w(s)ds+\int_0^t(\mathcal{C}+\varDelta\mathcal{C})e^{(t-s)(\mathcal{A}+\varDelta\mathcal{A})}\mathcal{F}(\xi(s))ds.$$

Let us estimate the $L^2(0,\infty;\mathbf{R}^{p_1})$ norm of z. First, it follows from the above equation that

$$\begin{aligned}\|z\|_{L^2(0,\infty;\mathbf{R}^{p_1})}\le&\left\|\int_0^{\cdot}(\mathcal{C}+\varDelta\mathcal{C})e^{(\cdot-s)(\mathcal{A}+\varDelta\mathcal{A})}(\mathcal{B}+\varDelta\mathcal{B})w(s)ds\right\|_{L^2(0,\infty;\mathbf{R}^{p_1})}\\&+\left\|\int_0^{\cdot}(\mathcal{C}+\varDelta\mathcal{C})e^{(\cdot-s)(\mathcal{A}+\varDelta\mathcal{A})}\mathcal{F}(\xi(s))ds\right\|_{L^2(0,\infty;\mathbf{R}^{p_1})}.\end{aligned}\tag{6.10}$$

Here, noting that

$$\begin{aligned}&\sup_{\substack{\tilde{w}\in L^2(0,\infty;\mathbf{R}^{m_1})\\ \tilde{w}\neq0}}\frac{\left\|\int_0^{\cdot}(\mathcal{C}+\varDelta\mathcal{C})e^{(\cdot-s)(\mathcal{A}+\varDelta\mathcal{A})}(\mathcal{B}+\varDelta\mathcal{B})\tilde{w}(s)ds\right\|_{L^2(0,\infty;\mathbf{R}^{p_1})}}{\|\tilde{w}\|_{L^2(0,\infty;\mathbf{R}^{m_1})}}\\&\qquad=\|G_{zw}(\cdot)\|_{\mathcal{H}_\infty(\mathcal{L}(\mathbf{C}^{m_1},\mathbf{C}^{p_1}))}<\delta,\end{aligned}$$

it is easy to see that the first term of the right-hand side of (6.10) is estimated as

$$\left\|\int_0^{\cdot}(\mathcal{C}+\varDelta\mathcal{C})e^{(\cdot-s)(\mathcal{A}+\varDelta\mathcal{A})}(\mathcal{B}+\varDelta\mathcal{B})w(s)ds\right\|_{L^2(0,\infty;\mathbf{R}^{p_1})}\le\delta\|w\|_{L^2(0,\infty;\mathbf{R}^{m_1})}\tag{6.11}$$

for all $w\in L^2(0,\infty;\mathbf{R}^{m_1})$. Therefore, we have only to consider the second term of the right-hand side of (6.10). Using (4.7) and (6.8), and using Hölder's inequality, we get

$$\begin{aligned}&\left\|\int_0^t(\mathcal{C}+\varDelta\mathcal{C})e^{(t-s)(\mathcal{A}+\varDelta\mathcal{A})}\mathcal{F}(\xi(s))ds\right\|_{\mathbf{R}^{p_1}}\\&\le\sqrt{2}km''\|\mathcal{C}+\varDelta\mathcal{C}\|_{\mathcal{L}(X,\mathbf{R}^{p_1})}\int_0^te^{-\mu(t-s)}\|\xi(s)\|_Xds\\&\le\sqrt{2}km''\|\mathcal{C}+\varDelta\mathcal{C}\|_{\mathcal{L}(X,\mathbf{R}^{p_1})}\left(\int_0^te^{-\mu(t-s)}ds\right)^{\frac12}\left(\int_0^te^{-\mu(t-s)}\|\xi(s)\|_X^2ds\right)^{\frac12}\\&\le\sqrt{2}km''\|\mathcal{C}+\varDelta\mathcal{C}\|_{\mathcal{L}(X,\mathbf{R}^{p_1})}\left(\frac{1}{\mu}\right)^{\frac12}\left(\int_0^te^{-\mu(t-s)}\|\xi(s)\|_X^2ds\right)^{\frac12}.\end{aligned}\tag{6.12}$$

Here, squaring the both sides of (6.12), integrating from 0 to ∞ with respect to t, and interchanging the order of integration, we have

$$\begin{aligned}&\int_0^\infty\left\|\int_0^t(\mathcal{C}+\varDelta\mathcal{C})e^{(t-s)(\mathcal{A}+\varDelta\mathcal{A})}\mathcal{F}(\xi(s))ds\right\|_{\mathbf{R}^{p_1}}^2dt\\&\le2k^2m''^2\|\mathcal{C}+\varDelta\mathcal{C}\|^2_{\mathcal{L}(X,\mathbf{R}^{p_1})}\frac{1}{\mu}\int_0^\infty\left(\int_0^te^{-\mu(t-s)}\|\xi(s)\|_X^2ds\right)dt\\&=2k^2m''^2\|\mathcal{C}+\varDelta\mathcal{C}\|^2_{\mathcal{L}(X,\mathbf{R}^{p_1})}\frac{1}{\mu}\int_0^\infty\left(\int_s^\infty e^{-\mu(t-s)}\|\xi(s)\|_X^2dt\right)ds\\&=2k^2m''^2\|\mathcal{C}+\varDelta\mathcal{C}\|^2_{\mathcal{L}(X,\mathbf{R}^{p_1})}\frac{1}{\mu}\int_0^\infty\|\xi(s)\|_X^2\left(\int_s^\infty e^{-\mu(t-s)}dt\right)ds\\&=2k^2m''^2\|\mathcal{C}+\varDelta\mathcal{C}\|^2_{\mathcal{L}(X,\mathbf{R}^{p_1})}\frac{1}{\mu^2}\|\xi\|^2_{L^2(0,\infty;X)}.\end{aligned}\tag{6.13}$$

From (6.13), we obtain

$$\begin{aligned}
&\left\| \int_0^{\cdot} (\mathcal{C}+\Delta\mathcal{C}) e^{(\cdot - s)(\mathcal{A}+\Delta\mathcal{A})} \mathcal{F}(\xi(s)) ds \right\|_{L^2(0,\infty;\mathbf{R}^{p_1})} \\
&\leq \sqrt{2} k m'' \|\mathcal{C}+\Delta\mathcal{C}\|_{\mathcal{L}(X,\mathbf{R}^{p_1})} \frac{1}{\mu} \|\xi\|_{L^2(0,\infty;X)}.
\end{aligned} \tag{6.14}$$

Finally we estimate $\|\xi\|_{L^2(0,\infty;X)}$ in the right-hand side of (6.14) by $\|w\|_{L^2(0,\infty;\mathbf{R}^{m_1})}$. By using (4.7) and (6.8), and by using Hölder's inequality, it follows from the first equation of (6.9) that

$$\begin{aligned}
\|\xi(t)\|_X &\leq \sqrt{2} k m'' \int_0^t e^{-\mu(t-s)} \|\xi(s)\|_X ds \\
&\quad + m'' \|\mathcal{B}+\Delta\mathcal{B}\|_{\mathcal{L}(\mathbf{R}^{m_1},X)} \int_0^t e^{-\mu(t-s)} \|w(s)\|_{\mathbf{R}^{m_1}} ds \\
&\leq \sqrt{2} k m'' \left(\int_0^t e^{-\mu(t-s)} ds \right)^{\frac{1}{2}} \left(\int_0^t e^{-\mu(t-s)} \|\xi(s)\|_X^2 ds \right)^{\frac{1}{2}} \\
&\quad + m'' \|\mathcal{B}+\Delta\mathcal{B}\|_{\mathcal{L}(\mathbf{R}^{m_1},X)} \left(\int_0^t e^{-\mu(t-s)} ds \right)^{\frac{1}{2}} \left(\int_0^t e^{-\mu(t-s)} \|w(s)\|_{\mathbf{R}^{m_1}}^2 ds \right)^{\frac{1}{2}} \\
&\leq \sqrt{2} k m'' \left(\frac{1}{\mu} \right)^{\frac{1}{2}} \left(\int_0^t e^{-\mu(t-s)} \|\xi(s)\|_X^2 ds \right)^{\frac{1}{2}} \\
&\quad + m'' \|\mathcal{B}+\Delta\mathcal{B}\|_{\mathcal{L}(\mathbf{R}^{m_1},X)} \left(\frac{1}{\mu} \right)^{\frac{1}{2}} \left(\int_0^t e^{-\mu(t-s)} \|w(s)\|_{\mathbf{R}^{m_1}}^2 ds \right)^{\frac{1}{2}}.
\end{aligned} \tag{6.15}$$

Here, squaring both sides of (6.15) and integrating from 0 to ∞ with respect to t yield

$$\begin{aligned}
\int_0^\infty \|\xi(t)\|_X^2 dt \leq \Bigg[& \sqrt{2} k m'' \frac{1}{\mu} \left(\int_0^\infty \|\xi(s)\|_X^2 ds \right)^{\frac{1}{2}} \\
& + m'' \|\mathcal{B}+\Delta\mathcal{B}\|_{\mathcal{L}(\mathbf{R}^{m_1},X)} \frac{1}{\mu} \left(\int_0^\infty \|w(s)\|_{\mathbf{R}^{m_1}}^2 ds \right)^{\frac{1}{2}} \Bigg]^2.
\end{aligned}$$

In the above, we interchange the order of integration and use Hölder's inequality. Therefore, we have

$$\left(1 - \sqrt{2} k m'' \frac{1}{\mu} \right) \|\xi\|_{L^2(0,\infty;X)} \leq m'' \|\mathcal{B}+\Delta\mathcal{B}\|_{\mathcal{L}(\mathbf{R}^{m_1},X)} \frac{1}{\mu} \|w\|_{L^2(0,\infty;\mathbf{R}^{m_1})}.$$

Hence, it follows under the assumption (6.4) that

$$\|\xi\|_{L^2(0,\infty;X)} \leq \frac{m'' \|\mathcal{B}+\Delta\mathcal{B}\|_{\mathcal{L}(\mathbf{R}^{m_1},X)}}{\mu - \sqrt{2} k m''} \|w\|_{L^2(0,\infty;\mathbf{R}^{m_1})}. \tag{6.16}$$

Here, combining (6.14) and (6.16) gives

$$\left\| \int_0^{\cdot} (\mathcal{C}+\Delta\mathcal{C}) e^{(\cdot - s)(\mathcal{A}+\Delta\mathcal{A})} \mathcal{F}(\xi(s)) ds \right\|_{L^2(0,\infty;\mathbf{R}^{p_1})} \leq \delta_k \|w\|_{L^2(0,\infty;\mathbf{R}^{m_1})}, \tag{6.17}$$

where

$$\delta_k = \frac{\sqrt{2} k m''^2 \|\mathcal{B}+\Delta\mathcal{B}\|_{\mathcal{L}(\mathbf{R}^{m_1},X)} \|\mathcal{C}+\Delta\mathcal{C}\|_{\mathcal{L}(X,\mathbf{R}^{p_1})}}{\mu(\mu - \sqrt{2} k m'')}.$$

From (6.10), (6.11), and (6.17), we finally obtain

$$\|z\|_{L^2(0,\infty;\mathbf{R}^{p_1})} \le (\delta + \delta_k)\|w\|_{L^2(0,\infty;\mathbf{R}^{m_1})}$$

for all $w \in L^2(0,\infty;\mathbf{R}^{m_1})$. □

The proof of Theorem 6.2 is thus complete.

Remark. From (6.6), it is easy to see that $\delta_k \to 0$ as $k \downarrow 0$ and $\delta_k \to \infty$ as $k \uparrow \frac{\mu}{\sqrt{2m''}}$. This shows that the finite-dimensional $\mathcal{H}_\infty$ controller constructed for the linear infinite-dimensional system also works as a finite-dimensional $\mathcal{H}_\infty$ controller for a semilinear infinite-dimensional system with sufficiently small Lipschitz constant k.

Appendix: Derivation of (4.8). Noting that

$$\begin{aligned}
&(sI-\mathcal{A})^{-1}\\
&= \begin{bmatrix} sI-A_1 & -B_{21}L & 0 & 0 & 0 \\ -NC_{21} & sI-M & 0 & 0 & -NC_{22} \\ 0 & -B_{22}L & sI-A_2 & 0 & 0 \\ 0 & 0 & 0 & sI-A_3 & 0 \\ 0 & 0 & 0 & 0 & sI-A_2 \end{bmatrix}^{-1}\\
&= \begin{bmatrix} \begin{bmatrix} sI-A_1 & -B_{21}L & 0 \\ -NC_{21} & sI-M & 0 \\ 0 & -B_{22}L & sI-A_2 \end{bmatrix}^{-1} & -\begin{bmatrix} sI-A_1 & -B_{21}L & 0 \\ -NC_{21} & sI-M & 0 \\ 0 & -B_{22}L & sI-A_2 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ 0 & -NC_{22} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} sI-A_3 & 0 \\ 0 & sI-A_2 \end{bmatrix}^{-1} \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} & \begin{bmatrix} sI-A_3 & 0 \\ 0 & sI-A_2 \end{bmatrix}^{-1} \end{bmatrix}
\end{aligned}$$

and

$$\begin{aligned}
&\begin{bmatrix} sI-A_1 & -B_{21}L & 0 \\ -NC_{21} & sI-M & 0 \\ 0 & -B_{22}L & sI-A_2 \end{bmatrix}^{-1}\\
&= \begin{bmatrix} \begin{bmatrix} sI-A_1 & -B_{21}L \\ -NC_{21} & sI-M \end{bmatrix}^{-1} & \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ -(sI-A_2)^{-1}\begin{bmatrix} 0 & -B_{22}L \end{bmatrix}\begin{bmatrix} sI-A_1 & -B_{21}L \\ -NC_{21} & sI-M \end{bmatrix}^{-1} & (sI-A_2)^{-1} \end{bmatrix},
\end{aligned}$$

$G(s) = \mathcal{C}(sI-\mathcal{A})^{-1}\mathcal{B}$ can be calculated as

$$G(s) = C_{12}(sI-A_2)^{-1}B_{12} + G_1(s) + G_2(s) + G_3(s) + G_4(s),$$

where

$$G_1(s) := \begin{bmatrix} C_{11} & D_{12}L \end{bmatrix} \begin{bmatrix} sI-A_1 & -B_{21}L \\ -NC_{21} & sI-M \end{bmatrix}^{-1} \begin{bmatrix} B_{11} \\ ND_{21} \end{bmatrix} = T_{zw}(s),$$

$$G_2(s) := C_{12}(sI-A_2)^{-1} \begin{bmatrix} 0 & B_{22}L \end{bmatrix} \begin{bmatrix} sI-A_1 & -B_{21}L \\ -NC_{21} & sI-M \end{bmatrix}^{-1} \begin{bmatrix} B_{11} \\ ND_{21} \end{bmatrix},$$

$$G_3(s) := \begin{bmatrix} C_{11} & D_{12}L \end{bmatrix} \begin{bmatrix} sI - A_1 & -B_{21}L \\ -NC_{21} & sI - M \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ NC_{22}(sI - A_2)^{-1}B_{12} \end{bmatrix},$$

$$G_4(s) := C_{12}(sI - A_2)^{-1} \begin{bmatrix} 0 & B_{22}L \end{bmatrix} \begin{bmatrix} sI - A_1 & -B_{21}L \\ -NC_{21} & sI - M \end{bmatrix}^{-1} \times \begin{bmatrix} 0 \\ NC_{22}(sI - A_2)^{-1}B_{12} \end{bmatrix}.$$

Here, by using the assumptions (A3) and (A4) (i.e., (A3′) and (A4′)), $G_2(s)$, $G_3(s)$, and $G_4(s)$ can be rewritten as follows:

$$\begin{aligned}
&G_2(s) \\
&= C_{12}(sI - A_2)^{-1}B_{22} \begin{bmatrix} 0 & L \end{bmatrix} \begin{bmatrix} sI - A_1 & -B_{21}L \\ -NC_{21} & sI - M \end{bmatrix}^{-1} \begin{bmatrix} B_{11} \\ ND_{21} \end{bmatrix} \\
&= C_{12}(sI - A_2)^{-1}B_{22} \begin{bmatrix} D_{12}^TC_{11} & D_{12}^TD_{12}L \end{bmatrix} \begin{bmatrix} sI - A_1 & -B_{21}L \\ -NC_{21} & sI - M \end{bmatrix}^{-1} \begin{bmatrix} B_{11} \\ ND_{21} \end{bmatrix} \\
&= C_{12}(sI - A_2)^{-1}B_{22}D_{12}^T \begin{bmatrix} C_{11} & D_{12}L \end{bmatrix} \begin{bmatrix} sI - A_1 & -B_{21}L \\ -NC_{21} & sI - M \end{bmatrix}^{-1} \begin{bmatrix} B_{11} \\ ND_{21} \end{bmatrix} \\
&= C_{12}(sI - A_2)^{-1}B_{22}D_{12}^TT_{zw}(s),
\end{aligned}$$

$$\begin{aligned}
&G_3(s) \\
&= \begin{bmatrix} C_{11} & D_{12}L \end{bmatrix} \begin{bmatrix} sI - A_1 & -B_{21}L \\ -NC_{21} & sI - M \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ N \end{bmatrix} C_{22}(sI - A_2)^{-1}B_{12} \\
&= \begin{bmatrix} C_{11} & D_{12}L \end{bmatrix} \begin{bmatrix} sI - A_1 & -B_{21}L \\ -NC_{21} & sI - M \end{bmatrix}^{-1} \begin{bmatrix} B_{11}D_{21}^T \\ ND_{21}D_{21}^T \end{bmatrix} C_{22}(sI - A_2)^{-1}B_{12} \\
&= \begin{bmatrix} C_{11} & D_{12}L \end{bmatrix} \begin{bmatrix} sI - A_1 & -B_{21}L \\ -NC_{21} & sI - M \end{bmatrix}^{-1} \begin{bmatrix} B_{11} \\ ND_{21} \end{bmatrix} D_{21}^TC_{22}(sI - A_2)^{-1}B_{12} \\
&= T_{zw}(s)D_{21}^TC_{22}(sI - A_2)^{-1}B_{12},
\end{aligned}$$

$$\begin{aligned}
&G_4(s) \\
&= C_{12}(sI - A_2)^{-1}B_{22} \begin{bmatrix} 0 & L \end{bmatrix} \begin{bmatrix} sI - A_1 & -B_{21}L \\ -NC_{21} & sI - M \end{bmatrix}^{-1} \\
&\qquad \times \begin{bmatrix} 0 \\ N \end{bmatrix} C_{22}(sI - A_2)^{-1}B_{12} \\
&= C_{12}(sI - A_2)^{-1}B_{22} \begin{bmatrix} D_{12}^TC_{11} & D_{12}^TD_{12}L \end{bmatrix} \begin{bmatrix} sI - A_1 & -B_{21}L \\ -NC_{21} & sI - M \end{bmatrix}^{-1} \\
&\qquad \times \begin{bmatrix} B_{11}D_{21}^T \\ ND_{21}D_{21}^T \end{bmatrix} C_{22}(sI - A_2)^{-1}B_{12} \\
&= C_{12}(sI - A_2)^{-1}B_{22}D_{12}^T \begin{bmatrix} C_{11} & D_{12}L \end{bmatrix} \begin{bmatrix} sI - A_1 & -B_{21}L \\ -NC_{21} & sI - M \end{bmatrix}^{-1} \\
&\qquad \times \begin{bmatrix} B_{11} \\ ND_{21} \end{bmatrix} D_{21}^TC_{22}(sI - A_2)^{-1}B_{12} \\
&= C_{12}(sI - A_2)^{-1}B_{22}D_{12}^TT_{zw}(s)D_{21}^TC_{22}(sI - A_2)^{-1}B_{12}.
\end{aligned}$$

Therefore, we finally obtain

$$\begin{aligned}G(s) = {} & C_{12}(sI - A_2)^{-1}B_{12} + T_{zw}(s) + C_{12}(sI - A_2)^{-1}B_{22}D_{12}^T T_{zw}(s) \\ & +T_{zw}(s)D_{21}^T C_{22}(sI - A_2)^{-1}B_{12} \\ & +C_{12}(sI - A_2)^{-1}B_{22}D_{12}^T T_{zw}(s)D_{21}^T C_{22}(sI - A_2)^{-1}B_{12}.\end{aligned}$$

REFERENCES

[1] M. J. Balas, *Finite-dimensional controllers for linear distributed parameter systems: Exponential stability using residual mode filters*, J. Math. Anal. Appl., 133 (1988), pp. 283–296.

[2] R. Y. Chiang and M. G. Safonov, *Robust Control TOOLBOX For Use with MATLAB*, The MathWorks, Inc., Natick, MA, 1992.

[3] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, *State-space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.

[4] B. Van Keulen, *Redheffer's lemma and H^∞-control for infinite-dimensional systems*, SIAM J. Control Optim., 32 (1994), pp. 261–278.

[5] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.

[6] Y. Sakawa, *Feedback stabilization of linear diffusion systems*, SIAM J. Control Optim., 21 (1983), pp. 667–676.

[7] H. Sano and N. Kunimatsu, *Feedback stabilization of infinite-dimensional systems with A^γ-bounded output operators*, Appl. Math. Lett., 7 (1994), pp. 17–22.

[8] H. Sano and N. Kunimatsu, *An application of inertial manifold theory to boundary stabilization of semilinear diffusion systems*, J. Math. Anal. Appl., 196 (1995), pp. 18–42.

[9] J. M. Schumacher, *Dynamic Feedback in Finite- and Infinite-Dimensional Linear Systems*, Mathematical Centre Tracts 143, Mathematisch Centrum, Amsterdam, 1981.

[10] K. Taira, *Analytic Semigroups and Semilinear Initial Boundary Value Problems*, London Math. Soc. Lecture Note Ser. 223, Cambridge University Press, Cambridge, 1995.

[11] H. Tanabe, *Equations of Evolution*, Iwanami, Tokyo, 1975 (in Japanese).

SIAM J. Control Optim.
Vol. 37, No. 2, pp. 429–460

GENERALIZED FOURIER AND TOEPLITZ RESULTS FOR RATIONAL ORTHONORMAL BASES*

BRETT NINNESS†, HÅKAN HJALMARSSON‡, AND FREDRIK GUSTAFSSON§

Abstract. This paper provides a generalization of certain classical Fourier convergence and asymptotic Toeplitz matrix properties to the case where the underlying orthonormal basis is not the conventional trigonometric one but rather a rational generalization which encompasses the trigonometric one as a special case. These generalized Fourier and Toeplitz results have particular application in dynamic system estimation theory. Specifically, the results allow a unified treatment of the accuracy of least-squares system estimation using a range of model structures, including those that allow the inclusion of prior knowledge of system dynamics via the specification of fixed pole or zero locations.

Key words. stochastic processes, prediction theory, system identification

AMS subject classifications. 42C15, 93E12, 47B35, 93E24, 60G35

PII. S0363012996305437

1. Introduction.

In the area of applied mathematics, a fundamental idea is that of approximating or exactly expressing solutions by expanding them in terms of orthogonal basis functions. Well-known classical examples are Fourier analysis, solutions of the wave equation and Schrödinger's equation in terms of (respectively) Legendre and Laguerre orthogonal polynomials, and solutions of self-adjoint operator equations such as Sturm–Liouville systems in terms of the orthogonal eigenfunctions of the operator. More recently, particularly for the solution of signal processing and other system theoretic problems, there has been an explosion of interest in the development and use of a wide class of new orthogonal bases called "wavelets" [7, 5].

Indeed, tackling system theoretic problems using orthonormal descriptions has a particularly rich history, going back at least as far as the work of Kolmogorov [23] and Wiener [52], who exploited them in developing their now famous theory on the prediction of random processes. In that work, the orthonormal basis was the trigonometric one, but as was shown by Szegö there is great utility in reexpressing the problem with respect to another orthonormal basis that is adapted to the random process; namely, a basis of polynomials orthogonal to a given positive function f which is the spectral density of the process [45, 11]. Such polynomials are called "Szegö polynomials."

This latter approach derives its utility from the fact that the nth order Szegö polynomial is in fact the mean-square best, order n, one step ahead predictor of the random process [14, 45]. By exploiting the orthonormality of the basis to derive what is called a "Christoffel–Darboux" formula for the "reproducing kernel" associated with

[*] Received by the editors June 21, 1996; accepted for publication (in revised form) May 12, 1998; published electronically November 23, 1998.

http://www.siam.org/journals/sicon/37-2/30543.html

[†] Centre for Integrated Dynamics and Control (CIDAC) and Department of Electrical and Computer Engineering, University of Newcastle, Callaghan 2308, Australia (brett@ee.newcastle.edu.au). The research of this author was supported by the Australian Research Council and CIDAC. Part of this research was completed while the author was visiting the Department of Sensors, Signals and Systems-Automatic Control, The Royal Institute of Technology, S-100 44 Stockholm, Sweden.

[‡] Department of Sensors, Signals and Systems-Signal Processing, The Royal Institute of Technology, S-100 44 Stockholm, Sweden (hakan.hjalmarsson@s3.e.kth.se). This research was partly completed while the author was at CIDAC.

[§] Department of Electrical Engineering, Linköping University, S-581 83, Linköping, Sweden (fredrik@isy.liu.se). This research was partly completed while the author was at CIDAC.

the Szegö polynomial basis, theoretical analysis of this predictor is greatly facilitated. For example, it was by this means that Szegö was able to derive his famous formula

$$\sigma^2 = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log f(\omega)\,\mathrm{d}\omega\right\}$$

for the asymptotic in order n variance σ^2 of the prediction error associated with a spectral density f.

In addition, use of the Christoffel–Darboux formula provides a recursive in n formula for the Szegö polynomials [45, 11], and this in turn allows a computationally efficient means for calculating predictors. This recursive formula is of course the famous Levinson recursion, which was developed independently of Szegö's work by exploiting the properties of Toeplitz matrices [26, 38]. In practice, the so-called "reflection coefficients" required in the Levinson recursions are calculated by the Schur algorithm [41], originally proposed by Schur [43] as a means for testing whether or not a function is bounded positive real (or "Carathéodory" as it is known in some literature). Here again orthonormal bases and Toeplitz matrices arise since another test for positive realness involves testing for the positive definiteness of the Toeplitz matrix formed from the Fourier coefficients of the function [44].

These several links between Toeplitz matrices and orthonormal bases arise since (subject to some regularity conditions) the ℓ,mth element of any $n \times n$ symmetric Toeplitz matrix $T_n(f)$ may be expressed using the orthonormal trigonometric basis $\{e^{j\omega n}\}$ as

$$[T_n(f)]_{\ell,m} = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{j\omega\ell}e^{-j\omega m} f(\omega)\,\mathrm{d}\omega \tag{1.1}$$

for some positive function f. By recognizing this, certain quadratic forms of Toeplitz matrices that arise naturally in the frequency domain analysis of least-squares estimation problems may instead be conveniently rewritten as

$$\frac{1}{n}\Gamma_n^\star(\omega)T_n(f)\Gamma_n(\omega) = \sum_{k=-n}^{n}\left(1 - \frac{|k|}{n}\right) c_k\, e^{j\omega k} \tag{1.2}$$

where $\cdot^\star$ denotes "conjugate transpose" and

$$c_k \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\omega)e^{-j\omega k}\,\mathrm{d}\omega$$

is the kth Fourier coefficient of f with $\Gamma_n(\omega)$ an $n \times 1$ vector defined as

$$\Gamma_n^\star(\omega) \triangleq [1, e^{-j\omega}, e^{-j2\omega}, \ldots, e^{-j(n-1)\omega}].$$

The right-hand side of (1.2) may be recognized as the Cesàro mean reconstruction of a Fourier series which is known [10], provided f is continuous, to converge uniformly to $f(\omega)$ on $[-\pi, \pi]$.

This latter fact has been exploited by Ljung and coauthors [29, 31, 16, 30, 53, 27], who, reminiscent of Szegö's approach of examining the asymptotic in order n nature of predictors, have provided asymptotic-in-model-order results describing the variability of the frequency response of least-squares system estimates in such a way as to elucidate how they depend on excitation and measurement noise spectral densities, model order, and observed data length; see section 7 for more detail on this point.

Such results have found wide engineering application; see for example [2, 12, 28, 17]. However, to derive them, another key ingredient pertaining to the properties of Toeplitz matrices is required, namely, that asymptotically in size n, Toeplitz matrices possess the algebraic structure [14, 50]

$$T_n(f)T_n(g) \sim T_n(fg), \tag{1.3}$$

where f and g are any continuous positive functions, and for $n \times n$ matrices A_n and B_n, the notation $A_n \sim B_n$ means that $\lim_{n\to\infty} |A_n - B_n| = 0$, where $|\cdot|$ is the Hilbert–Schmidt matrix norm defined by

$$|A|^2 \triangleq \frac{1}{n}\text{Trace}\{A^\star A\}. \tag{1.4}$$

The main results of this paper extend the results of the convergence of the Cesàro mean (1.2) and the algebraic structure of Toeplitz matrices (1.3) to more general cases wherein the underlying orthonormal basis is not the trigonometric one but rather a generalization of it. More specifically, this paper studies the use of the basis functions $\mathcal{B}_n(z)$ given by

$$\mathcal{B}_n(z) \triangleq \frac{\sqrt{1-|\xi_n|^2}}{1-\xi_n z}\prod_{k=0}^{n-1}\left(\frac{z-\overline{\xi_k}}{1-\xi_k z}\right), \tag{1.5}$$

where the $\{\xi_k\}$ may be chosen (almost) arbitrarily inside and (in some cases) on the boundary of the open unit disc $\mathbf{D} \triangleq \{z \in \mathbf{C} : |z| < 1\}$ ($\mathbf{C}$ is the field of complex numbers). These functions $\{\mathcal{B}_n\}$ are orthonormal on the unit circle $\mathbf{T} = \{z \in \mathbf{C} : |z| = 1\}$, and the trigonometric basis is a special case of them if all the $\{\xi_k\}$ are chosen as zero. Using them, a generalization

$$[M_n(f)]_{\ell,m} \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{B}_\ell(e^{j\omega})\overline{\mathcal{B}_m(e^{j\omega})}f(\omega)\,\mathrm{d}\omega \tag{1.6}$$

of Toeplitz matrices is considered, for which it is shown here that a generalization of (1.3) still holds, and with the redefinition

$$\Gamma_n^T(\omega) \triangleq [\mathcal{B}_0(e^{j\omega}),\ \mathcal{B}_1(e^{j\omega}),\ \dots, \mathcal{B}_{n-1}(e^{j\omega})] \tag{1.7}$$

it is also shown here that a generalization of the uniform convergence of the Cesàro mean (1.2) to $f(\omega)$ also holds.

In both cases, the generalization involves replacing the $1/n$ normalization appearing in (1.2) and in the definition of the matrix norm (1.4) with a frequency dependent term $K_n(\omega,\omega)$ which is the reproducing kernel associated with the linear space spanned by the basis functions $\{\mathcal{B}_0, \mathcal{B}_1, \dots, \mathcal{B}_{n-1}\}$.

Indeed, this reproducing kernel is the key to the results presented here. Classical derivations of Cesàro summability and Toeplitz matrix results rely heavily on the algebraic structure of the trigonometric basis, namely, that $e^{j\omega n}e^{j\omega m} = e^{j\omega(n+m)}$. In the cases considered here, since $\mathcal{B}_n\mathcal{B}_m \neq \mathcal{B}_{n+m}$ this algebraic structure is lost and pre-existing analysis techniques are not applicable. Instead, motivated by Szegö's approach to the study of orthogonal polynomials, this paper exploits a closed form expression for the appropriate reproducing kernel.

The utility of the new results presented here is that, just as the classical Fourier and Toeplitz results have been used by Ljung and coauthors to analyze estimation using finite impulse response (FIR) and certain other rational model structures, the results of this paper can be used to analyze estimation using generalizations of these model structures. As shown in [37] and as commented on in section 7, these generalized structures are actually quite common since they implicitly arise whenever the common practice of data prefiltering is performed.

There is much other work related to the results presented here. The study of the basis functions (1.5) dates back to Malmquist [34] and was taken up by Walsh [49] in the context of complex rational approximation theory and by other authors [52, 25, 21, 3, 9, 8, 32, 39] for system theoretic applications such as system approximation and network synthesis, including generalizations of Schur and Levinson recursions, lattice structures, and the concomitant solution of inverse scattering problems.

In the context of system identification, as well as pertaining to the aforementioned work [29, 31, 16, 30, 53, 27], the results of this paper also have close connections with much recent literature examining the use of model structures derived from orthonormal bases. In [22, 6, 20, 46, 48] the use of the so-called "Laguerre" basis is examined. This basis can be encompassed by the basis (1.5) by fixing all the poles at a common value $\xi_k = \xi \in \mathbf{R}$ ($\mathbf{R}$ denotes the field of real numbers) and with the substitution $z \mapsto 1/z$ so as to accommodate convention in the signal processing and control theory literature. In this case the name "Laguerre" derives from the ensuing functions being related to the classical Laguerre orthonormal polynomials via a Fourier and bilinear transform [35]. In [47] a generalization of this Laguerre case is analyzed wherein the common value ξ may be complex valued. In [18, 40], these analyses are again generalized to the case where a fixed set of poles $\{\xi_0, \ldots, \xi_r\}$ are cyclically repeated and orthonormal bases are generated with denominators given as $D_p(z) = \prod_{k=0}^{p-1}(z - \xi_k)$ and numerators as Szegö polynomials associated with the weight function $|D_p(e^{j\omega})|^{-2}$. The cyclic repetition of poles arises due to the latter numerator and denominator pair being multiplied by powers of the all-pass function $z^p D_p(1/z)/D_p(z)$ as the number of required basis functions increases beyond p.

In all these works, any analysis of estimation accuracy proceeds by exploiting the restriction on the choice of ξ_k to establish, via a bilinear transform [46, 47, 48] or a multilinear transform (dubbed a "Hambo" transform) [40] an algebra isomorphism to the trigonometric basis $\{e^{j\omega n}\}$. The utility of this is that the original results of Ljung [31] can then be employed, having been mapped through the isomorphism, to provide quantification of estimation accuracy.

In spite of the elegance of this approach, it suffers several drawbacks which are the motivation for the work at hand. First, the results pertain only to a restricted class of models in which all the poles $\{\xi_k\}$ either are chosen the same [46, 48, 47] or are cyclically repeated from a fixed set [40]. Second, and with particular reference to [40], the results are asymptotic not, as is the case here, to the number of poles $\{\xi_k\}$ chosen but to the number of times the whole set $\{\xi_0, \ldots, \xi_{p-1}\}$ is repeated. The results in this paper allow the avoidance of these limitations by eschewing a strategy of forcing an algebra isomorphism to the trigonometric case.

The presentation of these ideas is organized as follows. In section 2, the analysis begins by establishing that the general orthonormal bases (1.5) fundamental to this paper form a complete set in the Hilbert space $H_2(\mathbf{T})$. In order to study other approximating properties of the basis, a "reproducing kernel" approach is employed, and section 3 is devoted to explaining certain important principles relevant to this framework. Perhaps more importantly, section 3 also contains the derivation of a closed form "Christoffel–Darboux" type formula for the reproducing kernel. With these results in hand, section 4 then considers generalized Fourier analysis with respect to the basis (1.5), and using the reproducing kernel ideas establishes uniform convergence for generalized Cesàro mean reconstructions.

In fact, because of application demands, something more is derived: it is shown that for certain frequencies being different, uniform convergence to zero also ensues. The generalized Cesàro mean reconstruction is defined with respect to a generalized Toeplitz matrix, and section 5 is devoted to the study of the asymptotic algebraic properties of such matrices since, as already explained, these properties are of great utility in certain system theoretic applications.

Pertinent to this, section 5 defines a new notion of asymptotic equivalence between matrices and then uses this to establish that asymptotically, arbitrary products of generalized Toeplitz matrices and their inverses are equivalent to a single generalized Toeplitz matrix with a symbol equal to the product of the corresponding symbols and inverse symbols of the matrices in the product. This study of generalized Toeplitz matrix properties in terms of its symbol is continued in section 6, where the relationship between the spectrum of the matrices and the values of the symbol are explored and found to be intimately connected. Given these theoretical developments, section 7 provides a very brief overview of how the results here may be applied in the study of certain system identification problems that were, in fact, the original motivation for this work. More detail on this application is provided in the separate work [37]. Finally, section 8 provides a summary and concluding perspectives on the work presented here.

2. Completeness properties. The theme of this paper is examination of certain system theoretic issues pertaining to the use of the basis functions (1.5) for the purposes of describing discrete time dynamic systems. In what follows only bounded-input, bounded-output stable and causal systems will be of interest, so that it is natural to embed the analysis in the Hardy space $H_2(\mathbf{T})$ of functions $f(z)$ which are analytic on $\mathbf{D}$, square integrable on $\mathbf{T}$, and possess only a one-sided Fourier expansion. As is well known [19], $H_2(\mathbf{T})$ is a Hilbert space when endowed with the inner product

$$(2.1)\quad \langle f,g\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(e^{j\omega})\overline{g(e^{j\omega})}\,\mathrm{d}\omega = \frac{1}{2\pi j}\oint_{\mathbf{T}} f(z)\overline{g(z)}\,\frac{\mathrm{d}z}{z}, \quad f,g\in H_2(\mathbf{T}).$$

That the functions (1.5) form an orthonormal set in that $\langle \mathcal{B}_n,\mathcal{B}_m\rangle = \delta(n-m) =$ Kronecker delta may easily be shown [36] using the contour integral formulation of the inner product in (2.1) and Cauchy's residue theorem.

What must be of central interest if the functions (1.5) are to be useful in such a system theoretic setting is whether or not linear combinations of them can describe an arbitrary system in $H_2(\mathbf{T})$ to any degree of accuracy. This may be answered in the affirmative by the following completeness result which has been developed elsewhere, but is presented here for the sake of a self-contained presentation.

Theorem 2.1 (Ninness and Gustafsson [36]).

$$\overline{\mathrm{Span}\,\{\mathcal{B}_k(z)\}_{k\ge 0}} = H_2(\mathbf{T})$$

if and only if

$$\sum_{k=0}^{\infty}(1-|\xi_k|)=\infty,$$

where here $\overline{X}$ denotes the norm closure of the space X.

3. Reproducing kernels. Given the completeness result in Theorem 2.1, to further examine the properties of approximants formed as linear combinations of the basis functions (1.5), this paper utilizes the ideas of reproducing kernel spaces [1, 45], of which a brief overview of the key ideas is as follows.

If an approximant $f_n(z)$ of a function $f(z)$ is formed as a certain linear combination of the n basis functions $\{\mathcal{B}_0, \mathcal{B}_1, \ldots, \mathcal{B}_{n-1}\}$, then for any $\mu \in \mathbf{D}$ it is also possible to form a linear functional F_μ defined as follows:

$$F_\mu : X_n \triangleq \operatorname{Span}\{\mathcal{B}_0, \mathcal{B}_1, \ldots, \mathcal{B}_{n-1}\} \to \mathbf{C}, \quad f_n \mapsto f_n(\mu).$$

Since the setting is a Hilbert space $H_2(\mathbf{T})$, by the Riesz representation theorem [42] there exists a unique μ dependent element in X_n, called $K_n(z,\mu)$, such that

$$g(\mu) \triangleq F_\mu(g) = \langle g(z), K_n(z,\mu)\rangle \quad \forall g \in X_n.$$

This element $K_n(z,\mu)$ is called the "reproducing kernel" on account of its property of reproducing values of elements of X_n at the point μ via an inner product.

Certain basic properties of $K_n(z,\mu)$ important for the purposes of this paper are that it is "Hermitian symmetric" as can be easily seen according to

$$K_n(\sigma,\mu) = \langle K_n(z,\mu), K_n(z,\sigma)\rangle = \overline{\langle K_n(z,\sigma), K_n(z,\mu)\rangle} = \overline{K_n(\mu,\sigma)}$$

so that $K_n(\mu,\mu) \in \mathbf{R}$ and also $K_n(\mu,\mu) > 0$, since if it were not, then since $K_n(\mu,\mu) = \langle K_n(z,\mu), K(z,\mu)\rangle = \|K_n(z,\mu)\|^2$, then $\|K_n(z,\mu)\| = 0$ would be implied, which would further imply that $g(\mu) = 0$ for every $g \in X_n$ which is impossible since, for example, $\mathcal{B}_0(\mu) \neq 0$ for any $\mu \in \mathbf{D}$.

As will be illustrated in what follows, the reproducing kernel is enormously useful in the study of the approximating properties of the linear span $\{\mathcal{B}_0, \ldots, \mathcal{B}_{n-1}\}$. As a preluding example, in the linear prediction context mentioned in the introduction, consider the problem of finding the nth order, mean square optimal, one step ahead predictor $\varphi_n(z) \in X_n$ of a wide-sense stationary process with spectral density $f(\omega)$. Here z is interpreted as the backward shift operator so that if $\{u_k\}$ is a sequence in ℓ_2, then $\{\varphi_n(z)u_k\}$ denotes a filtered version of that sequence. With this notation in hand φ_n is given by

$$\varphi_n = \arg\min_{\varphi \in X_n} \|1 - \varphi\| \quad \text{subject to } \varphi(0) = 0.$$

The constraint is added to ensure the one step ahead nature of the predictor, and the norm is induced by the inner product (2.1) modified so as to be weighted according to the spectral density $f(\omega)$. This constrained optimization problem is easily solved using the reproducing kernel $K_n(z,\mu)$ associated with $\operatorname{Span}\{1, \mathcal{B}_0, \ldots, \mathcal{B}_{n-1}\}$ and with respect to the modified inner product by first noting that via the Cauchy–Schwarz inequality

$$\begin{aligned} 1 = |1 - \varphi_n(0)|^2 &= |\langle 1 - \varphi_n, K_n(z,0)\rangle|^2 \\ &\leq \langle 1 - \varphi_n, 1 - \varphi_n\rangle \langle K_n(z,0), K_n(z,0)\rangle \\ &= \|1 - \varphi_n\|^2 K_n(0,0). \end{aligned}$$

However, equality occurs in the Cauchy–Schwarz inequality if and only if $1 - \varphi_n(z) = cK(z,0)$ for some constant c. The constraint $\varphi_n(0) = 0$ implies the choice $c = 1/K_n(0,0)$ which leads to the solution

$$\varphi_n(z) = 1 - \frac{K_n(z,0)}{K_n(0,0)}.$$

Given this utility of the reproducing kernel $K_n(z, \mu)$, the natural question of calculating it arises. This may be easily achieved as

$$K_n(z, \mu) = \sum_{k=0}^{n-1} \mathcal{B}_k(z)\overline{\mathcal{B}_k(\mu)}. \tag{3.1}$$

That this formulation is valid may be quickly checked by noting that for any $0 \le m < n$

$$\langle \mathcal{B}_m(z), K_n(z, \mu)\rangle = \sum_{k=0}^{n-1} \mathcal{B}_k(\mu)\langle \mathcal{B}_m(z), \mathcal{B}_k(z)\rangle = \mathcal{B}_m(\mu).$$

However, for the purposes of the analysis in this paper this representation is too cumbersome, and a more succinct description is required. This is in common with the study of orthogonal polynomials [45, 11] via the use of reproducing kernels, where simpler closed form formulae for $K_n(z, \mu)$ are called "Christoffel–Darboux formulae." Borrowing from this literature, the following theorem presents a Christoffel–Darboux formula for $K_n(z, \mu)$ which in what follows will be central to the derivation of the generalized Fourier and Toeplitz matrix results of this paper.

THEOREM 3.1 (Christoffel–Darboux formula). *Define the modified Blaschke product*

$$\varphi_n(z) \triangleq \prod_{k=0}^{n-1} \frac{z - \overline{\xi_k}}{1 - \xi_k z}.$$

Then the reproducing kernel of the space spanned by $\{\mathcal{B}_0, \mathcal{B}_1, \ldots, \mathcal{B}_{n-1}\}$ can be expressed as

$$K_n(z, \mu) = \frac{1 - \overline{\varphi_n(\mu)}\varphi_n(z)}{1 - z\overline{\mu}}. \tag{3.2}$$

Proof. Take $z, \mu \in \mathbf{D}$, consider the function $\Lambda_n(z, \mu) : \mathbf{C} \times \mathbf{C} \to \mathbf{C}$ defined by

$$\Lambda_n(z, \mu) \triangleq \frac{1 - \overline{\varphi_n(\mu)}\varphi_n(z)}{1 - z\overline{\mu}},$$

and define the space $X_n \triangleq \operatorname{Span}\{\mathcal{B}_0(z), \ldots, \mathcal{B}_{n-1}(z)\} \subset H_2(\mathbf{T})$. Clearly, since the product $\overline{\varphi_n(\mu)}\varphi_n(1/\overline{\mu}) = 1$, then $1 - \overline{\varphi_n(\mu)}\varphi_n(z)$ possesses a zero at $z = 1/\overline{\mu}$ so that $\Lambda_n(z, \mu) \in X_n$. Furthermore, by Cauchy's integral theorem

$$\left\langle \mathcal{B}_m(z), \frac{1}{1 - \overline{\mu}z} \right\rangle = \frac{1}{2\pi j} \oint_{\mathbf{T}} \frac{\mathcal{B}_m(z)}{z - \mu}\, \mathrm{d}z = \mathcal{B}_m(\mu),$$

and also, for any $m = 0, 1, \ldots, n-1$, by the change of integration variable $z \mapsto 1/z$ and Cauchy's integral theorem

$$\begin{aligned}\left\langle \mathcal{B}_m(z), \frac{\overline{\varphi_n(\mu)}\varphi_n(z)}{1 - \overline{\mu}z} \right\rangle &= \frac{\varphi_n(\mu)}{2\pi j} \oint_{\mathbf{T}} \frac{\mathcal{B}_m(z)\overline{\varphi_n(z)}}{z - \mu}\, \mathrm{d}z \\ &= \frac{\varphi_n(\mu)}{2\pi j} \oint_{\mathbf{T}} \frac{\mathcal{B}_m(1/z)\overline{\varphi_n(1/z)}}{1 - \mu z}\, \mathrm{d}z = 0.\end{aligned}$$

Therefore, $\Lambda_n(z,\mu)$ given by (3.2) has the property that $f(\mu) = \langle f(z), \Lambda_n(z,\mu)\rangle$ for any $f \in X_n$. However, the reproducing kernel $K_n(z,\mu)$ is the unique function in X_n with this property, so it must be that $K_n(z,\mu) = \Lambda_n(z,\mu)$. □

Often the expression (3.2) will be used by setting $\mu = re^{j\sigma}, z = re^{j\omega}$ and letting $r \to 1$ from below. In this case, with some abuse of notation in the interests of cleanliness of exposition, the theorem will be used in the form

$$K_n(\omega,\sigma) = \frac{1 - \overline{\varphi_n(e^{j\sigma})}\varphi_n(e^{j\omega})}{1 - e^{j(\sigma-\omega)}}. \tag{3.3}$$

4. Generalized Fourier series convergence. Given a function $f \in H_2(\mathbf{T})$, an obvious way of approximating it in terms of the basis functions $\{\mathcal{B}_0, \mathcal{B}_1, \ldots, \mathcal{B}_{n-1}\}$ is as f_n given by

$$f_n(z) = \operatorname*{arg\,min}_{g\in X_n} \|f - g\| = \sum_{k=0}^{n-1} \langle f, \mathcal{B}_k\rangle \mathcal{B}_k(z). \tag{4.1}$$

Provided $\sum(1 - |\xi_k|) = \infty$ holds, by the completeness theorem, Theorem 2.1, the approximation error $\|f_n - f\|$ can then be made arbitrarily small for arbitrarily large approximation order n.

A natural question to then ask is how the approximant f_n behaves with respect to other norms, for example, the supremum norm on $[-\pi,\pi]$. The purpose of this section is to show that a modified approximant, closely related to the above one and deriving from the Cesàro (or Fejér) mean of classical Fourier analysis, is also supremum norm convergent to f under the same condition of $\sum(1 - |\xi_k|) = \infty$. This result will encompass the classical result for the trigonometric basis by simply setting all the poles $\{\xi_k\}$ to zero.

To proceed, it is expedient to revisit the classical case by indeed setting $\xi_k = 0$ in (4.1) and also temporarily shifting to the $L_2(\mathbf{T})$ setting so that the sum in (4.1) becomes two-sided to obtain

$$f_n(\omega) = \frac{1}{2\pi}\sum_{k=-n}^{n} e^{jk\omega}\int_{-\pi}^{\pi} f(\sigma)e^{-j\sigma k}\,\mathrm{d}\sigma = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\sigma)D_n(\omega-\sigma)\,\mathrm{d}\sigma, \tag{4.2}$$

where

$$D_n(\theta) \triangleq \frac{\sin(2n+1)\theta/2}{\sin\theta/2}$$

is known [10] as the "Dirichlet kernel." Perhaps one of the more surprising facets of applied mathematics is that even if f is continuous on $[-\pi,\pi]$, then $\lim_{n\to\infty}|f_n - f| = 0$ uniformly on $[-\pi,\pi]$ is not guaranteed; this is a century-old observation due to Du Bois-Reymond [24]. This undesirable behavior stems from the fact that although from (4.2) one would wish $D_n(\theta)$ to behave more and more like a Dirac delta function as n increases, it does not in the sense that

$$\liminf_{n\to\infty}\int_{|\theta|>\rho} |D_n(\theta)|\,\mathrm{d}\theta \neq 0$$

for arbitrarily small ρ. In fact, the quantity $\|D_n\|_1$ (called the nth Lebesgue constant) is known to be bounded below by $(4/\pi^2)\log n$, so that since the norm of any linear projection $L_n : C[-\pi, \pi] \to \mathrm{Span}\{e^{-j\omega n}, \ldots, e^{-j\omega}, 1, e^{j\omega}, \ldots, e^{j\omega n}\}$ is known [4] to be underbounded by $\|D_n\|_1$, then by Du Bois-Reymond's result there always exists an $f \in C[-\pi, \pi]$ such that $\|L_n f\|$ becomes unbounded as $n \to \infty$. In fact, this difficulty has been the genesis of much work in the system identification literature, of which [33] offers a survey.

A remedy for this problem of nonconvergence is to replace the approximation (4.2) with the so-called Cesàro mean defined on the unit circle by

$$f_n(\omega) = \sum_{k=-n}^{n} \left(1 - \frac{|k|}{n}\right) \langle f, e^{jk\omega} \rangle e^{jk\omega} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\sigma) K_n(\omega - \sigma)\, \mathrm{d}\sigma, \tag{4.3}$$

where now

$$K_n(\theta) \triangleq \frac{\sin^2 n\theta/2}{n \sin^2 \theta/2}$$

is known [10] as the "Fejér kernel" and does possess the "delta-like" property

$$\lim_{n\to\infty} \int_{|\theta|>\rho} K_n(\theta)\, \mathrm{d}\theta = 0 \quad \forall \rho > 0 \tag{4.4}$$

so that since it is also true that

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} K_n(\theta)\, \mathrm{d}\theta = 1,$$

then

$$2\pi |f_n(\omega) - f(\omega)| \le \int_{|\omega-\sigma|\le\rho} |f_n(\sigma) - f(\omega)| K_n(\sigma)\, \mathrm{d}\sigma + \int_{|\omega-\sigma|>\rho} |f_n(\sigma) - f(\omega)| K_n(\sigma)\, \mathrm{d}\sigma$$

in which case if f is continuous, then use of (4.4) allows the conclusion that since ρ may be made arbitrarily small, then $\lim_{n\to\infty} |f_n(\omega) - f(\omega)| = 0$ uniformly on $[-\pi, \pi]$.

To develop the analogue of this result for the general case of approximants formed using the general orthonormal basis (1.5), it is necessary to first develop a generalization of the Cesàro mean (4.3). This may be accomplished by the definition

$$f_n(\omega) \triangleq \frac{\Gamma_n^\star(\omega) M_n(f) \Gamma_n(\omega)}{K_n(\omega, \omega)}, \tag{4.5}$$

where Γ_n defined in (1.7) is an $n \times 1$ vector of general rational orthonormal basis functions (1.5) and $M_n(f)$ is a generalized Toeplitz matrix as defined in (1.6). If all the poles $\{\xi_k\}$ are set to zero in (3.1), then it is straightforward to verify that the formulation (4.5) reduces (since in this case $M_n(f) = T_n(f)$) to the usual Cesàro mean (4.3). To analyze the convergence properties of (4.5), note that by the formulation (3.1)

$$\begin{aligned}
\Gamma_n^\star(\omega) M_n(f) \Gamma_n(\omega) &= \sum_{m=0}^{n-1} \sum_{n=0}^{n-1} \overline{\mathcal{B}_m(e^{j\omega})} \mathcal{B}_n(e^{j\omega}) [M_n(f)]_{m,n} \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\sigma) \sum_{m=0}^{n-1} \sum_{n=0}^{n-1} \overline{\mathcal{B}_m(e^{j\omega})} \mathcal{B}_n(e^{j\omega}) \mathcal{B}_m(e^{j\sigma}) \overline{\mathcal{B}_n(e^{j\sigma})}\, \mathrm{d}\sigma \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\sigma) \left| K_n(\omega, \sigma) \right|^2 \mathrm{d}\sigma.
\end{aligned}$$

Therefore, since by the defining property of the reproducing kernel

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}|K_n(\omega,\sigma)|^2\,\mathrm{d}\sigma = \langle K_n(\omega,\sigma),K_n(\omega,\sigma)\rangle = K_n(\omega,\omega) = \sum_{m=0}^{n-1}|\mathcal{B}_m(e^{j\omega})|^2, \tag{4.6}$$

then

$$\begin{aligned}\frac{1}{2\pi}\left|\frac{\Gamma_n^\star(\omega)M_n(f)\Gamma_n(\omega)}{K_n(\omega,\omega)} - f(\omega)\right| &= \frac{1}{2\pi K_n(\omega,\omega)}\left|\Gamma_n^\star(\omega)M_n(f)\Gamma_n(\omega) - K_n(\omega,\omega)f(\omega)\right|\\ &= \frac{1}{2\pi K_n(\omega,\omega)}\left|\int_{-\pi}^{\pi}[f(\sigma)-f(\omega)]\,|K_n(\omega,\sigma)|^2\,\mathrm{d}\sigma\right| \qquad (4.7)\end{aligned}$$

and so, in analogy with classical Fourier analysis, convergence of the generalized Cesàro mean approximant (4.5) hinges on a kernel function (in this case depending on the reproducing kernel and being given by $|K_n(\omega,\sigma)|^2/K_n(\omega,\omega)$) behaving in some sense like the Dirac delta function $\delta(\omega-\sigma)$. Via use of the Christoffel–Darboux formula (3.3) for the reproducing kernel, it is possible to establish that this "delta-like" behavior does in fact occur in the following sense.

LEMMA 4.1. *For any $\rho > 0$ and provided*

$$\sum_{k=0}^{\infty}(1-|\xi_k|) = \infty,$$

then

$$\lim_{n\to\infty}\frac{1}{K_n(\omega,\omega)}\int_{\sigma\notin[\omega-\rho,\omega+\rho]}|K_n(\omega,\sigma)|^2\,\mathrm{d}\sigma = 0.$$

Proof. By Theorem 3.1 formulated as (3.3)

$$|K_n(\omega,\sigma)| \le \frac{2}{|e^{j(\sigma-\omega)}-1|} = \frac{1}{|\sin(\sigma-\omega)/2|}$$

so that

$$\int_{\sigma\notin[\omega-\rho,\omega+\rho]}|K_n(\omega,\sigma)|^2\,\mathrm{d}\sigma \le \frac{2\pi}{\sin^2\rho/2}.$$

Also, since $|1-\xi_k e^{j\omega}| \le 1+|\xi_k|$, then

$$K_n(\omega,\omega) = \sum_{k=0}^{n-1}\frac{1-|\xi_k|^2}{|1-\xi_k e^{j\omega}|^2} \ge \frac{1}{2}\sum_{k=0}^{n-1}(1-|\xi_k|) \tag{4.8}$$

so that

$$\frac{1}{K_n(\omega,\omega)}\int_{\sigma\notin[\omega-\rho,\omega+\rho]}|K_n(\omega,\sigma)|^2\,\mathrm{d}\sigma \le \frac{\pi}{\sin^2\rho/2}\left(\sum_{k=0}^{n-1}(1-|\xi_k|)\right)^{-1}$$

which tends to zero under the conditions of the lemma. □

Before using this result, some further notation is required since, motivated by the desire to provide results applicable to certain system theoretic problems, it is necessary to be more ambitious than to just prove convergence of f_n to f. Instead, it is also necessary to prove convergence to zero of the quadratic form

$$\frac{\Gamma_n^\star(\mu)M_n(f)\Gamma_n(\omega)}{K_n(\omega,\omega)}$$

when $\mu \neq \omega$, and this entails some delicacy in dealing with the distance between $e^{j\mu}$ and $e^{j\omega}$ on the unit circle, so that the distance between μ and ω must be considered modulo 2π. This is achieved by defining, for $-\pi \leq x, y \leq \pi$, the function $d(x,y) : [-\pi,\pi] \times [-\pi,\pi] \to [0,\pi]$ as

$$d(x,y) \triangleq \min(|x-y|, 2\pi - |x-y|). \tag{4.9}$$

Furthermore, letting Ω_1 and Ω_2 be subsets of $[-\pi,\pi]$, then

$$d(\Omega_1,\Omega_2) \triangleq \min_{x\in\Omega_1,\, y\in\Omega_2} d(x,y).$$

In what follows, it will also be useful to note that because

$$|\sin(x)| = |\sin(\pi - x)| = |\sin(-x)| = |\sin(-(\pi - x))|,$$

then

$$|\sin(x)| = \sin\frac{d(2x,0)}{2}, \quad -\pi \leq x \leq \pi$$

so that since $2x/\pi \leq \sin x \leq x$ for $0 \leq x \leq \pi/2$, then

$$\frac{2}{d(2x,0)} \leq \frac{1}{|\sin(x)|} \leq \frac{\pi}{d(2x,0)} \tag{4.10}$$

for any $x : |x| \leq \pi$, $x \neq 0$. The main use for these latter ideas is to develop further bounds such as the ones contained in the following lemma, which sharpen the interpretation (already begun in Lemma 4.1) of $K_n(\omega,\sigma)$ as behaving approximately like the Dirac delta $\delta(\omega,\sigma)$.

LEMMA 4.2. *Suppose that $|\xi_n| \leq 1 - \delta$ for some $\delta > 0$ and all n. Then for n large enough the following bounds apply:*

$$\frac{1}{2}\sum_{k=0}^{n-1}(1 - |\xi_k|) \leq |K_n(\omega,\sigma)| \leq \begin{cases} \dfrac{2n}{\delta} & \forall \sigma,\omega, \\[2ex] \dfrac{1}{|\sin(\omega-\sigma)/2|}, & \omega \neq \sigma. \end{cases} \tag{4.11}$$

Proof. Consider first the case of $\omega = \sigma$. Then by the expression (4.6) and the formulation (1.5)

$$K_n(\omega,\omega) = \sum_{k=0}^{n-1}\frac{1-|\xi_k|^2}{|1-\xi_k e^{j\omega}|^2} \leq \sum_{k=0}^{n-1}\frac{1+|\xi_k|}{1-|\xi_k|} \leq \frac{2}{\delta}n \tag{4.12}$$

so that using the Cauchy–Schwarz inequality the bound

$$|K_n(\omega,\sigma)| \leq \sqrt{|K_n(\omega,\omega)|}\ \sqrt{|K_n(\sigma,\sigma)|} \leq \frac{2}{\delta}n \tag{4.13}$$

applies for all ω, σ. For the case of $\omega \neq \sigma$, using (3.3) derived from Theorem 3.1 then leads to the bound

$$|K_n(\omega, \sigma)| = \frac{1}{\left|e^{j(\omega-\sigma)} - 1\right|} \leq \frac{1}{|\sin(\omega - \sigma)/2|}.$$

Finally, the underbound has already been established in (4.8). □

With these ideas in hand, the following result is available.

THEOREM 4.3. *Suppose $f(\omega)$ is a continuous, not necessarily real-valued function on $[-\pi, \pi]$. Then provided*

$$\sum_{k=0}^{\infty}(1 - |\xi_k|) = \infty$$

the following limit result holds:

$$\lim_{n\to\infty} \frac{\Gamma_n^\star(\omega) M_n(f) \Gamma_n(\omega)}{K_n(\omega, \omega)} = f(\omega)$$

uniformly in ω on $[-\pi, \pi]$. Under the strengthened condition that $|\xi_n| \leq 1 - \delta$ for some $\delta > 0$ and all n, then for $\mu \neq \omega$

$$\lim_{n\to\infty} \frac{\Gamma_n^\star(\mu) M_n(f) \Gamma_n(\omega)}{K_n(\omega, \omega)} = 0.$$

Proof. Consider first the case of $\mu = \omega$. Then from (4.7) and for arbitrary $\rho > 0$

$$\begin{aligned}\frac{1}{2\pi}\left|\frac{\Gamma_n^\star(\omega) M_n(f) \Gamma_n(\omega)}{K_n(\omega, \omega)} - f(\omega)\right| &= \frac{1}{2\pi K_n(\omega, \omega)}\left|\int_{-\pi}^{\pi} [f(\sigma) - f(\omega)]\, |K_n(\omega, \sigma)|^2 \,\mathrm{d}\sigma\right| \\ \leq \frac{1}{2\pi K_n(\omega, \omega)}&\left|\int_{\sigma\in[\omega-\rho,\omega+\rho]} [f(\sigma) - f(\omega)]\, |K_n(\omega, \sigma)|^2 \,\mathrm{d}\sigma\right| \\ + \frac{1}{2\pi K_n(\omega, \omega)}&\left|\int_{\sigma\notin[\omega-\rho,\omega+\rho]} [f(\sigma) - f(\omega)]\, |K_n(\omega, \sigma)|^2 \,\mathrm{d}\sigma\right|.\end{aligned}$$

Now, $f(\omega)$ is continuous, so for ρ sufficiently small

$$|f(\sigma) - f(\omega)| \leq \epsilon \text{ on } [\omega - \rho, \omega + \rho].$$

Using this and (4.6) gives that for sufficiently small ρ

$$\begin{aligned}\frac{1}{2\pi K_n(\omega, \omega)}\left|\int_{\sigma\in[\omega-\rho,\omega+\rho]} [f(\sigma) - f(\omega)]\, |K_n(\omega, \sigma)|^2 \,\mathrm{d}\sigma\right| &\leq \frac{\epsilon}{2\pi K_n(\omega, \omega)} \int_{-\pi}^{\pi} |K_n(\omega, \sigma)|^2 \,\mathrm{d}\sigma \\ &= \epsilon.\end{aligned}$$

Also, since f is continuous on compact $[-\pi, \pi]$, then $|f|$ is bounded by some $M/2 < \infty$. Therefore

$$\begin{aligned}&\frac{1}{2\pi K_n(\omega, \omega)}\left|\int_{\sigma\notin[\omega-\rho,\omega+\rho]} [f(\sigma) - f(\omega)]\, |K_n(\omega, \sigma)|^2 \,\mathrm{d}\sigma\right| \\ &\quad\leq \frac{M}{2\pi K_n(\omega, \omega)} \int_{\sigma\notin[\omega-\rho,\omega+\rho]} |K_n(\omega, \sigma)|^2 \,\mathrm{d}\sigma\end{aligned}$$

which provides

$$\frac{1}{2\pi}\left|\frac{\Gamma_n^\star(\omega)M_n(f)\Gamma_n(\omega)}{K_n(\omega,\omega)} - f(\omega)\right| \leq \epsilon + \frac{M}{2\pi K_n(\omega,\omega)}\int_{\sigma\notin[\omega-\rho,\omega+\rho]} |K_n(\omega,\sigma)|^2 \,\mathrm{d}\sigma.$$

Using Lemma 4.1 and the fact that ϵ is arbitrary then gives the result for $\mu = \omega$. Now consider the case $\mu \neq \omega$. Define the regions

$$\begin{aligned}
\Omega_1 &\triangleq \left\{\sigma \in [-\pi,\pi] : |\sigma-\mu| < K_n^{-\alpha}(\omega,\omega)\right\},\\
\Omega_2 &\triangleq \left\{\sigma \in [-\pi,\pi] : |\sigma-\omega| < K_n^{-\alpha}(\omega,\omega)\right\},\\
\Omega_3 &\triangleq \left\{\sigma \in [-\pi,\pi] : \sigma \notin \{\Omega_1\cup\Omega_2\}\right\},
\end{aligned}$$

where $\alpha \in (0,1/2)$ is arbitrary. In this case

$$\begin{aligned}
\left|\frac{\Gamma_n^\star(\mu)M_n(f)\Gamma_n(\omega)}{K_n(\omega,\omega)}\right| &= \left|\frac{1}{2\pi K_n(\omega,\omega)}\int_{-\pi}^{\pi} f(\sigma)K_n(\mu,\sigma)\overline{K_n(\omega,\sigma)}\,\mathrm{d}\sigma\right|\\
&\leq \frac{\|f\|_\infty}{2\pi K_n(\omega,\omega)}\int_{\Omega_1\cup\Omega_2\cup\Omega_3} |K_n(\mu,\sigma)\overline{K_n(\omega,\sigma)}|\,\mathrm{d}\sigma.
\end{aligned}$$

Consider the integrals over the various regions in turn. By Lemma 4.2 the bound $K_n(\omega,\omega) \geq 1/2\sum_{k=0}^{n-1}(1-|\xi_k|)$ holds, so that by the assumptions of the theorem, n can be taken large enough that Ω_1 and Ω_2 do not overlap. Assuming this to be the case, then $|\omega-\sigma| > K_n^{-\alpha}(\omega,\omega)$ on Ω_1 and hence using Lemma 4.2

$$|K_n(\omega,\sigma)| \leq \frac{1}{|\sin(\omega-\sigma)/2|} \leq \frac{\pi}{d^\alpha(\omega,\sigma)}, \quad \sigma \in \Omega_1.$$

Therefore, under the assumption that $|\xi_k| \leq 1-\delta, \delta > 0$, then by Lemma 4.2 $|K_n(\mu,\sigma)| \leq 2n/\delta$, so that assuming n is so large that $K_n^{-\alpha}(\omega,\omega) \leq d(\mu,\omega)/4$ gives, using Lemma B.1,

$$\int_{\Omega_1} |K_n(\mu,\sigma)\overline{K_n(\omega,\sigma)}|\,\mathrm{d}\sigma \leq \frac{2n}{\delta}\int_{\Omega_1}\frac{1}{|\sin(\omega-\sigma)/2|}\,\mathrm{d}\sigma \leq \frac{32\,n}{\delta K_n^\alpha(\omega,\omega)|\sin(\omega-\mu)/2|}.$$

Using an identical argument

$$\int_{\Omega_2} |K_n(\mu,\sigma)\overline{K_n(\omega,\sigma)}|\,\mathrm{d}\sigma \leq \frac{32\,n}{\delta K_n^\alpha(\omega,\omega)|\sin(\omega-\mu)/2|}.$$

Finally, by the definition of Ω_3 and Lemma 4.2

$$\int_{\Omega_3} |K_n(\mu,\sigma)\overline{K_n(\omega,\sigma)}|\,\mathrm{d}\sigma \leq \frac{2\pi}{\sin^2 K_n^{-\alpha}(\omega,\omega)/2} \leq 8\pi K_n^{2\alpha}(\omega,\omega).$$

Combining the bounds on the integrals over the various regions gives

$$\left|\frac{\Gamma_n^\star(\mu)M_n(f)\Gamma_n(\omega)}{K_n(\omega,\omega)}\right| \leq \frac{32n\|f\|_\infty}{\pi K_n^{1+\alpha}(\omega,\omega)|\sin(\omega-\mu)/2|} + \frac{4\|f\|_\infty}{K_n^{1-2\alpha}(\omega,\omega)}$$

which, according to the lower bound in Lemma 4.2 and since $\alpha \in (0,1/2)$, tends to zero as $n \to \infty$. □

5. Algebraic structure of generalized Toeplitz matrices. In applications [29, 31, 16, 30, 53, 27], the consideration of quadratic forms more complicated than (4.5) occur. In fact, what is of more interest are forms such as

$$\frac{\Gamma_n^\star(\omega) M_n(f) M_n(g) \Gamma_n(\omega)}{K_n(\omega,\omega)}.$$

In these aforementioned applications [29, 31, 16, 30, 53, 27], the underlying orthonormal basis is the trigonometric one $\{e^{j\omega n}\}$ in which case $M_n(f) = T_n(f)$ is a bonafide Toeplitz matrix for which classical results are at hand concerning their algebraic structure. Namely, following the notation defined in (1.3), the convenient property that $T_n(f)T_n(g) \sim T_n(fg)$ is assured [14, 50] (the meaning of the $\sim$ notation here is as described in conjunction with equation (1.3)).

The purpose of this section is to establish this same algebraic structure for the generalized Toeplitz matrices defined by (1.6), the classical results once again arising as the special case of $\xi_k = 0$ in (1.5). To begin, note that by the formulation (1.6)

$$\begin{aligned}[M_n(f)M_n(g)]_{m,\ell} &= \frac{1}{4\pi^2}\sum_{k=0}^{n-1}\int_{-\pi}^{\pi} \mathcal{B}_m(\omega)\overline{\mathcal{B}_k(\omega)} f(\omega)\,\mathrm{d}\omega \int_{-\pi}^{\pi} \overline{\mathcal{B}_\ell(\sigma)}\mathcal{B}_k(\sigma) g(\sigma)\,\mathrm{d}\sigma \\ &= \sum_{k=0}^{n-1} \langle \mathcal{B}_m f, \mathcal{B}_k\rangle \overline{\langle \mathcal{B}_\ell g, \mathcal{B}_k\rangle}\end{aligned}$$

and

$$[M_n(fg)]_{m,\ell} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{B}_m(\omega)\overline{\mathcal{B}_\ell(\omega)} f(\omega)\overline{g(\omega)}\,\mathrm{d}\omega = \langle \mathcal{B}_m f, \mathcal{B}_\ell g\rangle.$$

Therefore, by Parseval's theorem

$$\begin{aligned}\left|[M_n(f)M_n(g)]_{m,\ell} - [M_n(fg)]_{m,\ell}\right| &= \left|\sum_{k=0}^{n-1}\langle \mathcal{B}_m f, \mathcal{B}_k\rangle \overline{\langle \mathcal{B}_\ell g, \mathcal{B}_k\rangle} - \langle \mathcal{B}_m f, \mathcal{B}_\ell g\rangle\right| \\ &\le \|\mathcal{B}_m f - \widehat{f}_n\| \|\mathcal{B}_\ell g - \widehat{g}_n\|,\end{aligned}$$

where

$$\widehat{f}_n(z) = \sum_{k=0}^{n-1}\langle \mathcal{B}_m f, \mathcal{B}_k\rangle \mathcal{B}_k(z), \quad \widehat{g}_n(z) = \sum_{k=0}^{n-1}\langle \mathcal{B}_\ell g, \mathcal{B}_k\rangle \mathcal{B}_k(z)$$

so that if the Hilbert–Schmidt norm dependence of matrix equivalence (1.3), (1.4) is to be employed, then interest is centered on the behavior of the overbound

$$|M_n(f)M_n(g) - M_n(fg)| \le \frac{1}{n}\sum_{m=0}^{n-1}\sum_{\ell=0}^{n-1} \|\mathcal{B}_m f - \widehat{f}_n\| \|\mathcal{B}_\ell g - \widehat{g}_n\|. \tag{5.1}$$

Now, since $|\mathcal{B}_m|$ is bounded, then $\mathcal{B}_m f \in L_2(\mathbf{T})$; so since $\{\mathcal{B}_k\}$ is a basis for $H_2(\mathbf{T})$, since $\{z^{-k}\}, k > 0$, is a basis for the orthogonal complement $H_2(\mathbf{T})^\perp$ [19], and since $L_2 = H_2 \oplus H_2^\perp$, then $\mathcal{B}_m f$ can be expanded as

$$\mathcal{B}_m f = \sum_{k=0}^{\infty}\langle \mathcal{B}_m f, \mathcal{B}_k\rangle \mathcal{B}_k + \sum_{k=1}^{\infty}\langle \mathcal{B}_m f, z^{-k}\rangle z^{-k}$$

so that

$$\|\mathcal{B}_m f - \widehat{f}_n\|^2 = \sum_{k=n}^{\infty} |\langle \mathcal{B}_m f, \mathcal{B}_k\rangle|^2 + \sum_{k=1}^{\infty} \left|\langle \mathcal{B}_m f, z^{-k}\rangle\right|^2 \tag{5.2}$$

and the task then becomes to try to show that as n and m increase, the terms in these sums tend to zero sufficiently quickly. In the trigonometric case this is straightforward since, for example, $\langle \mathcal{B}_m f, \mathcal{B}_k\rangle$ becomes $\langle z^m f, z^k\rangle = \langle f, z^{k-m}\rangle$ which is the $k-m$th term in the Fourier expansion of f. Assuming f is sufficiently smooth that these Fourier components die at some exponential rate (say, $\eta^{|k-m|}$ with $|\eta| < 1$) then provides (with the same reasoning giving $|\langle z^m f, z^{-k}\rangle| = |\langle f, z^{-(m+k)}\rangle| \leq |\eta|^{m+k}$) $\|\mathcal{B}_m f - \widehat{f}_n\| \leq K(\eta^{n-m} + \eta^m)$ for some $K < \infty$ so that the sums in the overbound in (5.1) are convergent and hence (5.1) tends to zero with increasing n thereby establishing $T_n(f)T_n(g) \sim T_n(fg)$ as $n \to \infty$.

Generalizing this to the basis (1.5) is surprisingly more difficult. However, consider the simplifying assumption that both f and g have finite dimensional (say, nth order) spectral factors of the form $f(z) = H(z)H(1/z)$, where

$$H(z) = \sum_{r=0}^{\infty} h_r z^r, \quad h_r = \sum_{i=0}^{n-1} \gamma_i^r$$

with $|\gamma_i| < 1$ and where for expediency (but without loss of generality) it is assumed that the $\{\gamma_i\}$ are isolated. Then the $|\langle \mathcal{B}_m f, \mathcal{B}_k\rangle|$ term can be simply bounded by the calculation

$$\begin{aligned}
\langle \mathcal{B}_m f, \mathcal{B}_k\rangle &= \sum_{r=0}^{\infty}\sum_{\ell=0}^{\infty} h_r \overline{h_\ell} \langle \mathcal{B}_m \overline{\mathcal{B}_k} z^r, z^\ell\rangle \\
&= \sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\sum_{r=0}^{\infty}\sum_{\ell=r+1}^{\infty} \gamma_i^r \overline{\gamma_j^\ell} \langle \mathcal{B}_m \overline{\mathcal{B}_k} z^r, z^\ell\rangle \\
&= \sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\sum_{r=0}^{\infty} \gamma_i^r \overline{\gamma_j^r} \sum_{\ell=1}^{\infty} \overline{\gamma_j^\ell} \langle \mathcal{B}_m \overline{\mathcal{B}_k}, z^\ell\rangle \\
&= \sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\sum_{r=0}^{\infty} \gamma_i^r \overline{\gamma_j^r} \frac{\overline{\gamma_j}\sqrt{(1-|\xi_m|^2)(1-|\xi_k|^2)}}{(\overline{\gamma_j} - \overline{\xi_k})(1-\overline{\gamma_j}\xi_m)} \prod_{t=m+1}^{k} \left(\frac{\overline{\gamma_j} - \overline{\xi_t}}{1-\xi_t\overline{\gamma_j}}\right),
\end{aligned}$$

where in progressing to the last line it has been recognized that the inner sum in the second-to-last line is the evaluation at $z = \overline{\gamma_j}$ of a function $\mathcal{B}_m(z)\overline{\mathcal{B}_k(z)}$ with impulse response terms $\langle \mathcal{B}_m \overline{\mathcal{B}}_k, z^\ell\rangle$ and without loss of generality it has also been assumed that $k > m$. Therefore, since $|\gamma_j| < 1$ there exists $|\eta| < 1$, $K < \infty$ both independent of n such that (K is different in different parts of the following expressions)

$$\begin{aligned}
|\langle \mathcal{B}_m f, \mathcal{B}_k\rangle| &\leq K \sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\sum_{r=0}^{\infty} |\gamma_i|^r |\gamma_j|^r \eta^{k-m} \\
&= K\eta^{k-m} \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} \frac{1}{1-|\gamma_i\gamma_j|} \\
&\leq K\eta^{|k-m|},
\end{aligned}$$

where in the last line the fact that the same argument works for $k < m$ has been taken into account. Using the same method, one can also bound $|\langle \mathcal{B}_m f, z^{-k}\rangle|$ to arrive at $\|\mathcal{B}_m f - \widehat{f}\| \le K(\eta^{p-m} + \eta^m)$, and arguing as before for the trigonometric case, the desired algebraic result $M_n(f)M_n(g) \sim M_n(fg)$ can be extended to the case of generalized Toeplitz matrix (1.6).

In passing, on an intuitive level this extended result may be understood by noting that since the basis functions $\{\mathcal{B}_m\}$ contain order m all-pass factors φ_m as defined in Theorem 3.1, then they approximate the order m shift z^m in that they may be written as $\varphi_m(e^{j\omega}) = e^{j\psi_m(\omega)}$, where $\psi_m(\omega)$ is a monotonically nondecreasing function taking values between 0 and $2m\pi$.

Note also in passing that this simplified development illustrates why convergence in the strong matrix norm is not possible. Specifically, the bounds just developed are of the form such that the difference $|[M_n(f)M_n(g) - M_n(fg)]_{m,\ell}| \le K(\eta^{p-m} + \eta^m)(\eta^{p-\ell} + \eta^\ell)$ is a matrix with "corners" not tending to zero with increasing n, so that vectors exist that decay exponentially at rate η^k, have bounded norm, and when used in quadratic forms of the matrix difference produce nonzero results no matter how large n is.

Leaving this intuitive development aside, it may be considered that the assumption that f and g have finite dimensional spectral factorization is too restrictive. In this case a more general result may be offered, applicable to any Lipschitz continuous f and g, but at the expense of somewhat weakening the definition of equivalence over that discussed in section 1 to one in which two $n \times n$ matrices A_n and B_n are said to be asymptotically equivalent as $n \to \infty$ with notation $A_n \sim B_n$ as $n \to \infty$ if

$$\lim_{n\to\infty} \frac{\Gamma_n^\star(\omega)[A_n - B_n][A_n - B_n]^\star \Gamma_n(\omega)}{K_n(\omega,\omega)} = 0 \quad \forall \omega \in [-\pi, \pi].$$

Note that this refinement of the definition of matrix equivalence makes no difference for the system theoretic applications motivating this paper (see (7.4) and the accompanying discussion in section 7 following, or the work [37] for more detail on this point). With this definition in hand, the following result on the algebraic structure of generalized Toeplitz matrices is available.

THEOREM 5.1. *Consider two not necessarily real-valued functions f and g of which at least one of them is Lipschitz continuous of order $\varepsilon > 0$ and the other one bounded. Suppose that the poles $\{\xi_k\}$ of the basis functions $\{\mathcal{B}_k\}$ in* (1.5) *satisfy $|\xi_k| \le 1 - \delta$ for some $\delta > 0$. Then*

$$M_n(f)M_n(g) \sim M_n(fg) \quad \textit{as } n \to \infty$$

with convergence rate faster than $O(\log^4 n / n^{\varepsilon/(\varepsilon+2)})$ *as* $n \to \infty$.

Proof. Without loss of generality assume that g is Lipschitz continuous and that f is bounded. From the definition (1.6) of $M_n(f)$ and the formulation (3.1) of $K_n(\omega,\sigma)$ it follows that with the definition

$$\Delta_n(\omega) \triangleq [M_n(f)M_n(g) - M_n(fg)]\, \Gamma_n(\omega),$$

then with the representation (1.6) and the formulation (3.1) in mind

$$\begin{aligned}\Delta_n(\omega) = &\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \Gamma_n(\mu)\Gamma_n^\star(\mu) f(\mu)\,\mathrm{d}\mu\right)\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \Gamma_n(\sigma)\Gamma_n^\star(\sigma) g(\sigma)\,\mathrm{d}\sigma\right)\Gamma_n(\omega)\\ &- \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \Gamma_n(\mu)\Gamma_n^\star(\mu) f(\mu) g(\mu)\,\mathrm{d}\mu\right)\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \Gamma_n(\sigma)\Gamma_n^\star(\sigma)\,\mathrm{d}\sigma\right)\Gamma_n(\omega)\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\Gamma_n(\mu)f(\mu)\int_{-\pi}^{\pi}\Gamma_n^\star(\mu)\Gamma_n(\sigma)\Gamma_n^\star(\sigma)\Gamma_n(\omega)g(\sigma)\,\mathrm{d}\sigma\,\mathrm{d}\mu\\
&\quad -\frac{1}{4\pi^2}\int_{-\pi}^{\pi}\Gamma_n(\mu)f(\mu)\int_{-\pi}^{\pi}\Gamma_n^\star(\mu)\Gamma_n(\sigma)\Gamma_n^\star(\sigma)\Gamma_n(\omega)g(\mu)\,\mathrm{d}\sigma\,\mathrm{d}\mu\\
&= \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\Gamma_n(\mu)f(\mu)\int_{-\pi}^{\pi}K_n(\sigma,\mu)K_n(\omega,\sigma)[g(\sigma)-g(\mu)]\,\mathrm{d}\sigma\,\mathrm{d}\mu
\end{aligned}$$

or more compactly

$$\Delta_n(\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\Gamma_n(\mu)G_n(\omega,\mu)f(\mu)\,\mathrm{d}\mu, \tag{5.3}$$

where the following definition has been used:

$$G_n(\omega,\mu) \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi}K_n(\sigma,\mu)K_n(\omega,\sigma)[g(\sigma)-g(\mu)]\,\mathrm{d}\sigma. \tag{5.4}$$

Therefore, denoting when x is a vector the Euclidean norm of x as $\|x\|$, then

$$\|\Delta_n(\omega)\|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}\overline{G_n(\omega,\lambda)}H_n(\omega,\lambda)\overline{f(\lambda)}\,\mathrm{d}\lambda, \tag{5.5}$$

where

$$H_n(\omega,\lambda) \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi}G_n(\omega,\mu)K_n(\mu,\lambda)f(\mu)\,\mathrm{d}\mu. \tag{5.6}$$

However, by using Lemma 4.2

$$|K_n(\omega,\sigma)| \le \begin{cases} \dfrac{2n}{\delta} & \forall \sigma,\omega,\\[2ex] \dfrac{1}{|\sin(\omega-\sigma)/2|}, & \omega\neq\sigma, \end{cases} \tag{5.7}$$

so that Lemma A.1 may be applied to (5.4) with $f_n(\omega,\sigma) = K_n(\omega,\sigma)$, $g_n(\sigma,\mu) = K_n(\sigma,\mu)[g(\sigma)-g(\mu)]$ and $\beta=\gamma=1$ to conclude that

$$|G_n(\omega,\mu)| \le \begin{cases} Cn^{2/(2+\varepsilon)} & \forall \omega,\mu,\\[2ex] \dfrac{C}{|\sin(\omega-\mu)/2|}\log n, & \omega\neq\mu, \end{cases} \tag{5.8}$$

for n sufficiently large. Applying Lemma A.1 again, this time to (5.6) with the choices $f_n(\omega,\sigma) = K_n(\omega,\sigma)f(\omega)$, $g_n(\sigma,\mu) = G_n(\sigma,\mu)/\log n$ and $\beta = 1, \gamma = 2/(2+\varepsilon)$ provides

$$|H_n(\omega,\lambda)| \le \begin{cases} Cn^{2/(2+\varepsilon)}\log^2 n & \forall \omega,\lambda,\\[2ex] \dfrac{C}{|\sin(\omega-\mu)/2|}\log^2 n, & \omega\neq\lambda. \end{cases}$$

Applying Lemma A.1 a final time, this time to (5.5) with the choices $f_n(\omega,\sigma) = H_n(\omega,\sigma)/\log^2 n$ and $g_n(\sigma,\mu) = G_n(\sigma,\mu)/\log n$, $\beta = 1, \gamma = 2/(2+\varepsilon)$ provides

$$\|\Delta_n(\omega)\|^2 \le Cn^{2/(2+\varepsilon)}\log^4 n$$

for some $C < \infty$ and for n sufficiently large. Therefore, under the assumption that $|\xi_k| < 1 - \delta$ for some $\delta > 0$, then by (4.8) $K_n(\omega, \omega) \geq \kappa n$ for some $\kappa > 0$ so that for some $C < \infty$

$$\lim_{n\to\infty} \frac{\|\Delta_n(\omega)\|^2}{K_n(\omega,\omega)} \leq \lim_{n\to\infty} \frac{C \log^4 n}{n^{\varepsilon/(\varepsilon+2)}} = 0$$

for any $\varepsilon > 0$ and the theorem is proved. □

Again, something more than this result is actually required in system theoretic applications where one is often concerned with multiple products that also contain matrix inverses. Such cases may be handled by the following corollary to the preceding result. In what follows, matrix products are to be interpreted in a left-to-right fashion as $\prod_{k=1}^{n} A_k = A_1 A_2 \cdots A_n$.

COROLLARY 5.2. *Suppose that the family of possibly complex valued functions $\{f_k\}_{k=1}^m$ are all Lipschitz continuous of order $\varepsilon > 0$. Suppose that the poles $\{\xi_k\}$ of the basis functions $\{\mathcal{B}_k\}$ in* (1.5) *satisfy $|\xi_k| \leq 1 - \delta$ for some $\delta > 0$. Then with $\sigma_k = \pm 1$*

$$\prod_{k=1}^{m} M_n^{\sigma_k}(f_k) \sim M_n\left(\prod_{k=1}^{m} f_k^{\sigma_k}\right) \quad \textit{as } n \to \infty$$

with convergence rate faster than $O(\log^4 n / n^{\varepsilon/(\varepsilon+2)})$ as $n \to \infty$ and provided the functions $\{f_k\}$ are invertible where required by the values of σ_k.

Proof. An inductive argument will be used to obtain the result. Define the matrix difference

$$\Delta_n(m) \triangleq M_n\left(\prod_{k=1}^{m} f_k^{\sigma_k}\right) - \prod_{k=1}^{m} M_n^{\sigma_k}(f_k)$$

which may be reexpressed as

$$\Delta_n(m) = \Delta_n'(m) + \widetilde{\Delta}_n(m), \tag{5.9}$$

where

$$\Delta_n'(m) \triangleq M_n\left(\prod_{k=1}^{m} f_k^{\sigma_k}\right) - \prod_{k=1}^{m} M_n(f_k^{\sigma_k}),$$
$$\widetilde{\Delta}_n(m) \triangleq \prod_{k=1}^{m} M_n(f_k^{\sigma_k}) - \prod_{k=1}^{m} M_n^{\sigma_k}(f_k).$$

The terms $\Delta_n'(m)$ and $\widetilde{\Delta}_n(m)$ will be considered separately. First, note that

$$\Delta_n'(m) = \Delta_n'(m-1) M_n(f_m^{\sigma_m}) + \left[M_n\left(\prod_{k=1}^{m} f_k^{\sigma_k}\right) - M_n\left(\prod_{k=1}^{m-1} f_k^{\sigma_k}\right) M_n(f_m^{\sigma_m}) \right] \tag{5.10}$$

and by Theorem 5.1 with the substitution $f = f_m^{\sigma_m}$, $g = \prod_{k=1}^{m-1} f_k^{\sigma_k}$ the second term in the above expression obeys

$$M_n\left(\prod_{k=1}^{m} f_k^{\sigma_k}\right) \sim M_n\left(\prod_{k=1}^{m-1} f_k^{\sigma_k}\right) M_n(f_m^{\sigma_m}) \quad \text{as } n \to \infty. \tag{5.11}$$

Furthermore, denoting when A is a matrix the spectral norm of A as $\|A\|$, then by Lemma 6.1 $\|M_n(f)\| \leq \|f\|_\infty$ so that by the Cauchy–Schwarz inequality

$$\left|\frac{\Gamma_n^\star(\omega)\Delta_n'(m-1)M_n(f_m^{\sigma_m})\Gamma_n(\omega)}{K_n(\omega,\omega)}\right|^2 \leq \|f_m^{\sigma_m}\|_\infty^2 \left|\frac{\Gamma_n^\star(\omega)\Delta_n'(m-1)[\Delta_n'(m-1)]^\star\Gamma_n(\omega)}{K_n(\omega,\omega)}\right|$$

so that $\Delta_n'(m) \sim 0$ as $n \to \infty$ if $\Delta_n'(m-1) \sim 0$ as $n \to \infty$. However, $\Delta_n'(1) = 0$, so that by induction $\Delta_n'(m) \sim 0$ as $n \to \infty$ for any $m \geq 1$.

Now consider the term $\widetilde{\Delta}_n(m)$ and note that by labeling $k_1, \ldots, k_r$ as being the indices k for which $\sigma_k = -1$, then $\widetilde{\Delta}_n(m)$ may be written as (given that if the lower index of a matrix product is greater than the upper index, then the product is understood to be equal to the identity matrix)

$$\widetilde{\Delta}_n(m) = \sum_{\ell=1}^{r}\prod_{k=1}^{k_\ell-1} M_n^{\sigma_k}(f_k)M_n^{-1}(f_{k_\ell})\left[M_n(f_{k_\ell})M_n(f_{k_\ell}^{-1}) - I\right]\prod_{k=k_\ell+1}^{m} M_n(f_k^{\sigma_k}).$$

This expression may be decomposed as $\widetilde{\Delta}_n(m) = \Sigma_n(m) + \Lambda_n(m)$ with the definitions

$$\Sigma_n(m) \triangleq \sum_{\ell=1}^{r}\prod_{k=1}^{k_\ell-1} M_n^{\sigma_k}(f_k)M_n^{-1}(f_{k_\ell})\left[M_n(f_{k_\ell})M_n(f_{k_\ell}^{-1}) - I\right] M_n\left(\prod_{k=k_\ell+1}^{m} f_k^{\sigma_k}\right),$$

$$\Lambda_n(m) \triangleq \sum_{\ell=1}^{r}\prod_{k=1}^{k_\ell-1} M_n^{\sigma_k}(f_k)M_n^{-1}(f_{k_\ell})\left[M_n(f_{k_\ell})M_n(f_{k_\ell}^{-1}) - I\right] \times\left[\prod_{k=k_\ell+1}^{m} M_n(f_k^{\sigma_k}) - M_n\left(\prod_{k=k_\ell+1}^{m} f_k^{\sigma_k}\right)\right].$$

Dealing with $\Sigma_n(m)$ and $\Lambda_n(m)$ in turn, note that

$$\Sigma_n(m) = \sum_{\ell=1}^{r}\prod_{k=1}^{k_\ell-1} M_n^{\sigma_k}(f_k)M_n^{-1}(f_{k_\ell}) \times\left[M_n(f_{k_\ell})M_n(f_{k_\ell}^{-1})M_n\left(\prod_{k=k_\ell+1}^{m} f_k^{\sigma_k}\right) - M_n\left(\prod_{k=k_\ell+1}^{m} f_k^{\sigma_k}\right)\right]$$

and by Lemma 6.1, the submultiplicativity of the matrix norm, and the continuity and positive definiteness assumptions on the f_k

$$\left\|M_n^{-1}(f_{k_\ell})\prod_{k=1}^{k_\ell-1} M_n^{\sigma_k}(f_k)\right\| \leq \prod_{k=1}^{k_\ell}\|f_k^{\sigma_k}\|_\infty < \infty$$

so that since it has been shown inductively that $\Delta'(m) \sim 0$ as $n \to \infty$ for any $m \geq 1$, then by the Cauchy–Schwarz inequality $\Sigma_n(m) \sim 0$ as $n \to \infty$ for any $n \geq 1$. Finally, again notice that $\Delta_n'(m) \sim 0$ as $n \to \infty$ implies that

$$\prod_{k=k_\ell+1}^{m} M_n(f_k^{\sigma_k}) \sim M_n\left(\prod_{k=k_\ell+1}^{m} f_k^{\sigma_k}\right) \quad \text{as } n \to \infty$$

so that once again using Lemma 6.1, the Cauchy–Schwarz inequality, and the submultiplicativity of the matrix norm $\Lambda_n(m) \sim 0$ as $n \to \infty$, which completes the proof. □

Combining this corollary with Theorem 4.3 then provides a further corollary representing an extension of the generalized Fourier convergence of Theorem 4.3.

COROLLARY 5.3. *Suppose that the family of possibly complex valued functions* $\{f_k\}_{k=1}^m$ *are all Lipschitz continuous of order* $\varepsilon > 0$. *Suppose that the poles* $\{\xi_k\}$ *of the basis functions* $\{\mathcal{B}_k\}$ *in* (1.5) *satisfy* $|\xi_k| \leq 1 - \delta$ *for some* $\delta > 0$. *Then the following limit result holds:*

$$\lim_{n\to\infty} \frac{1}{K_n(\omega,\omega)} \Gamma_n^\star(\mu) \left(\prod_{k=1}^m M_n^{\sigma_k}(f_k) \right) \Gamma_m(\omega) = \begin{cases} \prod_{k=1}^m f_k^{\sigma_k}(\omega), & \mu = \omega, \\ 0, & \mu \neq \omega, \end{cases}$$

for any $\omega \in [-\pi, \pi]$ *and where* $\sigma_k = \pm 1$ *with the functions* $\{f_k\}$ *assumed to be invertible when required by the values of* σ_k.

Proof.

$$\Gamma_n^\star(\mu) \left(\prod_{k=1}^m M_n^{\sigma_k}(f_k) \right) \Gamma_m(\omega) = \Gamma_n^\star(\mu) M_n \left(\prod_{k=1}^m f_k^{\sigma_k} \right) \Gamma_n(\omega) + \Gamma_n^\star(\mu) \Delta_n \Gamma_n(\omega),$$

where

$$\Delta_n \triangleq \prod_{k=1}^m M_n^{\sigma_k}(f_k) - M_n \left(\prod_{k=1}^m f_k^{\sigma_k} \right).$$

Now, by the Cauchy–Schwarz inequality

$$\left| \frac{\Gamma_n^\star(\mu)\Delta_n\Gamma_n(\omega)}{K_n(\omega,\omega)} \right|^2 \leq \left| \frac{\Gamma_n^\star(\omega)\Delta_n\Delta_n^\star\Gamma_n(\omega)}{K_n(\omega,\omega)} \right| \left| \frac{K_n(\mu,\mu)}{K_n(\omega,\omega)} \right|.$$

However, by Lemma 4.2, the lower bound $K_n(\omega,\omega) \geq \delta n/2$ applies and the upper bound $K_n(\mu,\mu) \leq 2n/\delta$ also applies so that $|K_n(\mu,\mu)/K_n(\omega,\omega)| \leq 1/\delta^2 < \infty$ independently of n. As well, by Corollary 5.2

$$\Delta_n \sim 0 \quad \text{as } n \to \infty$$

so that

$$\lim_{n\to\infty} \left| \frac{\Gamma_n^\star(\mu)\Delta_n\Delta_n^\star\Gamma_n(\omega)}{K_n(\omega,\omega)} \right| = 0.$$

Use of Theorem 4.3 with the substitution $f = \prod_{k=1}^m f_k^{\sigma_k}$ then completes the proof. □

As a simple but important example of the utility of this corollary, it allows the conclusion that when all the poles $\{\xi_k\}$ are chosen in a closed subset of $\mathbf{D}$, then

$$\lim_{n\to\infty} \frac{\Gamma_n^\star(\omega) M_n^{-1}(f) \Gamma_n(\omega)}{K_n(\omega,\omega)} = \frac{1}{f(\omega)} \tag{5.12}$$

which has particular relevance to the study of reproducing kernels with respect to weighted inner products.

More specifically, the emphasis so far has been on studying the reproducing kernel $K_n(z,\mu)$ associated with the space $X_n = \text{Span}\{\mathcal{B}_0, \ldots, \mathcal{B}_{n-1}\}$ and with respect to the inner product (2.1). However, as was illustrated in section 3 via a linear prediction example, there is utility in examining a related kernel $K_n'(z,\mu)$ which is still associated with X_n, but exists with respect to an inner product which is (2.1) with the integrand weighted by a positive function $f(\omega)$. In this case, since

$$\langle \Gamma_n^\star(\mu) M_n^{-1}(f)\Gamma_n(z), \Gamma_n^T(z)\rangle = \Gamma_n^\star(\mu) M_n^{-1}(f)\langle \Gamma_n(z), \Gamma_n^T(z)\rangle = \Gamma_n^\star(\mu),$$

then $\langle \mathcal{B}_k(z), \Gamma_n^\star(\mu) M_n^{-1}(f)\Gamma_n(z)\rangle = \mathcal{B}_k(\mu)$ for every $k = 0, 1, \ldots, n-1$, and hence since $K_n'(z,\mu)$ is the unique function in X_n with this property, then in fact $K_n'(z,\mu) = \Gamma_n^\star(\mu) M_n^{-1}(f)\Gamma_n(z)$ and so the asymptotic result (5.12) provides a means for providing the closed form approximation $K_n(\omega,\omega)/f(\omega)$ for $K_n'(\omega,\omega)$.

As a concluding remark for this section, it should be noted that the only other work known to the authors which addresses issues of generalizing Fourier convergence and asymptotic Toeplitz matrix properties in a similar context to this paper is that of Gunnarsson and Ljung [15, 16], wherein the generalization involves matrices not necessarily being Toeplitz but at least being approximately so in some sense. The details of this are such that the generalized matrices (and hence generalized Fourier convergence) considered in [15, 16] are quite different from those of the form $M_n(f)$ considered here, which are not approximately Toeplitz in any sense except that the spectral formulation (1.6) is reminiscent of the classical Toeplitz one (1.1).

6. Spectral properties of generalized Toeplitz matrices. A typical system identification application of the orthonormal bases (1.5) would be to seek a parameter vector $\theta \in \mathbf{R}^n$ in order to model the input-output relationship between N samples of an observed input sequence $\{u_k\}$ and output sequence $\{y_k\}$ as [46, 47, 40, 36]

$$y_k = \sum_{n=0}^{m-1} \theta_n \mathcal{B}_n(q) u_k = \phi_k^T \theta, \quad k = 0, 1, \ldots, N-1, \tag{6.1}$$

where q is the backward time shift operator and

$$\phi_k^T \triangleq [\mathcal{B}_0(q)u_k, \mathcal{B}_1(q)u_k, \ldots, \mathcal{B}_{n-1}(q)u_k]$$

is a vector of filtered versions of the signal $\{u_k\}$, the filtering depending on the orthonormal basis functions chosen. The least-squares solution $\widehat{\theta}$ for θ is then given as

$$\widehat{\theta} = R_N^{-1} \frac{1}{N} \sum_{k=0}^{N-1} \phi_k y_k, \quad R_N \triangleq \frac{1}{N} \sum_{k=0}^{N-1} \phi_k \phi_k^T$$

provided that R_N exists. It is well known [13] that the numerical robustness of solving for $\widehat{\theta}$ is intimately related to the condition number $\kappa(R_N)$ of R_N. By Parseval's theorem, for large N the matrix R_N converges as [27]

$$\lim_{N\to\infty} R_N = M_n(f),$$

where f is the spectral density of the observed input $\{u_k\}$ which, if containing deterministic components, is defined in the sense of Wiener's generalized harmonic analysis [51] or the quasi-stationarity sense of Ljung [27]. From a numerical point of view there is therefore significant practical relevance in examining the spectrum of generalized Toeplitz matrices.

For the classical trigonometric case wherein $\xi_k = 0$ and $M_n(f)$ is in fact a bonafide Toeplitz matrix, it is well known [14, 50] that the eigenvalues of $M_n(f)$ may be bounded above and below by the maximum and minimum values of f. This result can be easily extended to the general case, the classical case again emerging as a special case by setting $\xi_k = 0$.

LEMMA 6.1. *For continuous and real-valued $f(\omega) > 0$ let $M_n(f)$ be defined by* (1.6)*. Then*

$$\min_{\omega\in[-\pi,\pi]} f(\omega) \le \lambda(M_n(f)) \le \max_{\omega\in[-\pi,\pi]} f(\omega),$$

where $\lambda(A)$ is any eigenvalue of the matrix A. In the case that f is complex valued, then the upper bound

$$|\lambda(M_n(f))| \le \max_{\omega\in[-\pi,\pi]} |f(\omega)|$$

applies.

Proof. Consider the case of real-valued $f > 0$ first and take $x \in \mathbf{R}^n$ arbitrary but such that $x^\star x = 1$. Then

$$\begin{aligned} x^\star M_n(f)x &= \frac{1}{2\pi}\sum_{r=0}^{n-1}\sum_{k=0}^{n-1}\overline{x_r}x_k\int_{-\pi}^{\pi}\mathcal{B}_r(e^{j\omega})\overline{\mathcal{B}_k(e^{j\omega})}f(\omega)\,\mathrm{d}\omega,\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\omega)\left|\sum_{r=0}^{n-1}x_r\mathcal{B}_r(e^{j\omega})\right|^2\mathrm{d}\omega,\\ &\le \max_{\omega\in[-\pi,\pi]}\frac{f(\omega)}{2\pi}\sum_{r=0}^{n-1}\sum_{k=0}^{n-1}\overline{x_r}x_k\int_{-\pi}^{\pi}\mathcal{B}_r(e^{j\omega})\overline{\mathcal{B}_k(e^{j\omega})}\,\mathrm{d}\omega\\ &= \max_{\omega\in[-\pi,\pi]} f(\omega). \end{aligned}$$

However, since $M_n(f)$ is symmetric positive definite then

$$\max_{x^\star x=1} x^\star M_n(f)x = \lambda_{\max}(M_n(f)).$$

Using a similar argument to underbound the eigenvalues of $M_n(f)$ then completes the first part of the lemma. For the case of complex valued f note that the upper bound can be generated as a slight modification to the above reasoning as

$$|x^\star M_n(f)x| \le \frac{1}{2\pi}\int_{-\pi}^{\pi}|f(\omega)|\left|\sum_{r=0}^{n-1}x_r\mathcal{B}_r(e^{j\omega})\right|^2\mathrm{d}\omega \le \max_{\omega\in[-\pi,\pi]}|f(\omega)|. \qquad \square$$

This result provides a guaranteed overbound

$$\kappa(M_n(f)) \triangleq \frac{\lambda_{\max}(M_n(f))}{\lambda_{\min}(M_n(f))} \le \frac{\max_{\omega\in[-\pi,\pi]} f(\omega)}{\min_{\omega\in[-\pi,\pi]} f(\omega)} \tag{6.2}$$

on the condition number $\kappa(M_n(f))$ governing the numerical robustness of the least-squares estimation problem. The existence of this overbound has been one of the prime motivators for the recent interest in the use of orthonormal bases such as (1.5) for system identification applications [46, 47, 18, 40].

It is then natural to examine how conservative the bound (6.2) is. Considering that numerical robustness is of greatest concern when the dimension n is large, it is not unreasonable to simplify the examination of conservatism by letting $n \to \infty$. This allows the following result showing that for large n the bounds in Lemma 6.1 are tight so that the condition number of $M_n(f)$ actually achieves the bound (6.2).

THEOREM 6.2. *Define for continuous $f(\omega) > 0$ the operator $M(f) : \{x_k\} \in \ell_2^+ \mapsto \{y_k\} \in \ell_2^+$ as*

$$M(f) \triangleq \lim_{n\to\infty} M_n(f),$$

where this is understood to mean that the infinite sequence $\{y_k\}$ is generated from the infinite sequence $\{x_k\}$ as a natural limit of how finite length n $\{y_k\}$ are generated from finite length n $\{x_k\}$ via matrix multiplication by $M_n(f)$. Specifically

$$y_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{B}_k(e^{j\omega})\overline{g(\omega)}\,\mathrm{d}\omega, \quad g(\omega) \triangleq f(\omega)\sum_{k=0}^{\infty} x_k \mathcal{B}_k(e^{j\omega}).$$

Then provided

$$\sum_{k=0}^{\infty}(1 - |\xi_k|) = \infty$$

it holds that

$$\lambda(M(f)) \triangleq \{\lambda \in \mathbf{C} : M(f) - \lambda I \text{ is not invertible}\} = \mathrm{Range}\{f(\omega)\}.$$

Proof. Take any $\mu \in [\min_\omega f(\omega), \max_\omega f(\omega)]$ and suppose $\mu \notin \lambda(M)$. Then since by orthogonality $M(1)$ is the identity operator, $M(f - \mu)$ is an invertible operator from $\ell_2^+ \to \ell_2^+$ so that in particular $\exists x \in \ell_2^+$ such that for $e_0 = (1, 0, 0, \ldots)$

$$x^T M(f - \mu) = e_0. \tag{6.3}$$

Therefore, defining $g(z) \in H_2(\mathbf{T})$ by $g(z) \triangleq (x_0 - 1)\mathcal{B}_0(z) + \sum_{k=1}^{\infty} x_k B_k(z)$ gives that from (6.3) $(f - \mu)g \perp \mathrm{Span}\{\mathcal{B}_k\}$. However, under the condition $\sum(1 - |\xi_k|) = \infty$ by Theorem 2.1, $\mathrm{Span}\{\mathcal{B}_k\} = H_2$ so that since $L_2 = H_2 \oplus H_2^\perp$ then $(f - \mu)\overline{g} \in H_2$. However, $g \in H_2$ by construction and the product of two H_2 functions is in H_1 [19]. Therefore $(f - \mu)|g|^2$ is a real-valued H_1 function, and the only such functions are constants [19]. However, since $\mu \in [\min_\omega f(\omega), \max_\omega f(\omega)]$, this function cannot be of constant sign; hence it cannot be a constant. This contradiction implies $[\min_\omega f(\omega), \max_\omega f(\omega)] \subset \lambda(M)$. Finally, Lemma 6.1 gives that $\lambda(M) \subset [\min_\omega f(\omega), \max_\omega f(\omega)]$. □

Again, this result represents an expansion to the case of generalized Toeplitz matrices (1.6) of results already known for conventional symmetric Toeplitz matrices [14, 50], the latter results being encompassed as a special case of Theorem 6.2. A practical conclusion arising from this theorem is that the numerical properties of the solution of (6.1) are governed solely by the spectral density f of $\{u_k\}$ and are independent of the particular orthonormal basis chosen in (6.1) via the selection of $\{\xi_k\}$.

Another conclusion arising from Theorem 6.2 is that in the previous results on the asymptotic algebraic structure of generalized Toeplitz matrices (a main conclusion of which was to conclude that $M_n^{-1}(f) \sim M_n(1/f)$ as $n \to \infty$) the assumptions imposed there of f being invertible cannot be weakened.

7. Applications. To put these results in context, this section presents a very brief outline of how they may be applied to certain system identification problems that have been previously alluded to. A much more detailed exposition of the issues raised here is contained in [37].

A result that has become of key importance to the intuitive understanding of various system identification methods [2, 12, 28, 17] is that the variability of the frequency response $G(e^{j\omega}, \widehat{\theta})$, of a model based on a least-squares estimate $\widehat{\theta}$ of an n dimensional parameter vector θ obtained from N observations of noise-corrupted input-output measurements, is approximately given by [31, 29]

$$\text{Var}\{G(e^{j\omega}, \widehat{\theta})\} \approx \frac{n}{N} \frac{\Phi_\nu(\omega)}{\Phi_u(\omega)}, \tag{7.1}$$

where $\Phi_u(\omega)$ is the spectral density of the observed input process and $\Phi_\nu(\omega)$ is the spectral density of the noise-corrupting process.

This is established in [29] for a wide range of model structures, unified by the requirement that a certain "shift" structure holds. In the interest of clarity, the discussion here will focus only on the case of measurement noise being white and the true system (including noise model) being encompassed by the model structure. In this case, it is established in [29] that with $\Phi_\nu(\omega) = \sigma^2$ being the constant white noise spectral density, then for large N

$$\frac{N}{n}\text{Var}\{G(e^{j\omega}, \widehat{\theta})\} \approx \frac{1}{n}\Gamma_n^\star(\omega) T_n^{-1}(\Phi_u) T_n(\sigma^2\Phi_u) T_n^{-1}(\Phi_u)\Gamma_n(\omega). \tag{7.2}$$

The reasoning of [29] then is that using the previously discussed classical results [14, 50] on the asymptotic algebraic structure of Toeplitz matrices, it can be argued that $T_n^{-1}(\Phi_u)T_n(\sigma^2\Phi_u)T_n^{-1}(\Phi_u) \approx T_n(\sigma^2/\Phi_u)$ so that the right-hand side of (7.2) is approximately an nth order Cesàro mean reconstruction of σ^2/Φ_u at frequency ω and hence should be approximately equal to $\sigma^2/\Phi_u(\omega)$.

However, it is very common, in the interest of concentrating estimation accuracy in certain frequency regions, to prefilter the measured data [27]. Identifying this filter with its transfer function $F(z)$ then allows the nature of the prefiltering to be characterized by the spectral density change of $\Phi_u(\omega) \mapsto |F(e^{j\omega})|^2\Phi_u(\omega)$. However, at the same time the use of data prefiltering implies a revision of the "noise model" as [29] $\Phi_\nu(\omega) \mapsto |F(e^{j\omega})|^2\Phi_\nu(\omega)$. In this case, the approximate estimation variability would be expected to be unchanged from (7.1) by reasoning that

$$\text{Var}\{G(e^{j\omega}, \widehat{\theta})\} \approx \frac{n}{N} \frac{|F(e^{j\omega})|^2\Phi_\nu(\omega)}{|F(e^{j\omega})|^2\Phi_u(\omega)} = \frac{n}{N}\frac{\Phi_\nu(\omega)}{\Phi_u(\omega)}. \tag{7.3}$$

However, as illustrated numerically in [37], the accuracy of this approximation depends very much on the relationship between the order of the filter F and the model order. The closer the two orders the more inaccurate the approximation.

In [37] this phenomenon is traced to the fact that as the filter order grows, the underlying Fourier reconstruction involved in (7.1) and hence (7.3) is with respect to a function with decreasing smoothness, and hence more terms (which grow with model order) are required in the Fourier expansion before convergence will approximately occur.

To circumvent this problem, a key observation of [37] is that the model can be reparameterized into one in which the prefilter poles are absorbed into the model structure. In this case (7.2) is replaced with

$$\frac{N}{K_n(\omega,\omega)}\mathrm{Var}\{G(e^{j\omega},\widehat{\theta})\} \approx \frac{\Gamma_n^\star(\omega)M_n^{-1}(\Phi_u)M_n(\sigma^2\Phi_u)M_n^{-1}(\Phi_u)\Gamma_n(\omega)}{K_n(\omega,\omega)}, \tag{7.4}$$

where the poles of the basis functions forming the generalized Toeplitz matrix $M_n(\Phi_u)$ are chosen the same as the prefilter. In this case, using Corollary 5.2 it can be argued that $M_n^{-1}(\Phi_u)M_n(\sigma^2\Phi_u)M_n^{-1}(\Phi_u) \sim M_n(\sigma^2/\Phi_u)$ as $n \to \infty$ so that

$$\frac{N}{K_n(\omega,\omega)}\mathrm{Var}\{G(e^{j\omega},\widehat{\theta})\} \approx \frac{1}{K_n(\omega,\omega)}\Gamma_n^\star(\omega)M_n(\sigma^2/\Phi_u)\Gamma_n(\omega) \tag{7.5}$$

which is a generalized Fourier reconstruction of a function σ^2/Φ_u that is *invariant* to the choice of prefilter. Therefore, using Theorem 4.3 to argue that (7.5) is approximately $\sigma^2/\Phi_u(\omega)$ the work [37] is able to suggest that

$$\mathrm{Var}\{G(e^{j\omega},\widehat{\theta})\} \approx \frac{K_n(\omega,\omega)}{N}\frac{\Phi_\nu(\omega)}{\Phi_u(\omega)} \tag{7.6}$$

is a more accurate approximation when all-pole prefiltering is employed and the ratio of model order to filter order is low. In other words, data prefiltering can affect the variability of the estimated model, and (7.6) quantifies how this occurs. The validity of these conclusions is illustrated numerically in [37]. As a final applications-oriented remark, note that when all poles $\{\xi_k\}$ are chosen at the origin (which corresponds to no prefiltering), then $K_n(\omega,\omega) = n$ so that the new approximation (7.6) becomes the "classical" one (7.1) as a special case.

8. Conclusion.

The purpose of this paper was to consider certain results in the study of Fourier series and Toeplitz matrices that have proved to be key to various system theoretic applications, and expand them to the case where the underlying orthonormal basis is not the classical trigonometric one but rather a rational formulation that encompasses the trigonometric basis as a special case. These results, and the ensuing generalizations developed in this paper, are summarized in Table 8.1.

One point worth clarifying is that in system theoretic settings for which these results will be applicable (control, signal processing, system identification) it is more common to associate the complex variable z with a forward time shift, rather than the backward shift association used here. This discrepancy is easily accommodated by simply transforming $z \mapsto 1/z$ in all the results presented here. A different basis function definition will result, which is in accordance with certain so-called Laguerre and Kautz bases studied in the control theory literature. However, the matrices $M_n(f)$ and the associated Fourier reconstruction formulas will be unchanged.

Appendix A. Bounds on integrals of kernel-like functions.

Throughout this appendix, C will denote a finite positive constant which may be different in different places of the same expression.

LEMMA A.1. *Let $f_n(\omega,\sigma) : [-\pi,\pi]\times[-\pi,\pi] \to \mathbf{C}$ be subject to*

$$|f_n(\omega,\sigma)| \le \begin{cases} Cn^\beta & \forall\omega,\sigma, \\ \dfrac{C}{|\sin(\omega-\sigma)/2|}, & \omega\neq\sigma, \end{cases} \tag{A.1}$$

TABLE 8.1
Summary of classical results and their relation to the generalizations derived here.

	Classical	Generalized
Basis	$e^{j\omega n}$	$\mathcal{B}_n(e^{j\omega}) \triangleq \frac{\sqrt{1-\lvert\xi_n\rvert^2}}{1-\xi_n e^{j\omega}} \prod_{k=0}^{n-1} \frac{e^{j\omega}-\overline{\xi_k}}{1-\xi_k e^{j\omega}}$
Completeness	$H_2(\mathbf{T})$	$H_2(\mathbf{T})$ provided $\sum(1-\lvert\xi_n\rvert)=\infty$
Assoc. matrix	Toeplitz matrix $[T_n(f)]_{k,\ell} = \int_{-\pi}^{\pi} e^{j\omega(k-\ell)} f(\omega)\,\frac{d\omega}{2\pi}$	Generalized Toeplitz matrix $[M_n(f)]_{k,\ell} = \int_{-\pi}^{\pi} \mathcal{B}_k(e^{j\omega})\overline{\mathcal{B}_\ell(e^{j\omega})} f(\omega)\,\frac{d\omega}{2\pi}$
Cesàro mean	$f_n(\omega) = \frac{1}{n}\sum_{k,\ell=0}^{n-1} e^{j\omega(\ell-k)}[T_n(f)]_{k,\ell}$	$f_n(\omega) = \sum_{k,\ell=0}^{n-1} \frac{\mathcal{B}_k(e^{j\omega})\overline{\mathcal{B}_\ell(e^{j\omega})}}{K_n(\omega,\omega)}[M_n(f)]_{k,\ell}$
Convergence	$\lim_{n\to\infty} \sup_{\omega\in[-\pi,\pi]} \lvert f(\omega)-f_n(\omega)\rvert = 0$	$\lim_{n\to\infty} \sup_{\omega\in[-\pi,\pi]} \lvert f(\omega)-f_n(\omega)\rvert = 0$
Def. equivalence $A_n \sim B_n$ as $n\to\infty$	$\lim_{n\to\infty} \lvert A_n - B_n\rvert = 0$	$\lim_{n\to\infty} \frac{\lVert(A_n-B_n)\Gamma_n(\omega)\rVert^2}{K_n(\omega,\omega)} = 0$
Algebraic properties	$T_n(f)T_n(g) \sim T_n(fg)$	$M_n(f)M_n(g) \sim M_n(fg)$
Extensions $\sigma_k = \pm 1$	$\prod_{k=1}^{m} T_n^{\sigma_k}(f_k) \sim T_n\left(\prod_{k=1}^{m} f_k^{\sigma_k}\right)$	$\prod_{k=1}^{m} M_n^{\sigma_k}(f_k) \sim M_n\left(\prod_{k=1}^{m} f_k^{\sigma_k}\right)$

for some $\beta \geq 0$ *and let* $g_n(\sigma,\mu) : [-\pi,\pi]\times[-\pi,\pi] \to \mathbf{C}$ *be subject to*

$$
|g_n(\sigma,\mu)| \leq \begin{cases} Cn^\gamma d^\varepsilon(\sigma,\mu) & \forall \mu,\sigma, \\ \dfrac{C}{|\sin(\mu-\sigma)/2|}, & \omega \neq \sigma, \end{cases} \tag{A.2}
$$

for some $\gamma,\varepsilon \geq 0$ *and where the meaning of* $d(\sigma,\mu)$ *is defined in* (4.9). *Then for* n *sufficiently large*

$$
\left|\int_{-\pi}^{\pi} f_n(\omega,\sigma) g_n(\sigma,\mu)\,d\sigma\right| \leq \begin{cases} C\min(n^\lambda, n^\delta \log n) & \forall \omega,\mu, \\ \dfrac{C}{|\sin(\omega-\mu)/2|}\log n, & \omega\neq\mu, \end{cases} \tag{A.3}
$$

where

$$
\lambda \triangleq \frac{\beta+\gamma}{2+\varepsilon}, \qquad \delta \triangleq \min(\beta,\gamma).
$$

Proof. Suppose, to begin with, that $\omega \neq \mu$. Then for any $\alpha > 0$, $d(\omega,\mu) \geq 6n^{-\alpha}$ for sufficiently large n. Assuming this is the case, define the following regions:

$$\begin{aligned}\Omega_1 &\triangleq \left\{\sigma \in [-\pi,\pi] : d(\omega,\sigma) < n^{-\alpha}\right\},\\ \Omega_2 &\triangleq \left\{\sigma \in [-\pi,\pi] : d(\mu,\sigma) \leq n^{-\alpha}\right\},\\ \Omega_3 &\triangleq \{\sigma \in [-\pi,\pi] : \sigma \notin \{\Omega_1 \cup \Omega_2\}\}.\end{aligned}$$

Therefore, using the assumed bounds on $f_n(\omega,\sigma), g_n(\sigma,\mu)$, noticing that by definition the regions Ω_1 and Ω_2 are disjoint, and using Lemma B.1 with $\varepsilon = 0$ lead to

$$\left|\int_{\Omega_1} f_n(\omega,\sigma)g_n(\sigma,\mu)\,\mathrm{d}\sigma\right| \leq \int_{\Omega_1} \frac{Cn^{\beta}}{|\sin(\mu-\sigma)/2|}\,\mathrm{d}\sigma \leq C\frac{n^{\beta-\alpha}}{|\sin(\mu-\omega)/2|}.$$

Similarly

$$\left|\int_{\Omega_2} f_n(\omega,\sigma)g_n(\sigma,\mu)\,\mathrm{d}\sigma\right| \leq \int_{\Omega_2} \frac{Cn^{\gamma}}{|\sin(\omega-\sigma)/2|}\,\mathrm{d}\sigma \leq C\frac{n^{\gamma-\alpha}}{|\sin(\mu-\omega)/2|}.$$

Finally, this time using Lemma B.2

$$\left|\int_{\Omega_3} f_n(\omega,\sigma)g_n(\sigma,\mu)\,\mathrm{d}\sigma\right| \leq \int_{\Omega_3} \frac{C}{|\sin(\sigma-\omega)/2\sin(\sigma-\mu)/2|}\,\mathrm{d}\sigma \leq \frac{C\alpha}{|\sin(\mu-\omega)|}\log n$$

so that for $\omega \neq \mu$, choosing $\alpha = \max(\beta,\gamma)$ provides the bound

$$\left|\int_{-\pi}^{\pi} f_n(\omega,\sigma)g_n(\sigma,\mu)\,\mathrm{d}\sigma\right| \leq \frac{C}{|\sin(\omega-\mu)/2|}\log n$$

for sufficiently large n. Now assume that $\omega = \mu$. In this case, again using the assumed bounds on $f_n(\omega,\sigma)$ and $g_n(\sigma,\mu)$,

$$\begin{aligned}\left|\int_{\Omega_1} f_n(\omega,\sigma)g_n(\sigma,\mu)\,\mathrm{d}\sigma\right| &\leq Cn^{\beta+\gamma}\int_{\Omega_1} d^{\varepsilon}(\sigma,\mu)\,\mathrm{d}\sigma \leq Cn^{\beta+\gamma}\int_0^{n^{-\alpha}} x^{\varepsilon}\,\mathrm{d}x\\ &= Cn^{\beta+\gamma-\alpha(\varepsilon+1)}.\end{aligned}$$

Also, using Lemma B.2

$$\left|\int_{\Omega_3} f_n(\omega,\sigma)g_n(\sigma,\mu)\,\mathrm{d}\sigma\right| \leq \int_{\Omega_3} \frac{C}{\sin^2(\sigma-\omega)/2}\,\mathrm{d}\sigma \leq \frac{C}{\sin n^{-\alpha}} \leq Cn^{\alpha},$$

so that when $\omega = \mu$ and hence $\Omega_1 = \Omega_2$, then for sufficiently large n

$$\left|\int_{-\pi}^{\pi} f_n(\omega,\sigma)g_n(\sigma,\mu)\,\mathrm{d}\sigma\right| \leq Cn^{\beta+\gamma-\alpha(\varepsilon+1)} + Cn^{\alpha}.$$

This bound is minimized (for large n) by the choice $\alpha = (\beta+\gamma)/(2+\varepsilon)$ as $Cn^{(\beta+\gamma)/(2+\varepsilon)}$. Alternatively, with the definition $\delta \triangleq \min(\beta,\gamma)$ the integral on Ω_3 can also be bounded using Lemma B.2 as

$$\left|\int_{\Omega_3} f_n(\omega,\sigma)g_n(\sigma,\mu)\,\mathrm{d}\sigma\right| \leq \int_{\Omega_3} \frac{Cn^{\delta}}{|\sin(\sigma-\omega)/2|}\,\mathrm{d}\sigma \leq C\alpha n^{\delta}\log n$$

to give the bound for $\omega = \mu$ and sufficiently large n of

$$\left|\int_{-\pi}^{\pi} f_n(\omega,\sigma)g_n(\sigma,\mu)\,\mathrm{d}\sigma\right| \leq Cn^{\beta+\gamma-\alpha(\varepsilon+1)} + C\alpha n^{\delta}\log n.$$

This bound is minimized (assuming without loss of generality that $\beta = \min(\beta,\gamma)$) by the choice (for large n) $\alpha = \gamma/(1+\varepsilon)$ as

$$\left|\int_{-\pi}^{\pi} f_n(\omega,\sigma)g_n(\sigma,\mu)\,\mathrm{d}\sigma\right| \leq Cn^{\delta}\log n.$$

Note that this latter bound will be smaller than the previous one whenever $\beta+\gamma > (2+\varepsilon)\min(\beta,\gamma)$ and for n sufficiently large. □

Appendix B. Integrals of reciprocals of sine functions.

LEMMA B.1. *Let* $0 < \alpha < \pi/8$ *and suppose that* $\omega,\mu \in [-\pi,\pi]$ *satisfies* $d(\mu,\omega) \geq 4\alpha$. *Then for any* $\Omega \subset [\mu-\alpha,\mu+\alpha]$, $\varepsilon \geq 0$ *and where the meaning of* $d(\mu,\omega)$ *is defined in* (4.9)

$$\int_{\Omega} \frac{d^{\varepsilon}(\sigma,\mu)}{|\sin(\sigma-\omega)/2|}\,\mathrm{d}\sigma \leq \frac{16\alpha^{\varepsilon+1}}{(\varepsilon+1)|\sin(\mu-\omega)/2|}.$$

Proof. It holds that

$$\begin{aligned}\sin\left(\frac{\mu-\omega}{2}+\frac{x}{2}\right) &= \cos\left(\frac{x}{2}\right)\sin\left(\frac{\mu-\omega}{2}\right) + \sin\left(\frac{x}{2}\right)\cos\left(\frac{\mu-\omega}{2}\right)\\ &= \cos\left(\frac{x}{2}\right)\sin\left(\frac{\mu-\omega}{2}\right)\left(1+\tan\left(\frac{x}{2}\right)\cot\left(\frac{\mu-\omega}{2}\right)\right),\end{aligned} \tag{B.1}$$

and for $x \in [-\alpha,\alpha]$

$$1+\tan\left(\frac{x}{2}\right)\cot\left(\frac{\mu-\omega}{2}\right) > 1 - \frac{\alpha/2}{\cos(\pi/8)}\,\frac{4}{d(\mu,\omega)} > 1/4. \tag{B.2}$$

Without loss of generality, assume Ω is such that $d(\sigma,\mu) = |\sigma-\mu|$ on Ω in which case the change of variables $x = \sigma-\mu$, together with (B.1)–(B.2), then gives

$$\begin{aligned}\int_{\Omega} \frac{d^{\varepsilon}(\sigma,\mu)}{|\sin(\sigma-\omega)/2|}\,\mathrm{d}\sigma &\leq \int_{-\alpha}^{\alpha} \frac{|x|^{\varepsilon}}{|\sin(\mu-\omega+x)/2|}\,\mathrm{d}x\\ &\leq \frac{1}{|\sin(\mu-\omega)/2|}\int_{-\alpha}^{\alpha} \frac{|x|^{\varepsilon}}{|\cos x/2|}\,\frac{1}{|1+\tan(x/2)\cot(\mu-\omega)/2|}\,\mathrm{d}x\\ &\leq \frac{1}{|\sin(\mu-\omega)/2|}\int_{-\alpha}^{\alpha} 2\,\times\,4|x|^{\varepsilon}\,\mathrm{d}x = \frac{16\alpha^{1+\varepsilon}}{(1+\varepsilon)|\sin(\mu-\omega)/2|}.\end{aligned}$$

□

LEMMA B.2. *Let* $-\pi \leq \alpha < \beta \leq \pi$ *and suppose that* $\omega \in [-\pi,\pi]$ *does not belong to* $[\alpha,\beta]$. *Then*

$$\int_{\alpha}^{\beta} \frac{1}{|\sin(\sigma-\omega)/2|}\,\mathrm{d}\sigma \leq 4\log\frac{8}{\gamma}, \tag{B.3}$$

where $\gamma \triangleq d(\omega, [\alpha, \beta])$*. Suppose also that* $\mu \in [-\pi, \pi]$ *does not belong to* $[\alpha, \beta]$*. Then*

$$\text{(B.4)}\quad \int_\alpha^\beta \frac{1}{|\sin(\sigma-\omega)/2\sin(\sigma-\mu)/2|}\,\mathrm{d}\sigma \leq \begin{cases} \dfrac{8}{|\sin(\mu-\omega)/2|}\log\dfrac{4}{\gamma}, & \omega \neq \mu, \\ \\ \dfrac{4}{\sin\gamma}, & \omega = \mu, \end{cases}$$

where in this latter case $\gamma \triangleq d(\{\omega, \mu\}, [\alpha, \beta])$*.*

Proof. To begin with assume that $\omega < \alpha$. The change of variables $x = (\sigma - \omega)/2$ then gives

$$\text{(B.5)}\qquad \int_\alpha^\beta \frac{1}{|\sin(\sigma-\omega)/2|}\,\mathrm{d}\sigma = 2\int_{\alpha'}^{\beta'} \frac{1}{\sin x}\,\mathrm{d}x,$$

where $0 < \alpha' \triangleq (\alpha-\omega)/2 < \beta' \triangleq (\beta-\omega)/2 < \pi$. Since $\sin(\beta'/2) > \sin(\alpha'/2)$ and $\cos(\alpha'/2) > \cos(\beta'/2)$ it follows from (B.5) that

$$\begin{aligned} \int_\alpha^\beta \frac{1}{|\sin(\sigma-\omega)/2|}\,\mathrm{d}\sigma &= 2\left[\log\left(\tan(x/2)\right)\right]_{\alpha'}^{\beta'}, \\ &= 2\log\left(\frac{\sin(\beta'/2)}{\sin(\alpha'/2)}\right) + 2\log\left(\frac{\cos(\alpha'/2)}{\cos(\beta'/2)}\right) \\ &\leq 2\log\left(\frac{1}{\sin(\alpha'/2)}\right) + 2\log\left(\frac{1}{\cos(\beta'/2)}\right) \\ &= 2\log\left(\frac{1}{\sin(\alpha'/2)}\right) + 2\log\left(\frac{1}{\sin(\pi/2-\beta'/2)}\right) \\ &\leq 2\log\left(\frac{4}{d(\alpha',0)}\right) + 2\log\left(\frac{4}{d(\pi-\beta',0)}\right) \\ &= 2\log\left(\frac{4}{d(\alpha',0)}\right) + 2\log\left(\frac{4}{d(\beta',0)}\right) \\ \text{(B.6)}\qquad &\leq 4\log\left(\frac{8}{\gamma}\right). \end{aligned}$$

The case where $\beta < \omega$ follows analogously, and the proof of the bound (B.3) is complete. Moving on to the proof of the bound (B.4), consider first the case $\omega \neq \mu$ and assume that

$$\text{(B.7)}\qquad -\pi \leq \omega < \mu < \alpha < \beta < \pi.$$

Let $0 < \alpha' \triangleq (\alpha-\mu)/2 < \beta' \triangleq (\beta-\mu)/2 < \pi$. The change of variables $x = (\sigma-\mu)/2$ then gives

$$\begin{aligned} &\int_\alpha^\beta \frac{1}{|\sin(\sigma-\omega)/2\sin(\sigma-\mu)/2|}\,\mathrm{d}\sigma \\ &= 2\int_{\alpha'}^{\beta'} \frac{1}{\sin\left((\mu-\omega)/2+x\right)\sin(x)}\,\mathrm{d}\sigma \\ &= 2\int_{\alpha'}^{\beta'} \frac{1}{\sin\left(\pi-(\mu-\omega)/2-x\right)\sin(x)}\,\mathrm{d}\sigma \end{aligned}$$

$$
\begin{aligned}
&= \frac{4}{\sin\left(\pi-(\mu-\omega)/2\right)}\left[\log\left(\frac{\sin(x)}{\sin\left(\pi-(\mu-\omega)/2-x\right)}\right)\right]_{\alpha'}^{\beta'}\\
&= \frac{4}{\sin(\mu-\omega)/2}\left(\log\left(\sin(\beta')\right) - \log\left(\sin\left(\pi-\frac{\mu-\omega}{2}-\beta'\right)\right)\right.\\
&\qquad\qquad \left. -\log\left(\sin(\alpha')\right) + \log\left(\sin\left(\pi-\frac{\mu-\omega}{2}-\alpha'\right)\right)\right)\\
&= \frac{4}{\sin(\mu-\omega)2}\left(\log\left(\sin\left(\frac{\beta-\mu}{2}\right)\right) - \log\left(\sin\left(\frac{\beta-\omega}{2}\right)\right)\right.\\
&\qquad\qquad \left. -\log\left(\sin\left(\frac{\alpha-\mu}{2}\right)\right) + \log\left(\sin\left(\frac{\alpha-\omega}{2}\right)\right)\right)\\
&\le \frac{4}{\sin(\mu-\omega)/2}\left(-\log\left(\sin\left(\frac{\beta-\omega}{2}\right)\right) - \log\left(\sin\left(\frac{\alpha-\mu}{2}\right)\right)\right)\\
&= \frac{4}{\sin(\mu-\omega)/2}\left(\log\left(\frac{1}{\sin(\beta-\omega)/2}\right) + \log\left(\frac{1}{\sin(\alpha-\mu)/2}\right)\right)\\
&\le \frac{4}{\sin(\mu-\omega)/2}\left(\log\left(\frac{4}{d(\beta,\omega)}\right) + \log\left(\frac{4}{d(\alpha,\mu)}\right)\right)
\end{aligned}
$$

$$
\le \frac{8}{\sin(\mu-\omega)/2}\log\left(\frac{4}{\gamma}\right), \tag{B.8}
$$

where use of (4.10) was made in the second-to-last inequality. This proves the lemma for the case (B.7). The other cases for $\omega \neq \mu$ follow analogously. Now suppose $\omega = \mu$. Let $0 < \alpha' \triangleq (\alpha-\omega)/2 < \beta' \triangleq (\beta-\omega)/2 < \pi$. Then the change of variables $x = (\sigma-\omega)/2$ gives

$$
\int_{\alpha}^{\beta} \frac{1}{\sin^2(\sigma-\omega)/2}\,\mathrm{d}\sigma = 2\int_{\alpha'}^{\beta'} \frac{1}{\sin^2 x}dx = \left[-2\,\frac{\cos x}{\sin x}\right]_{\alpha'}^{\beta'} \le \frac{4}{\sin\gamma} \tag{B.9}
$$

which proves the lemma when $\omega = \mu$. □

REFERENCES

[1] N. Aronszajn, *Theory of reproducing kernels*, Acta Math., (1950), pp. 337–404.

[2] R. Bitmead, M. Gevers, and V. Wertz, *Adaptive Optimal Control, The Thinking Man's GPC*, Prentice–Hall, Englewood Cliffs, NJ, 1990.

[3] P. W. Broome, *Discrete orthonormal sequences*, J. Assoc. Comput. Mach., 12 (1965), pp. 151–168.

[4] E. Cheney, *Introduction to Approximation Theory*, McGraw–Hill, New York, 1966.

[5] C. K. Chui, *Wavelets: A Tutorial in Theory and Applications*, Wavelet Analysis and Its Applications 2, Academic Press, Boston, 1992.

[6] P. R. Clement, *Laguerre functions in signal analysis and parameter identification*, J. Franklin Institute, 313 (1982), pp. 85–95.

[7] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Appl. Math., SIAM, Philadelphia, PA, 1992.

[8] P. DeWilde and H. Dym, *Schur recursions, error formulas, and convergence of rational estimators for stationary stochastic sequences*, IEEE Trans. Inform. Theory, IT-27 (1981), pp. 446–461.

[9] P. DeWilde, A. Vieira, and T. Kailath, *On a generalized Szegö–Levinson realization algorithm for optimal linear predictors based on a network synthesis approach*, IEEE Trans. Circuits Systems, CAS-25 (1978), pp. 663–675.

[10] R. Edwards, *Fourier Series, A Modern Introduction*, vol. 1, 2nd ed., Graduate Texts in Mathematics 64, Springer-Verlag, New York, 1979.

[11] L. Y. Geronimus, *Orthogonal Polynomials: Estimates, Asymptotic Formulas, and Series of Polynomials Orthogonal on the Unit Circle and on an Interval*, Consultants Bureau, New York, 1961. (Authorized translation from the Russian.)

[12] M. Gevers, L. Ljung, and P. M. J. Van den Hof, *Asymptotic variance expressions for closed–loop identification and their relevance in identification for control*, in Proc. 11th IFAC Symposium on System Identification, 1997, pp. 1449–1454.

[13] G. Golub and C. V. Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1989.

[14] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, University of California Press, Berkeley, CA, 1958.

[15] S. Gunnarsson, *Frequency Domain Aspects of Modeling and Control in Adaptive Systems*, Ph.D. thesis, Department of Electrical Engineering, Linköping University, Linköping, Sweden, 1988.

[16] S. Gunnarsson and L. Ljung, *Frequency domain tracking characteristics of adaptive algorithms*, IEEE Trans. Acoust. Speech Signal Process., 37 (1989), pp. 1072–1089.

[17] L. Guo and L. Ljung, *The role of model validation for assessing the size of the unmodeled dynamics*, IEEE Trans. Automat. Control, 42 (1997), pp. 1230–1239.

[18] P. Heuberger, P. M. J. Van den Hof, and O. Bosgra, *A generalized orthonormal basis for linear dynamical systems*, IEEE Trans. Automat. Control, AC-40 (1995), pp. 451–465.

[19] K. Hoffman, *Banach Spaces of Analytic Functions*, Prentice–Hall, Englewood Cliffs, NJ, 1962.

[20] C. Hwang and Y.-P. Shih, *Parameter identification via Laguerre polynomials*, Internat. J. Systems Sci., 13 (1982), pp. 209–217.

[21] W. H. Kautz, *Network Synthesis for Specified Transient Response*, Tech. report 209, Massachusetts Institute of Technology, Research Laboratory of Electronics, Boston, MA, 1952.

[22] R. King and P. Paraskevopoulos, *Digital Laguerre filters*, Circuit Theory Appl., 5 (1977), pp. 81–91.

[23] A. Kolmogorov, *Interpolation und extrapolation von stationären zufälligen folgen*, Bull. Acad. Sci. USSR, 5 (1941), pp. 3–14.

[24] T. Körner, *Fourier Analysis*, Cambridge University Press, Cambridge, 1988.

[25] Y. W. Lee, *Statistical Theory of Communication*, John Wiley, New York, 1960.

[26] N. Levinson, *The Wiener RMS (root mean square) error criterion in filter design and prediction*, J. Math. Phys., 25 (1947), pp. 261–278.

[27] L. Ljung, *System Identification: Theory for the User*, Prentice–Hall, Englewood Cliffs, NJ, 1987.

[28] L. Ljung, *Identification in closed loop: Some aspects on direct and indirect approaches*, in Proc. 11th IFAC Symposium on System Identification, 1997, pp. 141–146.

[29] L. Ljung, *Asymptotic variance expressions for identified black-box transfer function models*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 834–844.

[30] L. Ljung and B. Wahlberg, *Asymptotic properties of the least squares method for estimating transfer functions and disturbance spectra*, Adv. Appl. Probab., 24 (1992), pp. 412–440.

[31] L. Ljung and Z. D. Yuan, *Asymptotic properties of black-box identification of transfer functions*, IEEE Trans. Automat. Control, 30 (1985), pp. 514–530.

[32] P. Mäkila, *Approximation of stable systems by Laguerre filters*, Automatica, 26 (1990), pp. 333–345.

[33] P. Mäkilä, J. Partington, and T. Gustafsson, *Worst-case control-relevant identification*, Automatica, 31 (1995), pp. 1799–1820.

[34] F. Malmquist, *Sur la détermination d'une classe de fonctions analytiques par leurs valeurs dans un ensemble donné de points*, in C. R. Dixième Congrès Math. Scandinaves (Copenhagen), 1925, pp. 253–259.

[35] B. Ninness and F. Gustafsson, *A unifying construction of orthonormal bases for system identification*, Tech. report EE9432, Department of Electrical and Computer Engineering, the University of Newcastle, Callaghan, Australia, August 1994.

[36] B. Ninness and F. Gustafsson, *A unifying construction of orthonormal bases for system identification*, IEEE Trans. Automat. Control, 42 (1997), pp. 515–521.

[37] B. Ninness, H. Hjalmarsson, and F. Gustafsson, *The fundamental role of general orthonormal bases in system identification*, IEEE Trans. Automat. Control, to appear.

[38] A. Papoulis, *Levinson's algorithm, Wold's decomposition, and spectral estimation*, SIAM Rev., 27 (1985), pp. 405–441.

[39] J. R. Partington, *Approximation of delay systems by Fourier-Laguerre series*, Automatica, 27 (1991), pp. 569–572.

[40] P. M. J. Van den Hof, P. S. C. Heuberger, and J. Bokor, *System identification with generalized orthonormal basis functions*, Automatica, 31 (1995), pp. 1821–1834.

[41] P. Regalia, *Adaptive IIR Filtering in Signal Processing and Control*, Marcel Dekker, New York, 1995.

[42] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, Fredrick Ungar, New York, 1955.

[43] I. Schur, *Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen Veränderlichen*, J. Reine Angew. Math., 140 (1911), pp. 1–28.

[44] S. Pillai and T. Shim, *Spectrum Estimation and System Identification*, Springer–Verlag, New York, 1993.

[45] G. Szegö, *Orthogonal Polynomials*, Colloquium Publications, vol. 23, American Mathematical Society, New York, 1939.

[46] B. Wahlberg, *System identification using Laguerre models*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 551–562.

[47] B. Wahlberg, *System identification using Kautz models*, IEEE Trans. Automat. Control, AC-39 (1994), pp. 1276–1282.

[48] B. Wahlberg and E. Hannan, *Parametric signal modelling using Laguerre filters*, Ann. Appl. Prob., 3 (1993), pp. 467–496.

[49] J. Walsh, *Interpolation and Approximation by Rational Functions in the Complex Domain*, Colloquium Publications, vol. 20, American Mathematical Society, New York, 1935.

[50] H. Widom, *Studies in Real and Complex Analysis*, MAA Studies in Mathematics, Prentice–Hall, Englewood Cliffs, NJ, 1965, chap. 7, Toeplitz Matrices.

[51] N. Wiener, *The Fourier Integral and Certain of Its Applications*, Cambridge University Press, Cambridge, 1933.

[52] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series*, MIT Press, Boston, MA, 1949.

[53] Z. Yuan and L. Ljung, *Black box identification of multivariable transfer functions: Asymptotic properties and optimal input design*, Internat. J. Control, 40 (1984), pp. 233–256.

SIAM J. Control Optim.
Vol. 37, No. 2, pp. 461–485

INGHAM-TYPE THEOREMS FOR VECTOR-VALUED FUNCTIONS AND OBSERVABILITY OF COUPLED LINEAR SYSTEMS[*]

VILMOS KOMORNIK[†] AND PAOLA LORETI[‡]

Abstract. We propose a new approach to study the observability of coupled linear distributed systems. It is based on a generalization of some classical theorems of nonharmonic analysis to vector-valued functions. Applying this method we answer some questions of J.-L. Lions [*Contrôlabilité exacte et stabilisation de systèms distribués*, Vol. 2, Masson, 1988] on the boundary observability of coupled linear distributed systems in particular cases.

Key words. observability, wave equation, Petrovsky system, nonharmonic analysis, Bessel functions

AMS subject classifications. 35L05, 35Q72, 93B07, 93C20

PII. S0363012997317505

1. Introduction and formulation of the main results.

Let Ω be a bounded open domain in $\mathbb{R}^n$ having a boundary Γ of class C^4. Fix four real numbers A, B, C, D and consider the following coupled system:

$$
\begin{cases}
u_1'' - \Delta u_1 + Au_1 + Bu_2 = 0 \quad \text{in} \quad \mathbb{R}\times\Omega,\\
u_2'' + \Delta^2 u_2 + Cu_1 + Du_2 = 0 \quad \text{in} \quad \mathbb{R}\times\Omega,\\
u_1 = u_2 = \Delta u_2 = 0 \quad \text{on} \quad \mathbb{R}\times\Gamma,\\
u_i(0) = u_{i0} \quad \text{and} \quad u_i'(0) = u_{i1} \quad \text{in} \quad \Omega,\ i = 1, 2.
\end{cases}
\tag{1.1}
$$

One can readily verify by standard methods that (1.1) is well-posed in the following sense:

- Given

$$(u_{10}, u_{11}, u_{20}, u_{21}) \in H_0^1(\Omega) \times L^2(\Omega) \times H_0^1(\Omega) \times H^{-1}(\Omega)$$

arbitrarily, (1.1) has a unique *weak* solution $u = (u_1, u_2)$ satisfying

$$u_1 \in C(\mathbb{R}; H_0^1(\Omega)) \cap C^1(\mathbb{R}; L^2(\Omega))$$

and

$$u_2 \in C(\mathbb{R}; H_0^1(\Omega)) \cap C^1(\mathbb{R}; H^{-1}(\Omega)).$$

- If the initial data also satisfy the conditions

$$(u_{10}, u_{11}, u_{20}, u_{21}) \in H^2(\Omega) \times H^1(\Omega) \times H^2(\Omega) \times L^2(\Omega),$$

[*] Received by the editors February 26, 1997; accepted for publication (in revised form) March 17, 1998; published electronically December 2, 1998.
http://www.siam.org/journals/sicon/37-2/31750.html

[†] Institut de Recherche Mathématique Avancée, Université Louis Pasteur et CNRS, 7, rue René Descartes, 67084 Strasbourg cédex, France (komornik@math.u-strasbg.fr). Most of this research was done during the visit of this author to the Istituto per le Applicazioni del Calcolo "Mauro Picone" del Consiglio Nazionale delle Ricerche in Rome in September 1996.

[‡] Istituto per le Applicazioni del Calcolo "Mauro Picone," Consiglio Nazionale delle Ricerche, Viale del Policlinico, 137, 00161 Roma, Italy. Current address: Dipartimento MeMoMat, Università degli Studi di Roma "La Sapienza," Via A. Scarpa, 16, 00161 Roma, Italy (loreti@dmmm.uniroma1.it).

then the following corresponding *strong* solutions are smoother:

$$u_1 \in C(\mathbb{R}; H^2(\Omega)) \cap C^1(\mathbb{R}; H^1(\Omega))$$

and

$$u_2 \in C(\mathbb{R}; H^2(\Omega)) \cap C^1(\mathbb{R}; L^2(\Omega)).$$

Let us denote by E_0 the *initial energy* of the solution, defined by the formula

$$E_0 = \tfrac{1}{2}\big(\|u_{10}\|^2_{H^1_0(\Omega)} + \|u_{11}\|^2_{L^2(\Omega)} + \|u_{20}\|^2_{H^1_0(\Omega)} + \|u_{21}\|^2_{H^{-1}(\Omega)}\big).$$

Here, as usual, $L^2(\Omega)$ and $H^1_0(\Omega)$ are endowed with the norms defined by

$$\|v\|^2_{L^2(\Omega)} = \int_\Omega |v|^2 \, dx, \quad \|v\|^2_{H^1_0(\Omega)} = \int_\Omega |\nabla v|^2 \, dx,$$

and $H^{-1}(\Omega)$ is endowed with the dual norm of that of $H^1_0(\Omega)$.

Let us denote by ν the outward unit normal vector to Γ. We have the following proposition.

PROPOSITION 1.1. *Fix a positive number T such that Ω belongs to a ball of diameter $< T$. Then there exists a positive number α such that if $|A| < \alpha$, $|B| < \alpha$, $|C| < \alpha$, and $|D| < \alpha$, then the strong solutions of* (1.1) *satisfy the estimates*

$$c_1 E_0 \le \int_0^T \int_\Gamma |\partial_\nu u|^2 \, d\Gamma \, dt \le c_2 E_0 \tag{1.2}$$

with two positive constants c_1, c_2 depending on T and α but not on the choice of the initial data.

In the uncoupled case $A = B = C = D = 0$, Proposition 1.1 follows from earlier results of Lasiecka and Triggiani [12], Lebeau [13], Lions [14], and Zuazua [21]. (See also [8] for simplified proofs.) In the general case the theorem will be proved by applying a perturbation method of Lions [15]. A general question raised in [15] is whether the smallness of the coupling parameters is necessary for the validity of the results or whether this restriction is merely related to the technique of the proof. Our next result suggests the second alternative.

THEOREM 1.2. *Let Ω be an open ball of radius R in $\mathbb{R}^n$. There exist countably many hypersurfaces in $\mathbb{R}^4$ such that if (A,B,C,D) does not belong to any of them, then the strong solutions of* (1.1) *satisfy the estimates* (1.2) *for every $T > 2R$. The constants $c_1, c_2 > 0$ depend on A, B, C, D, and T but not on the choice of the initial data.*

Note that the exceptional parameters form a set of measure zero in $\mathbb{R}^4$.

The second inequality in (1.2) allows us to define the boundary integrals in these estimates for weak solutions by a standard density argument. Then (1.2) remains valid for weak solutions in both results.

Proposition 1.1 and Theorem 1.2 imply some boundary observability results. Indeed, if a solution of (1.1) vanishes identically in a neighborhood of Γ in the time interval $(0, T)$, then by the first inequality of (1.2) the solution of (1.1) is in fact identically zero. Taking into account the linearity of the system it follows that two solutions of (1.1) corresponding to different initial data can be distinguished by observing them only in a small neighborhood of the boundary Γ for $0 < t < T$. Indeed, it suffices to apply the preceding argument for their differences.

Our proof of Theorem 1.2 will not show whether there are effectively exceptional parameters or whether they are excluded only because of the method of the proof. In fact there are exceptional parameters indeed; see the following proposition.

PROPOSITION 1.3. *Let Ω be an open ball in $\mathbb{R}^n$. There exist countably many two-dimensional surfaces in $\mathbb{R}^4$ such that if (A,B,C,D) belongs to one of them, then for some nonzero initial data the solution of* (1.1) *satisfies the equality*

$$\partial_\nu u = 0 \quad on \quad \mathbb{R}\times\Gamma. \tag{1.3}$$

Hence the first inequality in (1.2) *cannot hold.*

It remains an interesting open problem to determine the exact dimension (between 2 and 3) of the set of exceptional quadruples.

J. E. Lagnese asked one of the authors whether the exceptional cases can be eliminated if we require the estimates (1.2) only for initial data belonging to a suitably chosen finite codimensional subspace. The answer is positive; see the following proposition.

PROPOSITION 1.4. *Let Ω be an open ball of radius R in $\mathbb{R}^n$, and fix A, B, C, D arbitrarily. There exists a positive integer N such that the estimates* (1.2) *are satisfied for every $T > 2R$ for all strong solutions of* (1.1) *whose initial data are orthogonal to the first N eigenspaces of $-\Delta$ in $H_0^1(\Omega)$. The constants $c_1, c_2 > 0$ depend on A, B, C, D, and T but not on the particular choice of the initial data.*

Now consider the same system with Neumann-type boundary conditions:

$$\begin{cases} u_1'' - \Delta u_1 + Au_1 + Bu_2 = 0 \quad \text{in} \quad \mathbb{R}\times\Omega,\\ u_2'' + \Delta^2 u_2 + Cu_1 + Du_2 = 0 \quad \text{in} \quad \mathbb{R}\times\Omega,\\ \partial_\nu u_1 = \partial_\nu u_2 = \partial_\nu \Delta u_2 = 0 \quad \text{on} \quad \mathbb{R}\times\Gamma,\\ u_i(0) = u_{i0} \quad \text{and} \quad u_i'(0) = u_{i1} \quad \text{in} \quad \Omega,\ i = 1,2. \end{cases} \tag{1.4}$$

Let us denote by Z the vector space of all quadruples whose components are finite linear combinations of the eigenfunctions of $-\Delta$ in Ω with homogeneous Neumann boundary conditions. Clearly, for every $(u_{10}, u_{20}, u_{11}, u_{21}) \in Z$ the system (1.4) has a unique solution, given by a natural trigonometric sum. Given a bounded interval I, we define a seminorm in Z by setting

$$p_I(u_{10}, u_{20}, u_{11}, u_{21}) := \|u\|_{L^2(I\times\Gamma)}.$$

It follows easily from classical unique continuation theorems that p_I is a *norm* if I is sufficiently long. It turns out that all these norms are equivalent if Ω is a ball; see the following theorem.

THEOREM 1.5. *Let Ω be an open ball of radius R in $\mathbb{R}^n$. There exist countably many hypersurfaces in $\mathbb{R}^4$ such that if (A,B,C,D) does not belong to any of them, then for any two bounded intervals I and J of length $> 2R$, p_I and p_J are equivalent norms in Z.*

Again, there exist exceptional parameters; see the following proposition.

PROPOSITION 1.6. *Let Ω be an open ball in $\mathbb{R}^n$. There exist countably many two-dimensional surfaces in $\mathbb{R}^4$ such that if (A,B,C,D) belongs to one of them, then for some nonzero initial data the solution of* (1.4) *satisfies the equality*

$$u = 0 \quad on \quad \mathbb{R}\times\Gamma.$$

Hence p_I is not a norm in Z for any bounded interval I.

On the other hand, we have a result analogous to Proposition 1.4. Given a positive integer N, denote by Z_N the subspace of those elements of Z, all of whose components are orthogonal to the first N eigenspaces of $-\Delta$ in Ω with homogeneous Neumann boundary conditions.

PROPOSITION 1.7. *Let Ω be an open ball of radius R in $\mathbb{R}^n$, and fix A, B, C, D arbitrarily. There exists a positive integer N such that for any two bounded intervals I and J of length $> 2R$, p_I and p_J are equivalent norms in Z_N.*

Applying the Hilbert uniqueness method of Lions [14], [15], Proposition 1.1 and Theorems 1.2 and 1.5 yield several exact controllability results. Similarly, applying an analogous method developed in [9], they yield uniform stabilizability results. In order to not make the paper too long we do not study these questions here.

The norm equivalence in Theorem 1.5 implies that for a corresponding boundary exact controllability problem the space of controllable states remains the same for all sufficiently large time intervals (see [15, Vol. 1, pp. 156 and 218]). While this type of equivalence is obtained automatically for Dirichlet-type boundary conditions where the multiplier method is available, for Neumann-type boundary conditions it seems that the only known method is the one applied in the present paper.

The proofs of Theorems 1.2 and 1.5 are based on the generalization of some results of nonharmonic analysis to vector-valued functions and to complex exponents. Our method is thus a generalization of the approach of Graham and Russell [3]. The following two abstract theorems may represent some interest themselves. They can be applied to the study of various other coupled distributed systems in the same way as in this paper.

First we have the following generalization of a classical theorem of Ingham [5].

THEOREM 1.8. *Let $(\omega_n)_{n=-\infty}^{\infty}$ be a sequence of real numbers satisfying for some positive integer M and for some $\gamma > 0$ the condition*

$$\omega_{n+M} - \omega_n \geq \gamma \tag{1.5}$$

for all n. Let $(u_n)_{n=-\infty}^{\infty}$ be a sequence of vectors in a complex Hilbert space H, satisfying for some $0 \leq \eta < 1/(2M-2)$ the inequalities

$$|(u_n, u_k)_H| \leq \eta \|u_n\| \, \|u_k\| \tag{1.6}$$

whenever $n \not\equiv k \mod M$. Then

(a) *for every bounded interval I there exists a constant C_1 such that*

$$\int_I \Big\| \sum_{n=-\infty}^{\infty} \alpha_n e^{i\omega_n t} u_n \Big\|^2 \, dt \leq C_1 \sum_{n=-\infty}^{\infty} |\alpha_n|^2 \|u_n\|^2 \tag{1.7}$$

for all sequences (α_n) of complex numbers satisfying

$$\sum_{n=-\infty}^{\infty} |\alpha_n|^2 \|u_n\|^2 < +\infty; \tag{1.8}$$

(b) *for every bounded interval I of length $2T$ with*

$$T > \pi/\gamma \quad \text{and} \quad \frac{T^2\gamma^2 - \pi^2}{2T^2\gamma^2 + \pi^2} > (M-1)\eta \tag{1.9}$$

there exists a constant $C_2 > 0$ such that

$$\int_I \Big\| \sum_{n=-\infty}^{\infty} \alpha_n e^{i\omega_n t} u_n \Big\|^2 \, dt \ge C_2 \sum_{n=-\infty}^{\infty} |\alpha_n|^2 \|u_n\|^2 \tag{1.10}$$

for all sequences (α_n) of complex numbers satisfying (1.8).

Remarks. 1. Since $(M-1)\eta < 1/2$, the condition (1.9) is satisfied if T is sufficiently large.

2. For $M = 1$, $H = \mathbb{C}$, and $u_n \equiv 1$, Theorem 1.8 reduces to Ingham's original theorem.

3. Ingham's theorem was later improved and generalized very much in several works; see, e.g., [2], [17], [20], [18]. These works concern the scalar case and do not directly imply our result. Concerning the vector coefficient case, the monograph of Avdonin and Ivanov [1] contains several deep theorems and also includes many useful references. In particular, Corollary II.2.2 in [1] states that a vector family has the Riesz basis property if the associated scalar family does. However, in the case of our Theorem 1.8 the associated scalar family does not necessarily have the Riesz basis property because the assumption (1.5) for $M > 1$ does not imply that the exponents are separated, so one cannot apply this result here. As it will be seen later, in the typical applications to coupled systems there can even be repeated exponents, so that the weakened assumption (1.5) of Theorem 1.8 is essential. On the other hand, using the deep complex analytic tools of [1] one could probably obtain more general results than Theorem 1.8. However, we prefer the present result because it is sufficient for most applications to coupled linear systems and because its proof is much shorter and more elementary than that of a more general case. Furthermore, this can be adapted easily to prove some variants of Theorem 1.8 which are needed for the proof of typical *partial* observability results, such as Theorem 1.10 below, and which could be proved less easily by the general complex analytic approach.

4. In the proof we shall use some ideas of another earlier generalization of Ingham's theorem due to Loreti and Valente [16].

In our application the conditions of Theorem 1.8 will be satisfied only after the removal of a finite numbers of exponents ω_n. Then we shall complete our proof by applying the following generalization of a former theorem of Haraux [4].

PROPOSITION 1.9. *Let $(\omega_n)_{n=-\infty}^{\infty}$ be a sequence of pairwise distinct complex numbers such that*

$$|\omega_n| \to +\infty \quad \textit{as} \quad n \to \pm\infty \tag{1.11}$$

and

$$\textit{the sequence} \quad (\Im \omega_n) \quad \textit{is bounded}, \tag{1.12}$$

and let (u_n) be a sequence of nonzero vectors in a complex Hilbert space H.

Let T_0 be a nonnegative number. Assume that for every bounded interval I of length $> T_0$ there exists a finite set N of integers such that the estimates (1.7) *and* (1.10) *are satisfied for all complex sequences (α_n) satisfying* (1.8) *and*

$$\alpha_n = 0 \quad \textit{for all} \quad n \in N. \tag{1.13}$$

Then the same conclusion holds without the condition (1.13).

Remarks. 1. Proposition 1.9 improves Haraux's earlier results in two aspects: we allow nonreal exponents ω_n and vector coefficients α_n.

2. Proposition 1.9 also extends an earlier result in [6] and in [8, Chap. 5], where only real exponents were considered. Although intuitively the presence of complex exponents does not present extra difficulties, in fact we had to change part of the earlier proof where the reality of the exponents was used in a crucial way.

Let us note that Theorem 1.2 remains valid in fact for every bounded domain of class C^4: this more general result can be proved by an indirect compactness-uniqueness method; see [11]. Unlike that approach, the proof given here provides explicit constants in the estimates (1.2). On the other hand, we do not know another approach to prove Theorem 1.5, and we conjecture that this theorem holds true only if Ω is a ball.

The approach developed in this paper can also be applied to the study of *partial* observability problems. For example, we have the following.

THEOREM 1.10. *Let Ω be an open ball of radius R in $\mathbb{R}^n$. There exist countably many hypersurfaces in $\mathbb{R}^4$ such that if (A,B,C,D) does not belong to any of them, then the following estimates hold true:*

(a) *For every $T > 2R$ there exist two positive constants c_1, c_2 such that*

$$c_1E_0 \le \int_0^T \int_\Gamma |\partial_\nu u_1|^2 \ d\Gamma \ dt \le c_2E_0$$

for all strong solutions of (1.1) *whose initial data satisfy the condition*

$$u_{20} = u_{21} = 0 \quad in \quad \Omega.$$

(b) *For every $T > 0$ there exist two positive constants c_1, c_2 such that*

$$c_1E_0 \le \int_0^T \int_\Gamma |\partial_\nu u_2|^2 \ d\Gamma \ dt \le c_2E_0$$

for all strong solutions of (1.1) *whose initial data satisfy the condition*

$$u_{10} = u_{11} = 0 \quad in \quad \Omega.$$

Due to space limitations we shall not prove Theorem 1.10 here and we refer to [10] for similar results. Its proof is analogous to that of Theorems 1.2 and 1.5, but it requires modified versions of Theorem 1.8 and Proposition 1.9. Contrary to Theorems 1.2 and 1.5, we do not know whether there are effectively exceptional values in this problem.

2. Proof of Proposition 1.1. Let us consider, along with (1.1), the following two related systems:

$$\begin{cases} v_1'' - \Delta v_1 = 0 \quad \text{in} \quad \mathbb{R}\times\Omega, \\ v_2'' + \Delta^2 v_2 = 0 \quad \text{in} \quad \mathbb{R}\times\Omega, \\ v_1 = v_2 = \Delta v_2 = 0 \quad \text{on} \quad \mathbb{R}\times\Gamma, \\ v_i(0) = u_{i0} \quad \text{and} \quad v_i'(0) = u_{i1} \quad \text{in} \quad \Omega, \ i = 1,2, \end{cases} \tag{2.1}$$

and

$$\begin{cases} w_1'' - \Delta w_1 + Au_1 + Bu_2 = 0 \quad \text{in} \quad \mathbb{R}\times\Omega, \\ w_2'' + \Delta^2 w_2 + Cu_1 + Du_2 = 0 \quad \text{in} \quad \mathbb{R}\times\Omega, \\ w_1 = w_2 = \Delta w_2 = 0 \quad \text{on} \quad \mathbb{R}\times\Gamma, \\ w_i(0) = w_i'(0) = 0 \quad \text{in} \quad \Omega, \ i = 1,2. \end{cases} \tag{2.2}$$

Observe that $u = v + w$.

Since the system (2.1) is uncoupled, we may apply the earlier results of Lions [14] and Zuazua [21] mentioned in the introduction. Hence

$$c_3 E_0 \le \int_0^T \int_\Gamma |\partial_\nu v|^2 \, d\Gamma \, dt \le c_4 E_0 \tag{2.3}$$

for some positive constants c_3, c_4, independent of A, B, C, D (because these parameters do not appear in (2.1)). (All constants in this section are independent of the particular choice of the initial data.)

Fix $\alpha > 0$ (to be chosen later) and let $|A| < \alpha$, $|B| < \alpha$, $|C| < \alpha$, $|D| < \alpha$. It follows from the well-posedness of the system (1.1) that

$$\begin{aligned} \|Au_1\|^2_{L^1(0,T;H^1_0(\Omega))} + \|Bu_2\|^2_{L^1(0,T;H^1_0(\Omega))} + \|Cu_1\|^2_{L^1(0,T;H^1_0(\Omega))} \\ + \|Du_2\|^2_{L^1(0,T;H^1_0(\Omega))} \le c_5(\alpha) E_0 \end{aligned}$$

for some constants $c_5(\alpha)$ such that $c_5(\alpha) \to 0$ as $\alpha \to 0$. Therefore, applying the direct inequalities given in [14] for the wave equation and for the above Petrovsky system, we obtain that

$$\int_0^T \int_\Gamma |\partial_\nu w|^2 \, d\Gamma \, dt \le c_6(\alpha) E_0 \tag{2.4}$$

for some constant $c_6(\alpha)$ such that $c_6(\alpha) \to 0$ as $\alpha \to 0$.

Since $u = v + w$, we deduce from (2.3) and (2.4) the following inequalities:

$$\int_0^T \int_\Gamma |\partial_\nu u|^2 \, d\Gamma \, dt \le 2 \int_0^T \int_\Gamma |\partial_\nu v|^2 + |\partial_\nu w|^2 \, d\Gamma \, dt \le 2(c_4 + c_6(\alpha)) E_0,$$

$$2 \int_0^T \int_\Gamma |\partial_\nu u|^2 \, d\Gamma \, dt \ge \int_0^T \int_\Gamma |\partial_\nu v|^2 - 2|\partial_\nu w|^2 \, d\Gamma \, dt \ge (c_3 - 2c_6(\alpha)) E_0.$$

These inequalities imply (1.2) with

$$c_1 = (c_3 - 2c_6(\alpha))/2 \quad \text{and} \quad c_2 = 2(c_4 + c_6(\alpha)).$$

If we choose a sufficiently small α, then the constants c_1, c_2 are positive and the proposition follows. □

3. Proof of Theorem 1.8. If (1.7) and (1.10) are satisfied for an interval $I = (-T, T)$, then they are also satisfied for every translate $(-T+\tau, T+\tau)$ of this interval. Indeed, applying (1.7) and (1.10) to the function

$$z(t) := u(t+\tau) = \sum_n \big(a_n e^{i\omega_n \tau}\big) e^{i\omega_n t} u_n$$

and using the equalities $|e^{i\omega_n \tau}| = 1$ (because the ω_n's are real), we obtain that (1.7) and (1.10) are satisfied if we replace $\|u(t)\|^2$ by $\|z(t)\|^2$ in the integral. We conclude by observing that

$$\int_{-T+\tau}^{T+\tau} \|u(t)\|^2 \, dt = \int_{-T}^{T} \|z(t)\|^2 \, dt.$$

Hence it is sufficient to consider intervals of the form $I = (-T, T)$.

Let us introduce for any fixed $T > \pi/\gamma$ the continuous nonnegative function

$$k(t) = \begin{cases} \cos \frac{\pi t}{2T} & \text{if } |t| \le T, \\ 0 & \text{if } |t| \ge T. \end{cases}$$

We shall prove instead of (1.7) and (1.10) the estimates

$$\int_{-\infty}^{\infty} k(t) \Big\| \sum_{n=-\infty}^{\infty} \alpha_n e^{i\omega_n t} u_n \Big\|^2 \, dt \le C_1 \sum_{n=-\infty}^{\infty} |\alpha_n|^2 \|u_n\|^2 \tag{3.1}$$

and

$$\int_{-\infty}^{\infty} k(t) \Big\| \sum_{n=-\infty}^{\infty} \alpha_n e^{i\omega_n t} u_n \Big\|^2 \, dt \ge C_2 \sum_{n=-\infty}^{\infty} |\alpha_n|^2 \|u_n\|^2. \tag{3.2}$$

Since $k \le \chi_{[-T,T]}$, (1.10) follows at once from (3.2). For the proof of (1.7) we apply (3.1) with some $T' > T$ instead of T. Then the corresponding function k satisfies

$$k \ge \cos \frac{\pi T}{2T'} \; \chi_{[-T,T]}$$

and hence (1.7) follows with $C_1 / \cos(\pi T/(2T')) > 0$ instead of C_1.

Turning to the proofs of (3.1) and (3.2), let us denote by $K(x)$ the Fourier transform of $k(t)$

$$K(x) = \int_{-\infty}^{\infty} k(t) e^{ixt} \, dt.$$

An easy computation shows that

$$K(x) = \frac{4\pi T \cos(xT)}{\pi^2 - 4T^2 x^2}.$$

Hence $K(-x) \equiv K(x)$ and the maximum of $|K(x)|$ over $\mathbb{R}$ is attained by $K(0) = 4T/\pi$. Indeed, one can verify directly by looking at the graph of $K(x)$ that $|K(x)| \le K(0)$ for all $-1/T \le x \le 1/T$. Furthermore, for $x \ge 1/T$ we have

$$\begin{aligned} |K(x)| &= \Big| \frac{4\pi T \cos(xT)}{\pi^2 - 4T^2 x^2} \Big| = \Big| \frac{\pi T \sin(xT - (\pi/2))}{(xT - (\pi/2))(xT + (\pi/2))} \Big| \\ &= \Big| \frac{2\pi T}{2xT + \pi} \frac{\sin(xT - (\pi/2))}{xT - (\pi/2)} \Big| \le \frac{2\pi T}{2xT + \pi} \le \frac{4T}{\pi}. \end{aligned}$$

Finally, for $x \le -1/T$ we have $|K(x)| = |K(-x)| \le 4T/\pi$.

Developing the scalar product inside the integral in (3.1) and (3.2) we obtain

$$\int_{-\infty}^{\infty} k(t) \Big\| \sum_n \alpha_n e^{i\omega_n t} u_n \Big\|^2 \, dt = \sum_{n,k} \alpha_n \overline{\alpha_k} (u_n, u_k)_H K(\omega_n - \omega_k),$$

from which

$$\int_{-\infty}^{\infty} k(t) \Big\| \sum_n \alpha_n e^{i\omega_n t} u_n \Big\|^2 \, dt - K(0) \sum_n |\alpha_n|^2 \|u_n\|^2 = S_1 + S_2 \tag{3.3}$$

with

$$S_1 = \sum_{n\equiv k, n\neq k} \alpha_n \overline{\alpha_k} (u_n, u_k)_H K(\omega_n - \omega_k)$$

and

$$S_2 = \sum_{n\not\equiv k} \alpha_n \overline{\alpha_k} (u_n, u_k)_H K(\omega_n - \omega_k),$$

where $n \equiv k$ means $n \equiv k \bmod M$.

Using (1.5) and the inequality $T > \pi/\gamma$ we have for every fixed n the following estimate:

$$\begin{aligned}\sum_{k\equiv n, k\neq n} |K(\omega_n - \omega_k)| &\le \sum_{n\equiv k, n\neq k} \frac{4\pi T}{4T^2|\omega_n - \omega_k|^2 - \pi^2} \\ \le \sum_{n\equiv k, n\neq k} \frac{4\pi T}{4T^2(n-k)^2 M^{-2}\gamma^2 - T^2\gamma^2} &= \sum_{j=1}^{\infty} \frac{8\pi T}{T^2\gamma^2(4j^2-1)} \\ &= \frac{4\pi}{T\gamma^2} \sum_{j=1}^{\infty} \Big(\frac{1}{2j-1} - \frac{1}{2j+1} \Big) = \frac{4\pi}{T\gamma^2}.\end{aligned} \tag{3.4}$$

Therefore

$$\begin{aligned}|S_1| &\le \sum_{n\equiv k, n\neq k} |\alpha_n|\ |\alpha_k|\ \|u_n\|\ \|u_k\|\ |K(\omega_n - \omega_k)| \\ &\le \frac{1}{2} \sum_{n\equiv k, n\neq k} (|\alpha_n|^2 \|u_n\|^2 + |\alpha_k|^2 \|u_k\|^2) |K(\omega_n - \omega_k)| \\ &= \sum_{n\equiv k, n\neq k} |\alpha_n|^2 \|u_n\|^2 |K(\omega_n - \omega_k)| \\ &\le \frac{4\pi}{T\gamma^2} \sum_n |\alpha_n|^2 \|u_n\|^2.\end{aligned} \tag{3.5}$$

Next we remark that using (3.4) we have for any fixed n the estimate

$$\sum_{k\not\equiv n} |K(\omega_n - \omega_k)| \le \sum_{i=1}^{M-1} \sum_{k\equiv n+i} |K(\omega_n - \omega_k)| \le (M-1)\Big(\frac{4\pi}{T\gamma^2} + 2K(0) \Big),$$

because by (1.5) for each fixed $1 \le i \le M-1$ there are at most two integers $k \equiv n+i$ with $|\omega_n - \omega_k| < \gamma$. Therefore

$$\begin{aligned}|S_2| &\le \eta \sum_{k\not\equiv n} |\alpha_n|\ |\alpha_k|\ \|u_n\|\ \|u_k\|\ |K(\omega_n - \omega_k)| \\ &\le \frac{\eta}{2} \sum_{k\not\equiv n} (|\alpha_n|^2 \|u_n\|^2 + |\alpha_k|^2 \|u_k\|^2) |K(\omega_n - \omega_k)| \\ &= \eta \sum_{k\not\equiv n} |\alpha_n|^2 \|u_n\|^2 |K(\omega_n - \omega_k)| \\ &\le \eta(M-1)\Big(\frac{4\pi}{T\gamma^2} + 2K(0) \Big) \sum_n |\alpha_n|^2 \|u_n\|^2.\end{aligned} \tag{3.6}$$

Using (3.5), (3.6), and the equality $K(0) = 4T/\pi$ we deduce from (3.3) the inequality

$$\begin{aligned}&\Big|\int_{-\infty}^{\infty} k(t)\Big\|\sum_n \alpha_n e^{i\omega_n t}u_n\Big\|^2 \, dt - \frac{4T}{\pi}\sum_n |\alpha_n|^2\|u_n\|^2\Big|\\ &\le \Big[\frac{4\pi}{T\gamma^2} + \eta(M-1)\Big(\frac{8T}{\pi} + \frac{4\pi}{T\gamma^2}\Big)\Big]\sum_n |\alpha_n|^2\|u_n\|^2,\end{aligned}$$

and (1.7) and (1.10) follow by the choice (1.9) of T. □

4. Proof of Proposition 1.9. Given $\delta > 0$ and $\omega \in \mathbb{C}$ arbitrarily, for every continuous function $u : \mathbb{R} \to H$ we define another function $I_{\delta,\omega}u : \mathbb{R} \to H$ by the formula

$$I_{\delta,\omega}u(t) = u(t) - \frac{1}{2\delta}\int_{-\delta}^{\delta} e^{-i\omega s}u(t+s)\, ds.$$

We shall need two lemmas.

LEMMA 4.1. (a) *If $u(t) = e^{i\omega t}u_0$ with $u_0 \in H$, then $I_{\delta,\omega}u = 0$.*

(b) *If $u(t) = e^{i\mu t}u_0$ with $\mu \neq \omega$ and $u_0 \in H$, then*

$$I_{\delta,\omega}u(t) = \left(1 - \frac{\sin(\mu-\omega)\delta}{(\mu-\omega)\delta}\right)u(t).$$

(c) *The linear operators $I_{\delta,\omega}$ commute, i.e.,*

$$I_{\delta,\omega}I_{\delta',\omega'}u = I_{\delta',\omega'}I_{\delta,\omega}u$$

for all $\delta, \omega, \delta', \omega'$, and u.

Observe that since the analytic function $1 - (\sin z/z)$ does not vanish identically, the function in front of $u(t)$ in part (b) cannot vanish for more than a countable set of exceptional values of $\delta > 0$.

Proof. (a) We have

$$I_{\delta,\omega}u(t) = u(t) - \frac{1}{2\delta}\int_{-\delta}^{\delta} e^{-i\omega s}e^{i\omega(t+s)}u_0\, ds = u(t) - e^{i\omega t}u_0 = 0.$$

(b) We have

$$\begin{aligned}I_{\delta,\omega}u(t) &= u(t) - \frac{1}{2\delta}\int_{-\delta}^{\delta} e^{-i\omega s}e^{i\mu(t+s)}u_0\, ds = u(t) - \frac{1}{2\delta}\left[\frac{e^{i(\mu-\omega)s}}{i(\mu-\omega)}\right]_{-\delta}^{\delta} e^{i\mu t}u_0\\ &= u(t) - \frac{e^{i(\mu-\omega)\delta} - e^{-i(\mu-\omega)\delta}}{2i(\mu-\omega)\delta}e^{i\mu t}u_0 = \left(1 - \frac{\sin(\mu-\omega)\delta}{(\mu-\omega)\delta}\right)u(t).\end{aligned}$$

(c) This follows at once from the definition of the operators $I_{\delta,\omega}$. □

LEMMA 4.2. *For every continuous function $u : \mathbb{R} \to H$ we have*

$$\int_{-T}^{T} \|I_{\delta,\omega}u(t)\|^2 \, dt \le \left(2 + 2e^{2|\Im\omega|\delta}\right)\int_{-T-\delta}^{T+\delta} \|u(t)\|^2 \, dt. \tag{4.1}$$

Proof. For every fixed $t \in \mathbb{R}$ we have

$$\begin{aligned}\|I_{\delta,\omega}u(t)\|^2 &\le 2\|u(t)\|^2 + 2\Big\|\frac{1}{2\delta}\int_{-\delta}^{\delta} e^{-i\omega s}u(t+s)\ ds\Big\|^2\\ &\le 2\|u(t)\|^2 + \frac{1}{2\delta^2}\int_{-\delta}^{\delta}|e^{-i\omega s}|^2\ ds\int_{-\delta}^{\delta}\|u(t+s)\|^2\ ds\\ &\le 2\|u(t)\|^2 + \delta^{-1}e^{2|\Im\omega|\delta}\int_{t-\delta}^{t+\delta}\|u(x)\|^2\ dx.\end{aligned}$$

Therefore

$$\begin{aligned}&\int_{-T}^{T}\|I_{\delta,\omega}u(t)\|^2\ dt\\ &\le 2\int_{-T}^{T}\|u(t)\|^2\ dt + \delta^{-1}e^{2|\Im\omega|\delta}\int_{-T}^{T}\int_{t-\delta}^{t+\delta}\|u(x)\|^2\ dx\ dt\\ &= 2\int_{-T}^{T}\|u(t)\|^2\ dt + \delta^{-1}e^{2|\Im\omega|\delta}\int_{-T-\delta}^{T+\delta}\int_{\max\{-T,x-\delta\}}^{\min\{T,x+\delta\}}\|u(x)\|^2\ dt\ dx\\ &\le 2\int_{-T}^{T}\|u(t)\|^2\ dt + 2e^{2|\Im\omega|\delta}\int_{-T-\delta}^{T+\delta}\|u(t)\|^2\ dt\\ &\le \big(2+2e^{|\Im\omega|\delta}\big)\int_{T-\delta}^{T+\delta}\|u(x)\|^2\ dx. \qquad \square\end{aligned}$$

Now we turn to the proof of the theorem. Fix $\varepsilon > 0$ arbitrarily. We proceed in four steps.

Step 1. Fix $\varepsilon/(2|N|) < \delta < \varepsilon/|N|$ (to be chosen later), where $|N|$ denotes the number of elements in the set N, and let us denote by $I = I_\delta$ the composition of the linear operators I_{δ,ω_j}, where j runs over N. By Lemma 4.1 the definition does not depend on the order of the operators I_{δ,ω_j}. Therefore, using Lemma 4.1, if

$$u(t) = \sum_{n=1}^{\infty}\alpha_n e^{i\omega_n t}u_n,$$

then

$$(Iu)(t) = \sum_{n\notin N}\alpha_n\left(\prod_{j\in N}\left(1-\frac{\sin(\omega_n-\omega_j)\delta}{(\omega_n-\omega_j)\delta}\right)\right)e^{i\omega_n t}u_n =: \sum_{n=-\infty}^{\infty}\alpha_n' e^{i\omega_n t}u_n.$$

Next we choose $\varepsilon/(2|N|) < \delta < \varepsilon/|N|$ such that none of the products

$$\prod_{j\in N}\left(1-\frac{\sin(\omega_n-\omega_j)\delta}{(\omega_n-\omega_j)\delta}\right), \qquad n \notin N,$$

vanishes. This is possible by the analyticity remark preceding the proof of Lemma 4.1 because the numbers $\omega_n - \omega_j$ are all different from zero. (We have to exclude only a countable set of values of δ.)

Then there exists a constant $C' > 0$ such that

$$\Big|\prod_{j\in N}\left(1-\frac{\sin(\omega_n-\omega_j)\delta}{|(\omega_n-\omega_j)\delta|}\right)\Big|^2 \ge C'$$

for all $n \notin N$. Indeed, it is sufficient to observe that for any fixed $j \in N$ we have

$$\Big|\frac{\sin(\omega_n - \omega_j)\delta}{(\omega_n - \omega_j)\delta}\Big| \le \frac{\cosh(\Im(\omega_n - \omega_j)\delta)}{|(\omega_n - \omega_j)\delta|} \to 0$$

as $n \to \pm\infty$ because of our assumptions (1.11) and (1.12). Hence the above product tends to 1 as $n \to \pm\infty$, so that its absolute value is minorized, e.g., by $1/2$ for all sufficiently large $|n|$.

The above argument also implies that the above products are bounded with respect to n so that (α_n') also satisfies (1.8). Furthermore, $\alpha_n' = 0$ for all $n \in N$ so that we may apply our hypothesis to the function Iu on any interval $(-T, T)$ with $T > T_0/2$. It follows that

$$\int_{-T}^{T} \|Iu(t)\|^2 \, dt \ge C_2 \sum_{n \notin N} |\alpha_n'|^2 \|u_n\|^2 \ge C_2 C' \sum_{n \notin N} |\alpha_n|^2 \|u_n\|^2.$$

Applying (4.1) repeatedly with $\omega = \omega_j$, $j \in N$, and taking into account that $|N|\delta < \varepsilon$, it follows that

$$\sum_{n \notin N} |\alpha_n|^2 \|u_n\|^2 \le C'' \int_{-T-\varepsilon}^{T+\varepsilon} \|u(t)\|^2 \, dt \tag{4.2}$$

with

$$C'' = \frac{1}{C_2 C'} \prod_{j \in N} \big(2 + 2e^{2|\Im\omega_j|\varepsilon/|N|}\big).$$

Step 2. We are going to prove (1.7) for the interval $I = (-T - \varepsilon, T + \varepsilon)$. Let us first show that

$$\int_{-T-\varepsilon}^{T+\varepsilon} \Big\| \sum_{n \notin N} \alpha_n e^{i\omega_n t} u_n \Big\|^2 \, dt \le C_1(\varepsilon) \sum_{n \notin N} |\alpha_n|^2 \|u_n\|^2 \tag{4.3}$$

with a suitable constant $C_1(\varepsilon)$ defined later. Indeed, let us cover $(-T - \varepsilon, T + \varepsilon)$ by a finite number of translates

$$(-T + \tau_j, T + \tau_j), \qquad 1 \le j \le m$$

of the interval $(-T, T)$ and apply the hypothesis (1.7) m times as follows:

$$\begin{aligned} \int_{-T-\varepsilon}^{T+\varepsilon} \Big\| \sum_{n \notin N} \alpha_n e^{i\omega_n t} u_n \Big\|^2 \, dt &\le \sum_{j=1}^{m} \int_{-T+\tau_j}^{T+\tau_j} \Big\| \sum_{n \notin N} \alpha_n e^{i\omega_n t} u_n \Big\|^2 \, dt \\ &= \sum_{j=1}^{m} \int_{-T}^{T} \Big\| \sum_{n \notin N} \big(\alpha_n e^{i\omega_n \tau_j}\big) e^{i\omega_n t} u_n \Big\|^2 \, dt \end{aligned}$$

$$\le C_1 \sum_{j=1}^{m} \sum_{n \notin N} |\alpha_n e^{i\omega_n \tau_j}|^2 \cdot \|u_n\|^2 = C_1 \sum_{n \notin N} |\alpha_n|^2 \left(\sum_{j=1}^{m} e^{-2\Im\omega_n \tau_j} \right) \cdot \|u_n\|^2.$$

Thanks to hypothesis (1.12), the set of numbers

$$\sum_{j=1}^{m} e^{-2\Im\omega_n \tau_j}, \quad n \notin N$$

is bounded and therefore (4.3) follows with

$$C_1(\varepsilon) = C_1 \sup_{n \notin N} \sum_{j=1}^{m} e^{-2\Im \omega_n \tau_j}.$$

Next we show that

$$\int_{-T-\varepsilon}^{T+\varepsilon} \Big\| \sum_{n\in N} \alpha_n e^{i\omega_n t} u_n \Big\|^2 \, dt \le C_1'(\varepsilon) \sum_{n\in N} |\alpha_n|^2 \|u_n\|^2 \tag{4.4}$$

with some constant $C_1'(\varepsilon)$. Indeed, applying the Cauchy–Schwarz inequality we have

$$\Big\| \sum_{n\in N} \alpha_n e^{i\omega_n t} u_n \Big\|^2 \le \left(\sum_{n\in N} |\alpha_n| |e^{i\omega_n t}| \|u_n\| \right)^2 \le |N| \sum_{n\in N} |\alpha_n|^2 |e^{i\omega_n t}|^2 \|u_n\|^2,$$

and (4.4) follows with

$$C_1'(\varepsilon) = |N| \max \left\{ \int_{-T-\varepsilon}^{T+\varepsilon} |e^{i\omega_n t}|^2 \, dt \;\Big|\; n \in N \right\}.$$

Using the triangle inequality we deduce from (4.3) and (4.4) that

$$\begin{aligned} &\int_{-T-\varepsilon}^{T+\varepsilon} \Big\| \sum_{n=-\infty}^{\infty} \alpha_n e^{i\omega_n t} u_n \Big\|^2 \, dt \\ &\le \int_{-T-\varepsilon}^{T+\varepsilon} 2\Big\| \sum_{n\notin N} \alpha_n e^{i\omega_n t} u_n \Big\|^2 + 2\Big\| \sum_{n\in N} \alpha_n e^{i\omega_n t} u_n \Big\|^2 \, dt \\ &\le C_1(\varepsilon) \sum_{n\notin N} |\alpha_n|^2 \|u_n\|^2 + C_1'(\varepsilon) \sum_{n\in N} |\alpha_n|^2 \|u_n\|^2, \end{aligned}$$

and (1.7) follows with

$$\max\{C_1(\varepsilon), C_1'(\varepsilon)\}$$

in place of C_1.

Step 3. Now we prove (1.10). First, using the triangle inequality, our hypothesis (1.7), and the inequality (4.2) we obtain that

$$\begin{aligned} \int_{-T}^{T} \Big\| \sum_{n\in N} \alpha_n e^{i\omega_n t} u_n \Big\|^2 \, dt &\le \int_{-T}^{T} 2\|u(t)\|^2 + 2\Big\| \sum_{n\notin N} \alpha_n e^{i\omega_n t} u_n \Big\|^2 \, dt \\ &\le 2\int_{-T}^{T} \|u(t)\|^2 \, dt + 2C_1 \sum_{n\notin N} |\alpha_n|^2 \|u_n\|^2 \\ &\le 2\int_{-T}^{T} \|u(t)\|^2 \, dt + 2C_1 C'' \int_{-T-\varepsilon}^{T+\varepsilon} \|u(t)\|^2 \, dt \\ &\le (2 + 2C_1 C'') \int_{-T-\varepsilon}^{T+\varepsilon} \|u(t)\|^2 \, dt. \end{aligned}$$

Let us observe that the expression

$$\int_{-T}^{T} \Big\| \sum_{n\in N} \alpha_n e^{i\omega_n t} u_n \Big\|^2 \ dt$$

is a positive semidefinite quadratic form of the variable $\{\alpha_n \mid n \in N\} \in \mathbb{C}^{|N|}$. Moreover, it is positive *definite* because the functions $e^{i\omega_n t}u_n$, $n \in N$, are linearly independent. (Let us recall that N is a finite set, so that there is only one natural notion of linear independence and of positive definiteness here; cf. [1, pp. 23–26].) Hence it is minorized by a positive multiple of

$$\sum_{n\in N} |\alpha_n|^2 \|u_n\|^2,$$

and we deduce from the above inequality that

$$\sum_{n\in N} |\alpha_n|^2 \|u_n\|^2 \le C''' \int_{-T-\varepsilon}^{T+\varepsilon} \|u(t)\|^2 \ dt \tag{4.5}$$

for a suitable constant $C''' > 0$. Now (1.10) follows from (4.2) and (4.5) with $1/(C'' + C''')$ in place of C_2.

Step 4. Since $T > T_0/2$ and $\varepsilon > 0$ were chosen arbitrarily, we have proved (1.7) and (1.10) for all bounded intervals I of length $> T_0$ which are symmetrical with respect to 0. Finally, we prove that they remain valid for every interval I of length $2T > T_0$ even if it is not symmetrical with respect to 0.

Let us write $I = (-T + \tau, T + \tau)$ and consider the function $z(t) := u(t + \tau)$. We have clearly

$$\int_{-T+\tau}^{T+\tau} \|u(t)\|^2 \ dt = \int_{-T}^{T} \|z(t)\|^2 \ dt$$

and

$$z(t) = \sum_n \big(\alpha_n e^{i\omega_n \tau}\big) e^{i\omega_n t} u_n.$$

By hypothesis (1.12) there is a constant M such that $|\Im \omega_n| \le M$ for all n. Applying (1.7) and (1.10) to $z(t)$ on the interval $(-T, T)$ and using the inequalities

$$e^{-M\tau} \le |e^{i\omega_n \tau}| \le e^{M\tau}$$

we obtain that $u(t)$ satisfies (1.7) and (1.10) on the interval $I = (-T + \tau, T + \tau)$ with C_1, C_2 replaced by $C_1 e^{2M\tau}$ and $C_2 e^{-2M\tau}$, respectively.

5. Proof of Theorem 1.2. We may assume without loss of generality that Ω is the unit ball of $\mathbb{R}^n$: the general case then follows easily by a linear change of variables.

We shall only consider the case $n \ge 2$. The proof of the one-dimensional case is similar and simpler; we shall indicate briefly the modifications at the end of this section.

Let us denote by

$$\rho_{m1} < \rho_{m2} < \cdots$$

the sequence of the (strictly) positive roots of the Bessel function $J_{m-1+\frac{n}{2}}(x)$ for $m = 0, 1, \ldots$, and let us introduce the eigenvalues

$$\lambda_{mk} = \tfrac{1}{2}\Big(\rho_{mk}^4 + \rho_{mk}^2 + A + D + \sqrt{(\rho_{mk}^4 - \rho_{mk}^2 + D - A)^2 + 4BC}\Big) \tag{5.1}$$

and

$$\mu_{mk} = \tfrac{1}{2}\Big(\rho_{mk}^4 + \rho_{mk}^2 + A + D - \sqrt{(\rho_{mk}^4 - \rho_{mk}^2 + D - A)^2 + 4BC}\Big) \tag{5.2}$$

of the matrices

$$A_{mk} = \begin{pmatrix} \rho_{mk}^2 + A & B \\ C & \rho_{mk}^4 + D \end{pmatrix}$$

for $m = 0, 1, \ldots$ and $k = 1, 2, \ldots$. We recall from [19] that

$$\rho_{mk} \to \infty \quad \text{as} \quad m + k \to \infty. \tag{5.3}$$

These relations imply that

$$\rho_{mk}^4(\lambda_{mk} - \rho_{mk}^4 - D) = \rho_{mk}^4(\rho_{mk}^2 + A - \mu_{mk}) \to BC \tag{5.4}$$

as $m + k \to \infty$.

We shall assume that

$$B \neq 0, \quad C \neq 0 \tag{5.5}$$

and that the numbers

$$0, \lambda_{m1}, \mu_{m1}, \lambda_{m2}, \mu_{m2}, \ldots \tag{5.6}$$

are all different for each m.

Observe that these assumptions exclude only a set of measure zero of the quadruples (A, B, C, D) in $\mathbb{R}^4$. Indeed, more precisely, each of the countably many equations

$$\begin{aligned} B &= 0, \\ C &= 0, \\ \lambda_{mk} &= 0, \ k = 1, 2, \ldots, \\ \mu_{mk} &= 0, \ k = 1, 2, \ldots, \\ \lambda_{mk} &= \mu_{ml}, \ k, l = 1, 2, \ldots, \\ \lambda_{mk} &= \lambda_{ml}, \ k, l = 1, 2, \ldots, \ k \neq l, \\ \mu_{mk} &= \mu_{ml}, \ k, l = 1, 2, \ldots, \ k \neq l \end{aligned}$$

defines a hypersurface in $\mathbb{R}^4$, so that the union of these hypersurfaces has zero measure, and the conditions (5.5) and (5.6) are satisfied for all quadruples (A, B, C, D) outside this set.

Using (5.1)–(5.6) one can show easily that the vectors v_{mk} and w_{mk} defined by the formulas

$$\begin{cases} v_{mk} = (v_{mk1}, v_{mk2}) := (C^{-1}(\lambda_{mk} - \rho_{mk}^4 - D), 1), \\ w_{mk} = (w_{mk1}, w_{mk2}) := (1, B^{-1}(\mu_{mk} - \rho_{mk}^2 - A)) \end{cases} \tag{5.7}$$

satisfy the following conditions:

$$A_{mk}v_{mk} = \lambda_{mk}v_{mk} \quad \text{and} \quad A_{mk}w_{mk} = \mu_{mk}w_{mk}, \tag{5.8}$$

$$v_{mk} \text{ and } w_{mk} \quad \text{are linearly independent,} \tag{5.9}$$

$$v_{mk2} = w_{mk1} = 1, \tag{5.10}$$

$$v_{mk1}, w_{mk2} = O(\rho_{mk}^{-4}) \quad \text{as} \quad m+k \to \infty. \tag{5.11}$$

Applying the Fourier method we obtain that the solution of (1.1) is given by the series (we use hyperspherical coordinates)

$$\begin{aligned} u(t,r,\varphi) = \sum_{m=0}^{\infty}\sum_{k=1}^{\infty} & \quad r^{1-\frac{n}{2}} J_{m-1+\frac{n}{2}}(\rho_{mk}r) \\ & \times\big\{\big(a_{mk}(\varphi)e^{i\sqrt{\lambda_{mk}}t} + b_{mk}(\varphi)e^{-i\sqrt{\lambda_{mk}}t}\big)v_{mk} \\ & +\big(c_{mk}(\varphi)e^{i\sqrt{\mu_{mk}}t} + d_{mk}(\varphi)e^{-i\sqrt{\mu_{mk}}t}\big)w_{mk}\big\} \end{aligned} \tag{5.12}$$

with suitable spherical harmonics a_{mk}, b_{mk}, c_{mk}, d_{mk} of order m, depending on the initial data, and such that

$$\sum_{m=0}^{\infty}\sum_{k=1}^{\infty} |\rho_{mk}J'_{m-1+\frac{n}{2}}(\rho_{mk})|^2 \int_\Gamma |a_{mk}|^2 + |b_{mk}|^2 + |c_{mk}|^2 + |d_{mk}|^2 \ d\Gamma < \infty.$$

Expanding the initial data according to the eigenfunctions of $-\Delta$ with the homogeneous Dirichlet boundary condition, we obtain the orthogonal expansions

$$\begin{cases} u_{10}(r,\varphi) = \sum_{m=0}^{\infty}\sum_{k=1}^{\infty} r^{1-\frac{n}{2}} J_{m-1+\frac{n}{2}}(\rho_{mk}r)u_{10mk}(\varphi), \\ u_{20}(r,\varphi) = \sum_{m=0}^{\infty}\sum_{k=1}^{\infty} r^{1-\frac{n}{2}} J_{m-1+\frac{n}{2}}(\rho_{mk}r)u_{20mk}(\varphi), \\ u_{11}(r,\varphi) = \sum_{m=0}^{\infty}\sum_{k=1}^{\infty} r^{1-\frac{n}{2}} J_{m-1+\frac{n}{2}}(\rho_{mk}r)u_{11mk}(\varphi), \\ u_{21}(r,\varphi) = \sum_{m=0}^{\infty}\sum_{k=1}^{\infty} r^{1-\frac{n}{2}} J_{m-1+\frac{n}{2}}(\rho_{mk}r)u_{21mk}(\varphi), \end{cases} \tag{5.13}$$

where u_{10mk}, u_{20mk}, u_{11mk}, u_{21mk} are suitable spherical harmonics of order m. Comparing (5.12) and (5.13) we obtain the algebraic relations

$$\begin{cases} u_{10mk} = (a_{mk}+b_{mk})v_{mk1} + (c_{mk}+d_{mk})w_{mk1}, \\ u_{20mk} = (a_{mk}+b_{mk})v_{mk2} + (c_{mk}+d_{mk})w_{mk2}, \\ u_{11mk} = i\sqrt{\lambda_{mk}}(a_{mk}-b_{mk})v_{mk1} + i\sqrt{\mu_{mk}}(c_{mk}-d_{mk})w_{mk1}, \\ u_{21mk} = i\sqrt{\lambda_{mk}}(a_{mk}-b_{mk})v_{mk2} + i\sqrt{\mu_{mk}}(c_{mk}-d_{mk})w_{mk2}. \end{cases} \tag{5.14}$$

In order to simplify the notations we shall write $f \approx g$ if there exist two positive constants c' and c'' such that

$$c'f \le g \le c''f.$$

The constants c' and c'' will be assumed to be independent of $\varphi \in \Gamma$ and of the choice of the initial data in (1.1). Also, we shall write $f \approx g$ *uniformly for* $m+k \ge K$ if we can choose the same constants c' and c'' for all m and k satisfying $m+k \ge K$.

Let us first evaluate the initial energy.

LEMMA 5.1. *We have*

$$\begin{aligned} E_0 \approx \sum_{m=0}^{\infty}\sum_{k=1}^{\infty} \rho_{mk}^2 \int_0^1 r|J_{m-1+\frac{n}{2}}(\rho_{mk}r)|^2 \ dr \\ \times \int_\Gamma |a_{mk}|^2 + |b_{mk}|^2 + |c_{mk}|^2 + |d_{mk}|^2 \ d\Gamma \end{aligned}$$

for all solutions of (1.1).

Proof. Using (5.13) we obtain by an easy computation the identity

$$E_0 = n|\Omega| \sum_{m=0}^{\infty} \sum_{k=1}^{\infty} \rho_{mk}^2 \int_0^1 r|J_{m-1+\frac{n}{2}}(\rho_{mk}r)|^2 \, dr \tag{5.15}$$

$$\times \int_\Gamma |u_{10mk}(\varphi)|^2 + |u_{20mk}(\varphi)|^2 + \rho_{mk}^{-2}|u_{11mk}(\varphi)|^2 + \rho_{mk}^{-4}|u_{21mk}(\varphi)|^2 \, d\varphi,$$

where $|\Omega|$ denotes the volume of Ω. Hence it suffices to show that

$$\begin{aligned} |u_{10mk}|^2 + |u_{20mk}|^2 + \rho_{mk}^{-2}|u_{11mk}|^2 + \rho_{mk}^{-4}|u_{21mk}|^2 \\ \asymp |a_{mk}|^2 + |b_{mk}|^2 + |c_{mk}|^2 + |d_{mk}|^2 \end{aligned} \tag{5.16}$$

uniformly for all m and k.

Since $\lambda_{mk} \neq 0$, $\mu_{mk} \neq 0$ and since the vectors v_{mk} and w_{mk} are linearly independent, we deduce easily from (5.14) the implications

$$\begin{aligned} u_{10mk} = u_{20mk} = u_{11mk} = u_{21mk} = 0 \\ \iff a_{mk} = b_{mk} = c_{mk} = d_{mk} = 0 \end{aligned} \tag{5.17}$$

for each pair (m, k).

Furthermore, using (5.3), (5.4), (5.10), and (5.11) we deduce from (5.14) that

$$\begin{aligned} u_{10mk} &= (c_{mk} + d_{mk}) + (a_{mk} + b_{mk})o(1), \\ u_{20mk} &= (a_{mk} + b_{mk}) + (c_{mk} + d_{mk})o(1), \\ u_{11mk} &= \rho_{mk}(c_{mk} - d_{mk})(i + o(1)) + (a_{mk} - b_{mk})o(1), \\ u_{21mk} &= \rho_{mk}^2(a_{mk} - b_{mk})(i + o(1)) + (c_{mk} - d_{mk})o(1) \end{aligned}$$

as $m + k \to \infty$. Hence

$$\begin{aligned} |u_{10mk}|^2 + |u_{20mk}|^2 + \rho_{mk}^{-2}|u_{11mk}|^2 + \rho_{mk}^{-4}|u_{21mk}|^2 \\ = \big(2 + o(1)\big)\big(|a_{mk}|^2 + |b_{mk}|^2 + |c_{mk}|^2 + |d_{mk}|^2\big) \end{aligned} \tag{5.18}$$

as $m + k \to \infty$.

Now (5.17) and (5.18) imply (5.16). □

Next we evaluate the boundary integral in (1.2).

LEMMA 5.2. *Fix $T > 2$ arbitrarily. Then*

$$\int_{-T}^{T} \int_\Gamma |\partial_\nu u|^2 \, d\Gamma \, dt$$

$$\asymp \sum_{m=0}^{\infty} \sum_{k=1}^{\infty} \rho_{mk}^2 |J'_{m-1+\frac{n}{2}}(\rho_{mk})|^2 \int_\Gamma |a_{mk}|^2 + |b_{mk}|^2 + |c_{mk}|^2 + |d_{mk}|^2 \, d\Gamma$$

for all solutions of (1.1).

We prove this lemma in several steps.

Step 1. It follows from (5.12) that

$$\partial_\nu u(t, 1, \varphi) = \sum_{m=0}^{\infty} \sum_{k=1}^{\infty} \rho_{mk} J'_{m-1+\frac{n}{2}}(\rho_{mk})$$

$$\times\big\{\big(a_{mk}(\varphi)e^{i\sqrt{\lambda_{mk}}t}+b_{mk}(\varphi)e^{-i\sqrt{\lambda_{mk}}t}\big)v_{mk}$$
$$+\big(c_{mk}(\varphi)e^{i\sqrt{\mu_{mk}}t}+d_{mk}(\varphi)e^{-i\sqrt{\mu_{mk}}t}\big)w_{mk}\big\}$$
$$=:\sum_{m=0}^{\infty}\sum_{k=1}^{\infty}r_{mk}f_{mk}(t,\varphi).$$

Since the spherical harmonics of different order are orthogonal in $L^2(\Gamma)$, we have

$$\int_\Gamma |\partial_\nu u|^2\ d\Gamma=\sum_{m=0}^{\infty}\int_\Gamma\Big|\sum_{k=1}^{\infty}r_{mk}f_{mk}(t,\varphi)\Big|^2\ d\varphi.$$

Hence, applying the Fubini–Tonelli theorem,

$$\int_0^T\int_\Gamma |\partial_\nu u|^2\ d\Gamma\ dt=\sum_{m=0}^{\infty}\int_\Gamma\left(\int_0^T\Big|\sum_{k=1}^{\infty}r_{mk}f_{mk}(t,\varphi)\Big|^2\ dt\right)\ d\varphi.$$

Therefore it suffices to prove that the estimates

$$\int_0^T\Big|\sum_{k=1}^{\infty}r_{mk}f_{mk}(t,\varphi)\Big|^2\ dt \tag{5.19}$$
$$\asymp\sum_{k=1}^{\infty}r_{mk}^2(|a_{mk}(\varphi)|^2+|b_{mk}(\varphi)|^2+|c_{mk}(\varphi)|^2+|d_{mk}(\varphi)|^2)$$

hold uniformly in m.

Step 2. Let us prove that the estimates (5.19) hold uniformly for $m\ge m'$ for some sufficiently large m'. We recall from [19] that if $m-1+(n/2)\ge 1/2$, then the difference sequence $(\rho_{m,k+1}-\rho_{mk})$ is nonincreasing and tends to π as $k\to\infty$. Choosing $\pi>\gamma>2\pi/T$ arbitrarily and using (5.3) and (5.4), we conclude that

$$2\sqrt{\lambda_{m1}},\ 2\sqrt{\mu_{m1}},\ \inf_k\sqrt{\lambda_{m,k+1}}-\sqrt{\lambda_{mk}},\quad\text{and}\quad\inf_k\sqrt{\mu_{m,k+1}}-\sqrt{\mu_{mk}}$$

are all larger than γ if m is sufficiently large, say $m\ge m''$. Furthermore, it follows from (5.3), (5.10), and (5.11) that $(v_{mk},w_{ml})\to 0$ uniformly in k and l if $m\to\infty$. Hence for any fixed $0<\eta<1/2$ we may apply Theorem 1.8 for each sufficiently large m with $M=2$,

$$\{\omega_n\}=\{\pm\sqrt{\lambda_{mk}},\pm\sqrt{\mu_{mk}}\}$$

and with

$$u_n=\begin{cases}v_{mk} & \text{if } \omega_n=\pm\sqrt{\lambda_{mk}},\\ w_{mk} & \text{if } \omega_n=\pm\sqrt{\mu_{mk}}.\end{cases}$$

It follows that

$$\int_0^T\Big|\sum_{k=1}^{\infty}r_{mk}f_{mk}(t,\varphi)\Big|^2\ dt$$
$$\asymp\sum_{k=1}^{\infty}r_{mk}^2\big((|a_{mk}(\varphi)|^2+|b_{mk}(\varphi)|^2)\|v_{mk}\|^2+(|c_{mk}(\varphi)|^2+|d_{mk}(\varphi)|^2)\|w_{mk}\|^2\big)$$

uniformly for all sufficiently large m.

It follows from (5.3), (5.10), and (5.11) that none of the vectors v_{mk} and w_{mk} vanishes and that $\|v_{mk}\| \to 1$ and $\|w_{mk}\| \to 1$ if $m+k \to \infty$. Therefore we deduce from the above relation that for a suitably chosen integer $m' \ge m''$ the relations (5.19) hold uniformly for all $m \ge m'$.

Step 3. Now it suffices to prove the estimates (5.19) for each fixed $m < M$. We recall from [19] that $\rho_{m,k+1} - \rho_{mk} \to \pi$ as $k \to \infty$, for each m. Using (5.4) we may therefore choose K such that

$$2\sqrt{\lambda_{m,K+1}} > 2\pi/T, \quad \inf_{k>K} \sqrt{\lambda_{m,k+1}} - \sqrt{\lambda_{mk}} > 2\pi/T$$

and

$$2\sqrt{\mu_{m,K+1}} > 2\pi/T, \quad \inf_{k>K} \sqrt{\mu_{m,k+1}} - \sqrt{\mu_{mk}} > 2\pi/T.$$

Furthermore, by (5.3), (5.10), and (5.11) we have $(v_{mk}, w_{ml}) \to 0$ if $k, l \to \infty$. Therefore, choosing a larger K if necessary, we may apply Theorem 1.8 for the sequence (ω_n) containing the numbers $\pm\sqrt{\lambda_{mk}}$ and $\pm\sqrt{\mu_{mk}}$ for all $k > K$ and for the sequence (u_n) defined as in the preceding step.

We conclude that (5.19) holds true under the extra hypothesis

$$a_{mk} = b_{mk} = c_{mk} = d_{mk} = 0 \tag{5.20}$$

for all $k \le K$.

Step 4. It remains to remove the hypothesis (5.20). Thanks to hypothesis (5.6) we may apply Proposition 1.9 for (ω_n), (u_n) as in Step 2 and with N given by

$$\{\omega_n \mid n \in N\} = \{\pm\sqrt{\lambda_{mk}}, \pm\sqrt{\mu_{mk}} \mid k \le K\}$$

and the proposition follows.

Now Theorem 1.2 follows from Lemmas 5.1 and 5.2 and from the identity (see [19] or [7])

$$2\int_0^1 r|J_{m-1+\frac{n}{2}}(\rho_{mk}r)|^2\,dr = |J'_{m-1+\frac{n}{2}}(\rho_{mk})|^2.$$

In dimension $n = 1$ the proof is similar. Assuming for simplicity that $\Omega = (0,\pi)$, now the solutions of (1.1) are given by the formula

$$\begin{aligned} u(t,x) = \sum_{m=1}^{\infty} \Big\{ &\big(a_m e^{i\sqrt{\lambda_m}t} + b_m e^{-i\sqrt{\lambda_m}t}\big) v_m \\ &+ \big(c_m e^{i\sqrt{\mu_m}t} + d_m e^{-i\sqrt{\mu_m}t}\big) w_m \Big\} \sin mx \end{aligned} \tag{5.21}$$

instead of (5.12), where

$$\lambda_m = \tfrac{1}{2}\Big(m^4 + m^2 + A + D + \sqrt{(m^4 - m^2 + D - A)^2 + 4BC}\Big)$$

and

$$\mu_m = \tfrac{1}{2}\Big(m^4 + m^2 + A + D - \sqrt{(m^4 - m^2 + D - A)^2 + 4BC}\Big)$$

are the eigenvalues of the matrices

$$A_m = \begin{pmatrix} m^2 + A & B \\ C & m^4 + D \end{pmatrix};$$

$$v_m = (v_{m1}, v_{m2}) := (C^{-1}(\lambda_m - m^4 - D), 1),$$

and

$$w_m = (w_{m1}, w_{m2}) := (1, B^{-1}(\mu_m - m^2 - A))$$

are corresponding eigenvectors; and a_m, b_m, c_m, d_m are arbitrary numbers such that

$$\sum_{m=1}^{\infty} |a_m|^2 + |b_m|^2 + |c_m|^2 + |d_m|^2 < \infty.$$

Then the proof given above can be adapted easily.

6. Proof of Proposition 1.3. As in the preceding section, assume first that Ω is the unit ball of $\mathbb{R}^n$, $n \ge 2$. Fix a nonnegative integer m and fix three positive roots $\rho_{mk_1} < \rho_{mk_2} < \rho_{mk_3}$ of $J_{m-1+\frac{n}{2}}(x)$ arbitrarily. We will show later that for some suitable choices of the parameters A, B, C, D, the matrices A_{mk_1}, A_{mk_2}, and A_{mk_3} (defined at the beginning of the preceding section) have a common eigenvalue μ.

Then the theorem will follow easily. Indeed, denoting by β_1, β_2, β_3 corresponding nonzero eigenvectors in $\mathbb{R}^2$, the formula

$$u(t, r, \varphi) = \sum_{j=1}^{3} r^{1-\frac{n}{2}} J_{m-1+\frac{n}{2}}(\rho_{mk_j} r) \delta_j \beta_j h(\varphi) \cos \sqrt{\mu} t$$

defines a solution of (1.1) for every spherical harmonics $h(\varphi)$ of order m and for every choice of real numbers δ_1, δ_2, δ_3. Furthermore, we have

$$\partial_\nu u(t, r, \varphi) = \sum_{j=1}^{3} \rho_{mk_j} J'_{m-1+\frac{n}{2}}(\rho_{mk_j}) \delta_j \beta_j h(\varphi) \cos \sqrt{\mu} t$$

on the boundary of Ω. Since the three vectors β_j cannot be linearly independent in $\mathbb{R}^2$, we can choose δ_1, δ_2, and δ_3 such that not all of them are zero but

$$\sum_{j=1}^{3} \rho_{mk_j} J'_{m-1+\frac{n}{2}}(\rho_{mk_j}) \delta_j \beta_j = 0.$$

Now choose an arbitrary nonzero spherical harmonics $h(\varphi)$ of order m. Then the corresponding function u satisfies (1.1) and (1.3). Finally, u does not vanish identically because the positive roots of the Bessel functions are simple and hence

$$\rho_{mk_j} J'_{m-1+\frac{n}{2}}(\rho_{mk_j}) \ne 0$$

for $j = 1, 2, 3$.

It remains to show that we can choose A, B, C, D so that the matrices A_{mk_1}, A_{mk_2}, and A_{mk_3} have a common eigenvalue. Writing $c_j = \rho_{mk_j}$ for brevity, first we choose two real numbers a and d such that $c_j^6 + ac_j^4 + dc_j^2$ has the same value (say, P) for $j = 1, 2, 3$. This is possible because the linear system

$$\begin{aligned} c_1^6 + ac_1^4 + dc_1^2 &= c_2^6 + ac_2^4 + dc_2^2, \\ c_1^6 + ac_1^4 + dc_1^2 &= c_3^6 + ac_3^4 + dc_3^2 \end{aligned}$$

is equivalent to

$$\begin{aligned} (c_1^2 + c_2^2)a + d + c_1^4 + c_1^2c_2^2 + c_2^4 &= 0, \\ (c_1^2 + c_3^2)a + d + c_1^4 + c_1^2c_3^2 + c_3^4 &= 0, \end{aligned}$$

whose determinant $c_2^2 - c_3^2$ is different from zero.

Now fix a real number μ arbitrarily. Set $A = a + \mu$, $D = d + \mu$, and choose B, C such that $BC = P$. Then we have

$$c_j^6 + (A - \mu)c_j^4 + (D - \mu)c_j^2 + (A - \mu)(D - \mu) - BC = 0$$

for $j = 1, 2, 3$. In other words, μ is a common root of the characteristic polynomials of the matrices A_{mk_1}, A_{mk_2}, A_{mk_3}, and our assertion follows.

Since the set of such quadruples (A, B, C, D) forms a two-dimensional surface in $\mathbb{R}^4$ for every choice of m and of the roots $\rho_{mk_1} < \rho_{mk_2} < \rho_{mk_3}$, the proposition follows.

The proof can be modified easily for dimension $n = 1$. Assume for simplicity that $\Omega = (0, \pi)$. Fix three positive integers $m_1 < m_2 < m_3$ *having the same parity*, and then choose A, B, C, D so that the matrices A_{m_1}, A_{m_2}, and A_{m_3} have a common eigenvalue μ. Denoting by β_1, β_2, β_3 corresponding nonzero eigenvectors in $\mathbb{R}^2$, the formula

$$u(t, x) = \sum_{j=1}^{3} \delta_j \beta_j \sin m_j x \cos \sqrt{\mu} t$$

defines a solution of (1.1) for any choice of δ_1, δ_2, δ_3. It remains to choose them so that

$$\sum_{j=1}^{3} \delta_j \beta_j m_j = 0,$$

although at least one of the numbers δ_j is different from zero.

For example, for $(m_1, m_2, m_3) = (2, 4, 6)$ and for any fixed real number μ we can choose the parameters

$$A = -56 + \mu, \quad B = -52, \quad C = 800, \quad D = 784 + \mu.$$

Then the nonzero function

$$u(t, x) = \left\{ \begin{pmatrix} 65 \\ -65 \end{pmatrix} \sin 2x + \begin{pmatrix} -52 \\ 40 \end{pmatrix} \sin 4x + \begin{pmatrix} 13 \\ -5 \end{pmatrix} \sin 6x \right\} \cos \sqrt{\mu} t$$

satisfies (1.1) and (1.3).

7. Proof of Proposition 1.4. For brevity, henceforth we only consider the case where Ω is the unit ball of $\mathbb{R}^n$, $n \ge 2$. We modify the proof of Theorem 1.2 as follows. For any fixed A, B, C, D we define the eigenvalues of the matrices A_{mk} by the same formulas (5.1) and (5.2). Then the relations (5.4) remain valid.

Furthermore, we keep the definition (5.7) of the eigenvectors unchanged if $B \ne 0$ and $C \ne 0$ (otherwise they make no sense), but we define them differently in the remaining cases: we set

$$v_{mk} := (B(\lambda_{mk} - \mu_{mk})^{-1}, 1) \quad \text{and} \quad w_{mk} := (1, 0)$$

if $B \ne 0$ and $C = 0$,

$$v_{mk} := (0, 1) \quad \text{and} \quad w_{mk} := (1, C(\mu_{mk} - \lambda_{mk})^{-1})$$

if $B = 0$ and $C \ne 0$, and finally

$$v_{mk} := (0, 1) \quad \text{and} \quad w_{mk} := (1, 0)$$

if $B = C = 0$. Then (5.8), (5.9), (5.10), and (5.11) remain true.

Although we do not have (5.6) in the general case, for any fixed $\eta > 0$ there exists a positive integer M such that

$$\begin{aligned}
0 \ne \lambda_{mk} \ne \mu_{mk} \ne 0 &\quad \text{whenever} \quad m + k > M,\\
\lambda_{m1}, \lambda_{m2}, \dots &\text{ has no repeated elements if } m > M,\\
\mu_{m1}, \mu_{m2}, \dots &\text{ has no repeated elements if } m > M,\\
\lambda_{mM+1}, \lambda_{mM+2}, \dots &\text{ has no repeated elements if } m \le M,\\
\mu_{mM+1}, \mu_{mM+2}, \dots &\text{ has no repeated elements if } m \le M,\\
|(v_{mk}, w_{ml})_H| \le \eta \|v_{mk}\| \, \|w_{ml}\| &\quad \text{whenever} \quad m + k, m + l > M.
\end{aligned}$$

Using these properties we can repeat the proof of Theorem 1.2 except for Step 4 in the proof of Lemma 5.2. The proposition follows if we choose N so large that all functions of the form

$$r^{1-\frac{n}{2}} J_{m-1+\frac{n}{2}}(\rho_{mk_j} r) h(\varphi), \quad m + k \le M,$$

where h runs over the spherical harmonics of order m, belong to the first N eigenspaces of $-\Delta$ in Ω with homogeneous Dirichlet boundary conditions.

8. Proof of Theorem 1.5. We consider only initial data whose integral over Ω is equal to zero. This orthogonality condition can be eliminated by applying the method of Proposition 1.9, but we omit the details in order to keep the length of this paper reasonable.

Now let us denote by

$$\rho_{m1} < \rho_{m2} < \cdots$$

the sequence of the (strictly) positive roots of

$$\left(1 - \frac{n}{2}\right) J_{m-1+\frac{n}{2}}(x) + x J'_{m-1+\frac{n}{2}}(x)$$

for $m = 0, 1, \dots$. Introducing the eigenvalues λ_{mk} and μ_{mk} by the same formulas (5.1) and (5.2), we still have (5.3) and (5.4). Furthermore, assuming (5.5), (5.6), and defining v_{mk}, w_{mk} by (5.7), we still have (5.8)–(5.12). It follows that the restriction of the solution of (1.4) to the boundary is given by the formula

$$\begin{aligned} u(t,1,\varphi) = \sum_{m=0}^{\infty}\sum_{k=1}^{\infty} & J_{m-1+\frac{n}{2}}(\rho_{mk}) \\ \times\big\{ & \big(a_{mk}(\varphi)e^{i\sqrt{\lambda_{mk}}t} + b_{mk}(\varphi)e^{-i\sqrt{\lambda_{mk}}t}\big)v_{mk} \\ + & \big(c_{mk}(\varphi)e^{i\sqrt{\mu_{mk}}t} + d_{mk}(\varphi)e^{-i\sqrt{\mu_{mk}}t}\big)w_{mk}\big\} \\ =: \sum_{m=0}^{\infty}\sum_{k=1}^{\infty} & R_{mk}f_{mk}(t,\varphi). \end{aligned}$$

(We used the above-mentioned orthogonality assumption here.) As in section 5, it follows that

$$\int_\Gamma |u|^2 \, d\Gamma = \sum_{m=0}^{\infty} \int_\Gamma \Big|\sum_{k=1}^{\infty} R_{mk}f_{mk}(t,\varphi)\Big|^2 \, d\varphi$$

and then

$$\int_I\int_\Gamma |u|^2 \, d\Gamma \, dt = \sum_{m=0}^{\infty} \int_\Gamma \left(\int_I \Big|\sum_{k=1}^{\infty} R_{mk}f_{mk}(t,\varphi)\Big|^2 \, dt\right) \, d\varphi.$$

If the length of I is larger than 2, then by repeating the proof of Lemma 5.2 we obtain that

$$\begin{aligned} &\int_I \Big|\sum_{k=1}^{\infty} R_{mk}f_{mk}(t,\varphi)\Big|^2 \, dt \\ &\qquad \approx \sum_{k=1}^{\infty} R_{mk}^2(|a_{mk}(\varphi)|^2 + |b_{mk}(\varphi)|^2 + |c_{mk}(\varphi)|^2 + |d_{mk}(\varphi)|^2) \end{aligned}$$

uniformly in m, and hence

$$\int_I\int_\Gamma |u|^2 \, d\Gamma \, dt \approx \sum_{m=0}^{\infty}\sum_{k=1}^{\infty} R_{mk}^2(|a_{mk}(\varphi)|^2 + |b_{mk}(\varphi)|^2 + |c_{mk}(\varphi)|^2 + |d_{mk}(\varphi)|^2).$$

Since the length of I does not appear on the right-hand side of this estimate, the theorem follows.

9. Proof of Proposition 1.6. As in the proof of Proposition 1.3, we are going to construct a nontrivial solution of (1.4) which vanishes identically on the boundary of Ω. Fix a nonnegative integer m arbitrarily and fix arbitrarily three different positive roots c_1, c_2, c_3 of the function

$$\Big(1-\frac{n}{2}\Big) J_{m-1+\frac{n}{2}}(x) + xJ'_{m-1+\frac{n}{2}}(x).$$

Repeating the proof of Proposition 1.3 we obtain for some exceptional values of the coupling parameters a positive number λ and three nonzero vectors β_1, β_2, β_3 in $\mathbb{R}^2$ such that the function

$$u(t,r,\varphi) := \big[\delta_1 J(c_1 r)\beta_1 + \delta_2 J(c_2 r)\beta_2 + \delta_3 J(c_3 r)\beta_3\big] h(\varphi)\cos\sqrt{\lambda}t \tag{9.1}$$

is a solution of (1.4) (with suitable initial data) for any spherical harmonics h of order m and for every choice of real numbers δ_1, δ_2, and δ_3.

Fix a nonzero spherical harmonics h of order m arbitrarily and choose δ_1, δ_2, and δ_3 such that not all of them vanish but that the linear combination

$$\delta_1 J(c_1)\beta_1 + \delta_2 J(c_2)\beta_2 + \delta_3 J(c_3)\beta_3$$

is equal to zero. This is possible because β_1, β_2, β_3 cannot be linearly independent in $\mathbb{R}^2$. Then the function (9.1) has the required properties.

10. Proof of Proposition 1.7. The proof is similar to that of Proposition 1.4, and hence it is omitted.

Acknowledgments. The authors thank the referees for their suggestions and J. E. Lagnese for his question leading to Propositions 1.4 and 1.7. Vilmos Komornik is grateful to the Istituto per le Applicazioni del Calcolo “Mauro Picone” del Consiglio Nazionale delle Ricerche in Rome for its hospitality.

REFERENCES

[1] S. A. Avdonin and S. A. Ivanov, *Families of Exponentials*, Cambridge University Press, Cambridge, UK, 1995.

[2] J. N. J. W. L. Carleson and P. Malliavin, eds., *The Collected Works of Arne Beurling*, Vol. 2, Birkhäuser, Basel, 1989.

[3] K. D. Graham and D. L. Russell, *Boundary value control of the wave equation in a spherical region*, SIAM J. Control, 13 (1975), pp. 174–196.

[4] A. Haraux, *Séries lacunaires et contrôle semi-interne des vibrations d'une plaque rectangulaire*, J. Math. Pures Appl. (9), 68 (1989), pp. 457–465.

[5] A. E. Ingham, *Some trigonometrical inequalities with applications in the theory of series*, Math. Z., 41 (1936), pp. 367–379.

[6] V. Komornik, *A new method of exact controllability in short time and applications*, Ann. Fac. Sci. Toulouse Math. (6), 10 (1989), pp. 415–464.

[7] V. Komornik, *On the zeros of Bessel type functions and applications to exact controllability problems*, Asymptotic Anal., 5 (1991), pp. 115–128.

[8] V. Komornik, *Exact Controllability and Stabilization. The Multiplier Method*, Masson, Paris and John Wiley, Chichester, UK, 1994.

[9] V. Komornik, *Rapid boundary stabilization of linear distributed systems*, SIAM J. Control Optim., 35 (1997), pp. 1591–1613.

[10] V. Komornik and P. Loreti, *Partial observability of coupled linear systems*, Acta Math. Hungar., submitted.

[11] V. Komornik, P. Loreti, and E. Zuazua, *On the control of coupled linear systems*, in Proceedings of the Conference on Control Theory (Vorau 1996), Internat. Ser. Numer. Math. 126, Birkhäuser, Boston, MA, 1998, pp. 183–189.

[12] I. Lasiecka and R. Triggiani, *Regularity of hyperbolic equations under $L_2(0,T;L_2(\Gamma))$ boundary terms*, Appl. Math. Optim., 10 (1983), pp. 275–286.

[13] G. Lebeau, *Contrôle de l'équation de Schrödinger*, J. Math. Pures Appl. (9), 71 (1992), pp. 267–291.

[14] J.-L. Lions, *Exact controllability, stabilization, and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[15] J.-L. Lions, *Contrôlabilité exacte et stabilisation de systèmes distribués*, Vols. 1–2, Masson, Paris, 1988.

[16] P. Loreti and V. Valente, *Partial exact controllability for spherical membranes*, SIAM J. Control Optim., 35 (1997), pp. 641–653.

[17] N. K. Nikolskii, *A Treatise on the Shift Operator*, Springer-Verlag, Berlin, 1986.

[18] K. Seip, *On the connection between exponential bases and certain related sequences in $l_2(-\pi,\pi)$*, J. Funct. Anal., 130 (1995), pp. 131–160.

[19] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, Cambridge University Press, Cambridge, UK, 1962.

[20] R. M. YOUNG, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

[21] E. ZUAZUA, *Contrôlabilité exacte en un temps arbitrairement petit de quelques modèles de plaques*, in Contrôlabilité exacte et stabilisation de systèmes distribués, Vol. 1, Masson, Paris, Appendix 1.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 2, pp. 486–501

A DUALITY APPROACH IN THE OPTIMIZATION OF BEAMS AND PLATES*

J. SPREKELS† AND D. TIBA†‡

Abstract. We introduce a class of nonlinear transformations called "resizing rules" which associate with optimal shape design problems certain equivalent distributed control problems while preserving the state of the system. This puts into evidence the duality principle that the class of system states that can be achieved, under a prescribed force, via modifications of the structure (shape) of the system can be obtained as well via the modifications of the force action, under a prescribed structure.

We apply such transformations to the optimization of beams and plates, and in the simply supported or cantilevered cases, the obtained control problems are even convex. In all cases, we establish existence theorems for optimal pairs by assuming only boundedness conditions. Moreover, in the simply supported case, we also prove the uniqueness of the global minimizer. A general algorithm that iterates between the original transformed problems is introduced and studied. The applications also include the case of variational inequalities.

Key words. optimal design, nonconvex duality, resizing rules

AMS subject classification. 49D05

PII. S036301299732540X

1. Introduction.

It is our aim to study a class of *control into the coefficients* problems. The state equation has the form

$$\Delta(bu^3\Delta y) = f \quad \text{in } \Omega, \tag{1.1}$$

where Ω is a smooth bounded domain in $R^n, n \geq 1, f \in L^2(\Omega), u \in L^\infty(\Omega)$, and $b > 0$ is a constant. If $n \leq 2$, such models are used in the literature for the deflection y of plates or beams of thickness $u > 0$ almost everywhere (a.e.) in Ω and are subject to the transverse load f. The coefficient b is a material constant, and we shall fix $b = 1$ in what follows. For instance, (1.1) was recently proposed by Bendsoe [3] for modeling plates made of a material with special properties. We also quote Hlavacek, Bock, and Lovisek [12], [13], Haslinger and Neittaanmäki [11], Casas [4], Neto and Polak [21], and Langenbach [17] for more complex beams or plate equations. To derive (1.1), smoothness conditions are imposed on the thickness u. However, in the associated minimization problems, we show that this assumption is not necessary and that is why we take just $u \in L^\infty(\Omega)$.

To (1.1) we add various boundary conditions:

$$y = \Delta y = 0 \quad \text{on } \partial\,\Omega \tag{1.2}$$

(simply supported plates), and

$$y = \frac{\partial y}{\partial n} = 0 \quad \text{on } \partial\,\Omega \tag{1.3}$$

*Received by the editors August 1, 1997; accepted for publication (in revised form) March 19, 1998; published electronically December 2, 1998.
http://www.siam.org/journals/sicon/37-2/32540.html

†Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, D–10117 Berlin, Germany (sprekels@wias-berlin.de).

‡Permanent address: Institute of Mathematics, Romanian Academy of Sciences, P.O. Box 1–764, RO–70700 Bucharest, Romania (dtiba@imar.ro).

(clamped plates; $\frac{\partial}{\partial n}$ denotes the outward normal derivative to $\partial\Omega$).

In space dimension one, cantilevered beams or unilaterally supported beams (variational inequalities) will be discussed as well.

We associate with (1.1) various optimization problems:

$$\text{Min} \int_\Omega u(x)\,dx \tag{1.4}$$

(minimization of the weight or volume), and

$$\text{Min} \int_\Omega \big(y(x) - y_d(x)\big)^2\,dx \tag{1.5}$$

(identification-type problems: the function $y_d \in L^2(\Omega)$ is a "desired" or "observed" deflection).

Moreover, natural control and state constraints will be imposed on u, y:

$$0 \le m \le u(x) \le M \quad \text{a.e. in } \Omega, \tag{1.6}$$

$$y(x) \ge -\tau \quad \text{a.e. in } \Omega, \tag{1.7}$$

(m, M, τ are positive constants),

$$y \in A. \tag{1.8}$$

$A \subset L^2(\Omega)$ is a prescribed closed subset, not necessarily convex.

Problems of this type are well known in the literature and their difficulty, from both a theoretical and a numerical point of view, was put into evidence in the works of Neto and Polak [21] (with an example of approximating local minimizers converging to a nonstationary point of the original problem), Murat [20] (indicating counterexamples to the existence of minimizers for control into coefficients problems governed by second order equations) and Cheng and Olhoff [5] and Rozvany et al. [22], where comprehensive numerical experiments are discussed.

In general, in nonconvex minimization problems one may expect just approximation of stationary points. In the case of optimal design of beams, this is discussed by Polak and Neto [21] via the use of consistent approximations.

In this work, we introduce a class of nonlinear transformations which may be applied to any of the problems (1.1)–(1.8). We call them "resizing rules" with reference to a partial similarity that exists with the fully stressed design method (FSD) appearing in the engineering literature (see Haftka, Gürdal, and Kamat [10, Chap. 9].

Via the resizing rules, the control into coefficients problem is transformed into an equivalent distributed control problem. In this way, we see that the problems corresponding to simply supported or cantilevered boundary conditions are convex or even strictly convex (after transformation). This gives the uniqueness of the global minimum in the original problem. The optimization problem associated with clamped plates remains nonconvex after the transformation. Moreover, this approach allows us to relax the compactness assumptions on the set of admissible controls needed to show the existence of the minimizers. The boundedness condition (1.6) is sufficient for our method to work. In this respect, similar results were previously obtained by Cox and Overton [6] and Senatorov [23] in the case of ordinary differential equations and by a different approach.

In sections 2 and 3 such results are proved for simply supported, respectively, clamped, plates and beams. In section 4, an algorithmic approach is used for the optimization of a unilaterally supported beam, and a numerical example is discussed. This shows the multiple possibilities of the "resizing rule" method. Such algorithms were previously used by Sprekels and Tiba [25] for classical types of beams (simply supported, cantilevered, and clamped).

Finally, we point out that ours is a duality-type method: associated with the original minimization problem is another optimization problem which is simpler and gives relevant information on the first problem. From a theoretical point of view, the equivalence results are essential in proving convexity, uniqueness, or existence. From a numerical point of view, simple "dual" problems may be considered that provide efficient approximations in the examples. Let us also notice that this duality approach has a mechanical background and is not inspired by the convex duality theory or its nonconvex extensions. A detailed comparison (from this point of view) was performed by Sprekels and Tiba (see [25, sections 1 and 2].

2. Simply supported plates and beams. We start with a general equivalence result which, roughly speaking, says that the set of deflections obtained under a given load and for various thicknesses is the same as the set of deflections obtained for a fixed thickness, but with variable load. Namely, we consider the following two "state" systems:

$$\Delta(u^3\Delta y) = f \quad \text{in } \Omega, \tag{2.1}$$

$$y = \Delta y = 0 \quad \text{on } \partial\Omega, \tag{2.2}$$

$$0 < m \le u(x) \le M \quad \text{a.e. in } \Omega, \tag{2.3}$$

$$y \in A \tag{2.4}$$

and

$$\Delta\Delta y = h \quad \text{in } \Omega, \tag{2.5}$$

$$y = \Delta y = 0 \quad \text{on } \partial\Omega, \tag{2.6}$$

$$\min\{m^{-3}z(x), M^{-3}z(x)\} \le \Delta y(x) \le \max\{m^{-3}z(x),\, M^{-3}z(x)\} \text{ a.e. in } \Omega, \tag{2.7}$$

$$y \in A. \tag{2.8}$$

In (2.1)–(2.4), $f \in L^2(\Omega)$ is fixed, $u \in L^\infty(\Omega)$ is the optimization parameter, $A \subset L^2(\Omega)$ is closed, and m, M are positive real constants. No sign conditions are imposed on f, and the unique weak solution y satisfies $y \in V := H^2(\Omega) \cap H_0^1(\Omega), u^3\Delta y \in V$. One (convex) example for the set A is obtained via the constraint

$$y \ge -\tau \quad \text{in } \Omega \tag{2.9}$$

with $\tau > 0$ given.

In (2.5)–(2.8) we assume that $h \in V^*$, and we define $g \in L^2(\Omega)$ as the unique transposition solution (see Lions and Magenes [19]) to

$$\Delta g = h \quad \text{in } \Omega, \tag{2.10}$$

$$g = 0 \quad \text{on } \partial\Omega; \tag{2.11}$$

that is, by definition, we have

$$\int_\Omega g\,\Delta\rho\,dx \;=\; \int_\Omega h\,\rho\,dx \quad \forall\;\rho\;\in\;V. \tag{2.12}$$

Then, $y\;\in\;V$ is the strong solution to

$$\Delta y \;=\; g \quad \text{in } \Omega, \tag{2.13}$$

$$y \;=\; 0 \quad \text{on } \partial\,\Omega. \tag{2.14}$$

The second boundary condition, $\Delta y\;=\;0$ on $\partial\,\Omega$, is included in the choice of test mappings ρ in (2.12) and is not explicit. The mapping $z \in V$ from (2.7) is the strong solution to (2.10), (2.11), corresponding to $h = f$. We also mention that (2.10) is valid in the sense of distributions although $C_0^\infty(\Omega)$ is not dense in V. The constraint (2.7) shows that, for admissible y, the boundary condition $\Delta y = 0$ on $\partial\Omega$ has an explicit meaning.

THEOREM 2.1. *For any admissible pair* $[y, u]$ *for* (2.1)–(2.4), *there is some* $h \in V^*$ *such that the pair* $[y, h]$ *is admissible for* (2.5)–(2.8). *The converse is also true if meas* $\{x \in \Omega;\; z(x) = 0\} = 0$.

Proof. If $[y, u]$ is admissible for (2.1)–(2.4), then

$$\Delta y \;=\; \frac{1}{u^3}\,z\;\in\;L^2(\Omega). \tag{2.15}$$

We denote by $\tilde h \in V^*$ the linear bounded functional on V defined by

$$\langle \tilde h, \rho\rangle_{V^*\times V} \;=\; \int_\Omega \frac{1}{u^3}\,z\,\Delta\rho\,dx \quad \forall\;\rho\;\in\;V. \tag{2.16}$$

Then, (2.16), (2.12) show that $\tilde g\;=\;(1/u^3)z$ is the transposition solution to (2.10)–(2.11) associated with this $\tilde h$. By (2.15) and (2.2), it follows that y satisfies (2.5), (2.6) with $\tilde h$ given by (2.16). Then, (2.7) is a clear consequence of (2.15) and (2.3).

Conversely, taking $[\hat y, \hat h]$ admissible for (2.5)–(2.8), and $\hat g$ satisfying (2.10), (2.11) with $h = \hat h$, we see that

$$\Delta\hat y \;=\; \hat g \quad \text{a.e. in } \Omega. \tag{2.17}$$

We shall multiply (2.17) by $z(x) \cdot [\hat g(x)]^{-1}$ which we denote $v(x)$. By (2.7) and (2.17), we notice that $v \in L^\infty(\Omega)$, and $\hat u = v^{1/3}$ satisfies constraint (2.3). To see this, we analyze in (2.7) the situations $z(x) > 0, z(x) < 0$ to get

$$0\;<\;M^{-3}\;\le\;\frac{\hat g(x)}{z(x)}\;=\;[\hat u(x)]^{-3}\;\le\;m^{-3}. \tag{2.18}$$

Under our hypothesis, (2.18) is valid a.e. in Ω, and we obtain (2.3). Moreover,

$$\hat u^3(x)\Delta\hat y(x)\;=\;z(x) \quad \text{a.e. in } \Omega. \tag{2.19}$$

The definition of z and (2.19) show that $\hat y$ is a weak solution of (2.1) as well, and the proof is finished. □

Corollary 2.2. *For any admissible pair* $[y, u]$ *for* (2.1)–(2.4)*, there is some* $l \in L^2(\Omega)$ *such that the pair* $[y, l]$ *is admissible for the system*

$$\Delta y = zl \quad \textit{in } \Omega, \tag{2.20}$$

$$y = 0 \quad \textit{on } \partial\Omega, \tag{2.21}$$

$$M^{-3} \leq l(x) \leq m^{-3} \quad \textit{a.e. in } \Omega, \tag{2.22}$$

$$y \in A. \tag{2.23}$$

The converse is also true.

The proof is just a variant of the proof of Theorem 2.1.

Remark. While Theorem 2.1 has a physical interpretation which we have stressed from the very beginning, Corollary 2.2 represents a mathematical equivalence trick. Its advantages are to transform the fourth order equation into a second order one and to replace the "state" constraint (2.7) by the "control" constraints (2.22). Notice that no special assumption on z is necessary.

Remark. Theorem 2.1 and Corollary 2.2 are controllability-type results. They say that the reachable set of states is the same in the systems (2.1)–(2.4), (2.5)–(2.8), or (2.20)–(2.23).

Remark. One basic property for the above results is that the set of admissible pairs $[y, h]$ defined by (2.5)–(2.7), as well as the set of admissible $[y, l]$ given by (2.20)–(2.22), are convex. If A is convex (which is generally the case; see (2.9)), then the systems (2.5)–(2.8) or (2.20)–(2.23) define convex pair sets in the appropriate product spaces.

This fundamental property is not valid, in general, for the original set of admissible pairs $[y, u]$ since the transformation that we use is nonlinear. However, there is one example, due to Kawohl [15] and Kawohl and Lang [16], where the system (2.1)–(2.3) and (2.9) define a convex set of admissible "control" mappings u in $L^2(\Omega)$.

Example 2.3. We assume that $f \leq 0$ a.e. in Ω. Then the maximum principle gives that $z > 0$ in Ω, and we have the representation formula

$$y(x) = -\int_\Omega \sigma(x, y) \frac{z(y)}{u^3(y)} dy, \tag{2.24}$$

with σ being the Green function, again positive. Let u_1, u_2 be two admissible thicknesses for the system (2.1)–(2.3), (2.9), and $u_\lambda(x) = \lambda\, u_1(x) + (1 - \lambda)\, u_2(x)$ $\forall x \in [0, 1]$, $\forall\, \lambda \in [0, 1]$. We denote by y_1, y_2, y_λ the solutions of (2.1), (2.2) corresponding to u_1, u_2, u_λ, respectively.

Then, (2.24) and the positivity of σ, z give

$$y_\lambda(x) \geq \lambda\, y_1(x) + (1 - \lambda)\, y_2(x) \geq -\tau, \tag{2.25}$$

since the function $-u^{-3}$ is concave. We conclude that u_λ is admissible for any $\lambda \in [0, 1]$, that is, the admissible set of controls is convex. For the set of admissible states y this is also true since the function $\bar{y}_\lambda = \lambda\, y_1 + (1 - \lambda)\, y_2$ corresponds to the thickness $(\bar{u}_\lambda)^3 = (\lambda\, u_1^{-3} + (1 - \lambda) u_2^{-3})^{-1}$ which satisfies (2.3) and (2.9). However, since the operator $u \mapsto y$ is nonlinear, we cannot expect $\bar{y}_\lambda = y_\lambda$ in (2.25), and the set of admissible pairs $[y, u]$ is not convex. Moreover, such properties do not extend beyond the condition (2.9) to general convex state constraints expressed by (2.4) or to nonnegative f.

It is our aim now to apply this equivalence, especially in the form given by Corollary 2.2, which is the simplest one, to certain optimization problems. Let us associate with the system (2.1)–(2.4) one of the following cost functionals, to be minimized:

$$\text{Min} \int_\Omega u(x)\, dx, \tag{2.26}$$

$$\text{Min} \int_\Omega \left(-\, u^{-\,3}(x)\right)\, dx, \tag{2.27}$$

$$\text{Min} \int_\Omega \left(y(x) \,-\, y_d(x)\right)^2\, dx. \tag{2.28}$$

The minimization parameter is $u \in L^\infty(\Omega)$, and we denote by $(\mathbf{P}_i), i = \overline{1,3}$, the obtained minimization problems, in this order. Obviously, $(\mathbf{P}_1)$ is the minimization of weight (volume) problem, subject to the given constraints. $(\mathbf{P}_2)$ is related to this question, as will be explained later, and $(\mathbf{P}_3)$ is an identification-type problem ($y_d \in L^2(\Omega)$ is an "observed" or "desired" deflection of the plate).

With the system (2.20)–(2.23) we associate the following cost functionals:

$$\text{Min} \int_\Omega l^{-\,\frac{1}{3}}(x)\, dx, \tag{2.29}$$

$$\text{Min} \int_\Omega \left(-\, l(x)\right)\, dx, \tag{2.30}$$

$$\text{Min} \int_\Omega \left(y(x) \,-\, y_d(x)\right)^2 dx. \tag{2.31}$$

The minimization distributed control is the mapping $l \in L^2(\Omega)$, and we denote by $(\mathbf{D}_i)$, $i = \overline{1,3}$, the obtained optimization problems, in this order.

THEOREM 2.4. *The problems $(\mathbf{P}_i)$ are equivalent to the problems $(\mathbf{D}_i)$, $i = \overline{1,3}$, in the sense that if $[y,u]$ is admissible for $(\mathbf{P}_i)$, then $[y,l]$, $l = 1/u^3$, is admissible for $(\mathbf{D}_i)$ with the same cost, and conversely.*

This follows directly from Corollary 2.2 and the definitions (2.26)–(2.31).

COROLLARY 2.5. *Under admissibility assumptions, if A is convex, then problem $(\mathbf{P}_1)$ has a unique global minimum $u^* \in L^\infty(\Omega)$.*

Proof. The existence of u^* can be established from standard estimates in (2.1), (2.2) and the boundedness of minimizing sequences given by (2.3). The passage to the limit is a simplified variant of the one performed in Theorem 3.2. By Theorem 2.4, $l^* = (u^*)^{-3}$ is the (global) minimizer for $(\mathbf{D}_1)$. By Corollary 2.2 and its subsequent remarks, the set of admissible pairs $[y,l]$ is convex if A is convex. Since $l^{-\frac{1}{3}}$ is a strictly convex function for $l > 0$, then the integral cost (2.29) is strictly convex and the uniqueness of l^*, u^* follows. ☐

Remark. Instead of solving the nonconvex problems $(\mathbf{P}_i)$, $i = \overline{1,3}$, we suggest solving the equivalent convex problems $(\mathbf{D}_i)$, $i = \overline{1,3}$. In numerical experiments, this avoids the "trap" of local minimum points, and the uniqueness of the global optimum enhances the numerical stability.

Remark. It is known that, in discussing weight minimization problems, any increasing function $\mu(u)$ may be relevant as an integrand in the cost functional. The problem $(\mathbf{P}_2)$ uses the increasing mapping

$$\mu(u) = -\frac{1}{u^3}, \quad u > 0,$$

which has the advantage that the equivalent problem $(\mathbf{D}_2)$ is a linear optimization problem.

Remark. Similar results may be obtained in dimension one for the simply supported beam and the cantilevered beam, i.e., for the boundary conditions

$$y(0) = y'(0) = 0,$$

$$y''(1) = (u^3 y'')'(1) = 0.$$

One basic property which is important for the above analysis is that the state system can be decoupled into two independent second order differential equations. In the next sections, this property is no longer true; however, the results can be extended.

3. Clamped plates and beams. We investigate first the classical optimal shape design problem:

$$\text{Min} \int_\Omega u(x)\,dx \tag{3.1}$$

subject to

$$\Delta(u^3\,\Delta y) = f \quad \text{in } \Omega, \tag{3.2}$$

$$y = \frac{\partial y}{\partial n} = 0 \quad \text{on } \partial\Omega, \tag{3.3}$$

$$0 < m \le u(x) \le M \quad \text{a.e. in } \Omega, \tag{3.4}$$

$$y \in A\,. \tag{3.5}$$

As usual, $f \in L^2(\Omega)$, $A \subset L^2(\Omega)$ are given, and $u \in L^\infty(\Omega)$ is the thickness of the plate, the minimization parameter.

The existence of a unique weak solution $y \in H_0^2(\Omega)$ to (3.2), (3.3) is obvious since the bilinear form

$$a(u,\, y,\, v) = \int_\Omega u^3\,\Delta y\,\Delta v\,dx$$

is coercive on $H_0^2(\Omega) \times H_0^2(\Omega)$.

Fix the mapping $g \in H^2(\Omega) \cap H_0^1(\Omega)$ by

$$\Delta g = f \quad \text{in } \Omega, \tag{3.6}$$

$$g = 0 \quad \text{on } \partial\Omega. \tag{3.7}$$

Here, we use that $f \in L^2(\Omega)$ and use the regularity of linear elliptic equations.

THEOREM 3.1. (a) *The equation* (3.2)*,* (3.3) *is equivalent to*

$$\Delta y = g\,l + h\,l \quad \text{in } \Omega \tag{3.8}$$

and (3.3) *(both conditions), where* $h \in L^2(\Omega)$ *is a harmonic mapping in* Ω *and* $l = u^{-3} \in L^\infty(\Omega)$.

(b) *The optimization problem* (3.1)–(3.5) *is equivalent to*

$$\text{Min} \int_\Omega l^{-\frac{1}{3}}(x)\,dx \tag{3.9}$$

subject to (3.8)*,* (3.3)*,* (3.5)*, and*

$$M^{-3} \le l(x) \le m^{-3} \quad \text{a.e. in } \Omega. \tag{3.10}$$

Proof. (a) By (3.2), (3.3), and the definition of $a(u,\cdot,\cdot)$, we see that

$$\int_\Omega (u^3\,\Delta y - g)\,\Delta v\,dx = 0 \quad \forall\; v \in H_0^2(\Omega). \tag{3.11}$$

We denote $h = u^3\,\Delta y - g \in L^2(\Omega)$, and (3.11) gives $\Delta h = 0$ in the sense of distributions. The converse is obvious.

(b) This is a clear consequence of (a) and of $l^{-1/3} = u$. □

Remark. The above transformation shows that the obtained problem remains nonconvex. Conceptually, the harmonic mapping h may be determined from the "supplementary" boundary condition $\frac{\partial y}{\partial n} = 0$ on $\partial\,\Omega$. One such situation is explained in Corollary 3.4. In general, we interpret h as an extra control variable and $\frac{\partial y}{\partial n} = 0$ on $\partial\,\Omega$ as a new state constraint.

THEOREM 3.2. *Under admissibility assumptions, the problem* $(\mathbf{P}_4)$ *given by* (3.1)–(3.5) *has at least one solution* $\tilde{u} \in L^\infty(\Omega)$.

Proof. By admissibility, there exists a minimizing sequence $\{u_n\} \subset L^\infty(\Omega)$ such that

$$\int_\Omega u_n(x)\,dx \to \inf(\mathbf{P_4}) \tag{3.12}$$

for $n \to \infty$. We denote by $l_n = u_n^{-3}$ and $y_n \in H_0^2(\Omega)$ the corresponding weak solution of (3.2), (3.3). Conditions (3.4), (3.10) show that $\{u_n\}$, $\{l_n\}$ are bounded in $L^\infty(\Omega)$, and hence we may assume that $u_n \rightharpoonup \hat{u}$, $l_n \rightharpoonup \hat{l}$ weakly* in $L^\infty(\Omega)$. In general, $\hat{l} \neq \hat{u}^{-3}$!

We also notice that $\{y_n\}$ is bounded in $H_0^2(\Omega)$:

$$m \int_\Omega \left[\Delta y_n(x)\right]^2 dx \le \int_\Omega u_n^3 (\Delta y_n)^2\,dx = \int_\Omega f\,y_n\,dx \le |f|_{L^2(\Omega)}\,|y_n|_{L^2(\Omega)}.$$

We may, as well, assume that $y_n \rightharpoonup \tilde{y}$ weakly in $H_0^2(\Omega)$, where $\tilde{y} \in A$ since A is closed in $L^2(\Omega)$. Moreover, by (3.8), we see that $h_n = u_n^3\,\Delta y_n - g$ is bounded in $L^2(\Omega)$, and we may write $h_n \rightharpoonup \tilde{h}$ weakly in $L^2(\Omega)$. We now use a lemma that will be proved later.

LEMMA 3.3. *If a sequence of harmonic mappings is weakly convergent in $L^1(\Omega)$, then it is pointwisely convergent.*

We remark that the right-hand side in (3.8) is bounded in $L^2(\Omega)$, and hence we may assume that, with some $z \in L^2(\Omega)$,

$$g\, l_n + h_n\, l_n \rightharpoonup z \quad \text{weakly in } L^2(\Omega). \tag{3.13}$$

The difficulty is just to identify z, that is, the limit of the product $h_n\, l_n$. By Lemma 3.3 and the Egorov theorem, for any $\varepsilon > 0$, there is $\Omega_\varepsilon \subset \Omega$ measurable, such that meas $(\Omega \backslash \Omega_\varepsilon) < \varepsilon$ and $h_n \to \tilde{h}$ uniformly in Ω_ε. Then, we can pass to the limit in (3.13) on Ω_ε, and we get $z = g\,\hat{l} + \tilde{h}\,\hat{l}$ in Ω_ε. Since ε is arbitrarily small, we obtain that $z(x) = g(x)\,\hat{l}(x) + \tilde{h}(x)\,\hat{l}(x)$ a.e. in Ω. Hence we can pass to the limit in (3.8) to obtain

$$\Delta\tilde{y} = g\,\hat{l} + \tilde{h}\,\hat{l} \quad \text{in } \Omega. \tag{3.14}$$

Using Theorem 3.1 in (3.14), we see that $\tilde{u} = \hat{l}^{-1/3}$ is the thickness in (3.2) which generates the deflection $\tilde{y}$. Obviously, the pair $[\tilde{y}, \tilde{u}]$ is admissible for the problem $(\mathbf{P}_4)$, and (3.12) yields

$$\inf\,(\mathbf{P}_4) = \lim \int_\Omega u_n(x)dx = \lim \int_\Omega l_n^{-\frac{1}{3}}(x)dx \geq \liminf \int_\Omega l_n^{-\frac{1}{3}}(x)\,dx \,\geq\, \int_\Omega \hat{l}^{-\frac{1}{3}}(x)\,dx$$

$$= \int_\Omega \tilde{u}(x)\,dx \,\geq\, \inf\,(\mathbf{P}_4).$$

Here, we also use the weak lower semicontinuity of the integral functional (3.9). This ends the proof. □

Proof of Lemma 3.3. Since $h_n, \tilde{h}$ are harmonic in Ω, the Weyl lemma, (see Hörmander [14]) shows that they belong to $C^\infty(\Omega)$. For any $x \in \Omega$ and any ball centered in x and of radius ρ, $B_\rho(x) \subset \Omega$, we have the solid mean property

$$h_n(x) = \frac{m}{w_m\,\rho^m} \int_{B_\rho(x)} h_n(y)\,dy \to \frac{m}{w_m\,\rho^m} \int_{B_\rho(x)} \tilde{h}(y)\,dy = \tilde{h}(x).$$

Here, m is the dimension of Ω, and w_m denotes the area of the unit ball in R^m. □

Remark. The passage to the limit in Theorem 3.2 is based on the following general property: If $\{w_n\}$ is bounded in $L^p(\Omega)$, $p > 1$ and $w_n(x) \to w(x)$ a.e. in Ω, then $w_n \to w$ strongly in $L^s(\Omega)$ for any s such that $1 < s < p$.

Proof. Let $\varepsilon > 0$ be fixed and let $\Omega_\varepsilon \subset \Omega$ measurable, with meas $(\Omega \setminus \Omega_\varepsilon) < \varepsilon$ be such that $w_n \to w$ uniformly in Ω_ε (by Egorov's theorem). We have

$$\int_\Omega |w_n - w|^s\,dx = \int_{\Omega_\varepsilon} |w_n - w|^s\,dx + \int_{\Omega\backslash\Omega_\varepsilon} |w_n - w|^s\,dx \leq \int_{\Omega_\varepsilon} |w_n - w|^s\,dx$$

$$+ \left(\int_{\Omega \setminus \Omega_\varepsilon} |w_n - w|^p \right)^{\frac{s}{p}} \text{meas}\,(\Omega \setminus \Omega_\varepsilon)^{\frac{p-s}{p}} \leq \int_{\Omega_\varepsilon} |w_n - w|^s\,dx + C\,\varepsilon^{\frac{p-s}{p}}.$$

If $n \geq N(\varepsilon)$, we get $\int_\Omega |w_n - w|^s\,dx \leq c(\varepsilon)$, where $c(\varepsilon) \to 0$ for $\varepsilon \to 0$.
This is a slight extension of Lemma 1.3 of Lions [18]. □

Remark. By Theorem 3.2, we see that the "optimal" thickness $\tilde{u}$ is obtained by twice inverting the minimizing sequence $\{u_n\}$. If u_n is pointwisely convergent, then $\tilde{u} = \hat{u} = \lim u_n$. This is the case used in the existing literature; see Haslinger and Neittaanmäki [11], Casas [4], Hlavacek, Bock, and Lovisek [12], [13], Neto and Polak [21], and Bendsoe [2]. Our result just shows that the strong compactness assumption (the boundedness of $\{\nabla u_n\}$) is not necessary to get existence in the optimal shape design problem. The numerical experiments from [5], [22] put into evidence the so-called "stiffeners" into the process of optimization of beams and plates, which correspond to unbounded gradients.

Remark. Obviously, the same argument applies to the cost functionals (2.27) or (2.28).

COROLLARY 3.4. *In the case of beams, the equation*

$$(u^3 y'')'' = f \quad in\ ]0,1[,$$

$$y(0) = y(1) = y'(0) = y'(1) = 0$$

is equivalent to

$$y'' = g\,l + (a_l\,x + b_l)\,l \quad in\ ]0,1[$$

with the same boundary conditions and with a_l, $b_l \in R$, *g satisfying* (3.6), (3.7), *and* $l = u^{-3}$.

Remark. It is clear, by direct integration, that the harmonic mapping $h_l = a_l\,x + b_l$ can be uniquely determined from the "supplementary" boundary conditions $y'(0) = y'(1) = 0$. In general, by a finite element approximation, the discretization of h will introduce a finite number of new entries into the state system (3.8) that can, in principle, be determined from the discretization of $\frac{\partial y}{\partial n} = 0$, which will generate the same finite number of nonlinear algebraic equations.

In the recent work of Arnautu et al. [1], a finite element approach is examined which combines the decomposition (3.8) together with a penalization of the boundary condition $\frac{\partial y}{\partial n} = 0$ in the cost.

Remark. Sprekels and Tiba [25] proved that if $f \leq 0$ in $[0,1]$, y'' has exactly two distinct roots in $[0,1]$, and that $y \leq 0$ in $[0,1]$ (see also Theorem 4.5). For general $f \in L^2(0,1)$, it is easy to see that y'' has at least two distinct roots in $[0,1]$. Otherwise $u^3\,y''$ (which is continuous) has at most one change of sign in $[0,1]$, and the maximum principle together with the Hopf maximum principle will contradict the boundary conditions.

Then, denoting by $\xi < \zeta$ two such roots, one can find a_l, b_l, and h_l from the simple relations

$$g(\xi) + a_l \cdot \xi + b_l = 0,$$

$$g(\zeta) + a_l \cdot \zeta + b_l = 0.$$

In general, the determination of h is related to the zeros of Δy in Ω. This is an extension to the case of the clamped plate of the relation (2.7) which ensures (in the case of simply supported plates) that the zeros and the sign of Δy remain unmodified via the resizing transformation. The roots distribution is connected to the famous conjecture of Hadamard [9] on the positivity of the Green function for the biharmonic operator. While Duffin [7] provided a first counterexample, he also noticed that the sign of Δy in a neighborhood of $\partial\Omega$ is the same as that of y. Later, Garabedian [8] and Shapiro and Tegmark [24] obtained counterexamples in eccentric ellipses. By reworking this last one, which has an elementary character, we see that Δy may change sign on an interior subdomain but also in the neighborhood of $\partial\Omega$ (even with f of constant sign). Therefore, the properties of Δy in dimension two are essentially different from [25, Thm. 3.1] in the one-dimensional case.

4. Variational inequalities. We consider the elastic beam with a unilateral obstacle at the right end:

$$(4.1)\qquad \left(u^3 y'', y'' - z''\right)_{L^2(0,1)} \le (f, y - z)_{V^* \times V} \quad \forall\, z \in \mathcal{K},$$

$$(4.2)\qquad y \in \mathcal{K} = \{w \in V\,;\, y(1) \ge \alpha\} \quad (\alpha \in R \text{ given}),$$

$$(4.3)\qquad V = \{y \in H^2(0,1)\,;\, y(0) = y'(0) = 0\}.$$

The beam is clamped at the left end.

To any $u \in L^\infty(0,1)$ we associate the linear bounded operator $A(u) : V \to V^*$ via the bilinear form on V

$$(4.4)\qquad a\,(u, y, z) = \int_0^1 u^3\, y''\, z''\, dx \quad \forall\, y, z \in V.$$

Then the variational inequality (4.1), (4.2) may be rewritten in the abstract form

$$(4.5)\qquad \big(A\,(u)\, y, y - z\big) = a\,(u, y, y - z) \le (f, y - z)_{V^* \times V}$$

for any $z \in \mathcal{K}$ and with $y \in \mathcal{K}$.

If $u \in L^\infty(0,1)$ is positive, $A(u)$ is strictly maximal monotone, and if $u(x) \ge m > 0$ in $[0,1]$, then $A(u)$ is strongly monotone and coercive. This gives a unique weak solution $y \in V$ to the variational inequality (4.5), for any $f \in V^*$.

We now define two auxiliary problems. First, we consider a cantilevered beam (without unilateral conditions):

$$(4.6)\qquad \begin{aligned} \left(u^3 y_1''\right)'' &= f \quad \text{in }]0,1[,\\ y_1(0) = y_1'(0) &= 0,\\ y_1''(1) = 0, \quad \left(u^3 y_1''\right)'(1) &= 0. \end{aligned}$$

Second, we introduce a clamped–simply supported beam:

$$(4.7)\qquad \begin{aligned} \left(u^3 y_2''\right)'' &= f \quad \text{in }]0,1[,\\ y_2(0) = y_2'(0) &= 0,\\ y_2''(1) = 0, \quad y_2(1) &= \alpha. \end{aligned}$$

It is simple to check by direct integration that both y_1, y_2 are in $H^2(0,1)$ and $u^3\, y_1''$, $u^3\, y_2'' \in H^2(0,1)$ for $f \in L^2(0,1)$.

Theorem 4.1. *If $f \in L^2(0,1)$, then the solution y of the variational inequality* (4.1) *is either the solution of* (4.6) *or the solution of* (4.7). *It satisfies $u^3\, y'' \in H^2(0,1)$.*

Proof. Assume first that $y_1(1) \geq \alpha$ (that is, $y_1 \in \mathcal{K}$). We multiply (4.6) by $y_1 - z$ for any $z \in \mathcal{K}$, and we see (by partial integration) that y_1 is also a solution of (4.1), $y = y_1$, and the claimed regularity is clear.

Assume now that $y_1 \notin \mathcal{K}$. By (4.7), it is obvious that $y_2 \in \mathcal{K}$. We multiply (4.7) by $y_2 - z$, $z \in \mathcal{K}$, and integrate by parts:

$$(f, y_2 - z)_{L^2(0,1)} = \big(u^3 y_2''\big)'(1)\big(\alpha - z(1)\big) + \int_0^1 u^3 y_2'' (y_2'' - z)\, dx. \tag{4.8}$$

Assume that

$$\gamma = \big(u^3 y_2''\big)'(1) > 0, \tag{4.9}$$

and denote $w = y_2 - y_1$. By (4.6), (4.7), (4.9), we see that w satisfies

$$\big(u^3 w''\big)'' = 0 \quad \text{in }]0,1[,$$

$$w(0) = w'(0) = 0,$$

$$w''(1) = 0, \quad \big(u^3 w''\big)'(1) = \gamma > 0.$$

Then, $u^3(x)\, w''(x) = \gamma x - \gamma \leq 0$ in $[0,1]$. That is, w is a concave function, and $w(0) = w'(0) = 0$ gives $w \leq 0$ in $[0,1]$. Therefore, $y_2(1) \leq y_1(1) < \alpha$, according to the assumption $y_1 \notin \mathcal{K}$. However, this is a contradiction to $y_2(1) = \alpha$, and it follows that (4.9) is false. Then (4.8) gives that y_2 is now the solution of (4.1); i.e., $y = y_2$ again has the claimed regularity. □

Remark. The boundary conditions in $x = 1$, associated with (4.5), are

$$y''(1) = 0, \quad y(1) \geq \alpha, \quad \big(u^3 y''\big)'(1) \leq 0,$$

$$\big(y(1) - \alpha\big)\big(u^3 y''\big)'(1) = 0.$$

We formulate the optimization problem $(\mathbf{P}_5)$:

$$\text{Min} \int_0^1 u(x)\, dx, \tag{4.10}$$

subject to (4.1) and to

$$m \leq u(x) \leq M \quad \text{a.e. in } [0,1], \tag{4.11}$$

$$y(x) \geq -\tau \quad \text{in } [0,1]. \tag{4.12}$$

Without loss of generality, we may assume

$$\alpha > -\tau. \tag{4.13}$$

Otherwise, all the admissible pairs of $(\mathbf{P}_5)$ correspond to an inactive variational inequality (the case $y = y_1$), that is, to a cantilevered beam (by Theorem 4.1), and we can refer to section 2. We call *extremal* for $(\mathbf{P}_5)$ any admissible "thickness" $u \in L^\infty(0,1)$ such that the associated state is active with respect to the constraint (4.12).

PROPOSITION 4.2. *If* $\alpha \geq 0$ *and* $m = 0$*, any local minimum of* $(\mathbf{P}_5)$ *is an extremal of* $(\mathbf{P}_5)$.

Proof. If $[u,y]$ is local optimum for $(\mathbf{P}_5)$, but not extremal, there is some $\lambda > 1$ such that the pair $[\lambda^{-1/3}\, u,\ \lambda\, y]$ is admissible for $(\mathbf{P}_5)$—it clearly satisfies the constraints and the variational inequality since $\lambda\, y \in \mathcal{K}$ by $\alpha \geq 0$.

Obviously $\lambda^{-\frac{1}{3}}\, u$ gives a lower cost which contradicts the local optimality of u when $\lambda \to 1+$. □

Remark. The case $\alpha = 0$ was considered by Hlavacek, Bock, and Lovisek [12].

PROPOSITION 4.3. *Assume that* $f < 0$ *in* $[0,1]$*. Then any extremal pair has exactly one active point in* $]0,1[$.

Proof. The existence of at least one point $x_u \in [0,1]$ such that $y(x_u) = -\tau$ is obvious by the definition. Assume that there are at least two such points $x_u \neq \bar{x}_u$, i.e., $y(x_u) = y(\bar{x}_u) = -\tau$. Again by definition, x_u and $\bar{x}_u$ are minimum points for y, different from 0 and 1, that is, $y'(x_u) = y'(\bar{x}_u) = 0$. Then $y + \tau$ satisfies the clamped beam conditions on $[x_u,\, \bar{x}_u]$. By [25, Thm. 3.1], we see that $y \leq -\tau$ on $[x_u,\, \bar{x}_u]$; therefore $y \equiv -\tau$ on $[x_u,\, \bar{x}_u]$. This contradicts $f < 0$ a.e. in $[0,1]$. □

Remark. Notice that by $y(0) = 0$ and $y(1) \geq \alpha > -\tau$ (by (4.13)) the end points cannot be active with respect to the state constraint.

COROLLARY 4.4. *If* $f \leq 0$*, any extremal of* $(\mathbf{P}_5)$ *satisfies* (4.7) *and* $(u^3 y'')'(1) \leq 0$.

Proof. If $f \leq u$, then the cantilevered beam y_1 has the global minimum in $x = 1$ and $y_1(1) \geq \alpha$ by (4.2). Condition (4.13) shows that y_1 cannot be extremal and Theorem 4.1 gives that y should satisfy (4.7) in order to be extremal for $(\mathbf{P}_5)$. Then, relation (4.9) is false, as in the proof of Theorem 4.1, and this completes the argument. □

Remark. By Corollary 4.4 and Proposition 4.2, the shape optimization problem $(\mathbf{P}_5)$, governed by variational inequalities, is reduced to the linear state system (4.7). Some cases of control problems governed by variational inequalities of obstacle type which can be equivalently reformulated as convex control problems with state constraints are discussed in Tiba [26, Chap. III.5], by different approaches.

We formulate the "dual" problem

$$(\mathbf{D}_5) \qquad \text{Min} \int_0^1 f(x)\, dx,$$

$$\begin{aligned}
(4.14) \qquad (\bar{u}^3\, y'')'' &= f \quad \text{in }]0,1[,\\
y(0) = y'(0) &= 0,\\
y(1) = \alpha, \quad y''(1) &= 0,\\
f &\leq 0 \quad \text{a.e. in } [0,1],\\
(4.15) \qquad y &\geq -\tau \quad \text{in } [0,1].
\end{aligned}$$

Notice that this is again a linear optimization problem ($\bar{u}$ is a prescribed thickness).

Remark. The control constraint $f \leq 0$ is a simplified stronger variant of (2.7), due to the maximum principle. Then, the equivalence results from Theorems 2.1 and 2.4 are not valid in this setting. However, we put into evidence that between the problems $(\mathbf{P}_5)$ and $(\mathbf{D}_5)$ there still exists a very useful relationship explained in Algorithm 4.6

and in Theorem 4.7. In the cases discussed in sections 2 and 3 (only for beams), this weaker relationship was studied in Sprekels and Tiba [25]. The problem $(\mathbf{D}_5)$ is, in fact, a slightly simplified variant of the problem $(\mathbf{D}_2)$ of section 2. Moreover, this approach allows us to consider $m = 0$ and gives another form for the resizing transformation.

THEOREM 4.5. *Assume that $\bar{u}$ is continuous and let $[y, f]$ be extremal for $(\mathbf{D}_5)$. Then y'' has exactly one root in $[0,1[$. Moreover, $y \leq \max\{0; \alpha\}$ in $[0,1]$.*

Proof. We have $\bar{u}^3\, y'' = g$ in $[0,1]$, where, moreover, $g'' = f$ in $[0,1]$ and $g(1) = 0$. Since $f \leq 0$, then g is concave in $[0,1]$ and it may have at most one root in $[0,1[$, unless it is identically 0 in some subinterval.

In the last subcase, by concavity and $g(1) = 0$, there is some $\xi \in]0,1[$ such that $g(x) \equiv 0$, $x \in [\xi, 1]$, and $g(x) < 0$ in $[0, \xi[$. Then $y'' < 0$ in $[0, \xi[$ and, since $y(0) = y'(0) = 0$, we see that $y(\xi) < 0$, $y'(\xi) < 0$, and $y(x) = y'(\xi)\,(x - \xi) + y(\xi)$ for $x \in [\xi, 1]$. We obtain that

$$\alpha \;=\; y(1) \;<\; y(x) \quad \forall\, x \,\in\, [0,1],$$

which contradicts the extremality of $[y, f]$ and (4.13).

Therefore g has at most one root in $[0,1[$. Since $[y, f]$ is extremal, there is some $\bar{\xi}$ in $]0,1[$ such that $y(\bar{\xi}) = -\tau$, and this is a minimum point for y on $]0,1[$. Then $y'(\bar{\xi}) = 0$, and there is some $\eta \in]0, \bar{\xi}[$ such that $y''(\eta) = 0$, since $y'(0) = 0$ and y'' is continuous by the assumption on $\bar{u}$.

We conclude that y'' has exactly one root in $]0, \bar{\xi}[$. Let η be this root. Then $y'' \leq 0$ in $[0, \eta]$ and $y'' \geq 0$ in $[\eta, 1]$. By $y(0) = y'(0) = 0$ and the concavity of y, we get $y \leq 0$ in $[0, \eta]$. By $y(\eta) \leq 0$, $y(1) = \alpha$, and the convexity of y, we get $y \leq \max\{0; \alpha\}$ in $[\eta, 1]$ as well. This ends the proof. □

Based on Theorem 4.5, we can formulate the following algorithm.

[4]ALGORITHM 4.6 ($m = 0$, $M = +\infty$).

```
1. $n = 0$,   $u_0$ admissible for $(\mathbf{P}_5)$, continuous.
2. Min$(\tilde{D}_n)$ gives $[y_n, f_n]$, where $(\tilde{D}_n)$ is given by (4.14) with $\bar{u}$ replaced by $u_n$.
3. If $f_n - f$ "small," then STOP! Otherwise
4. ("resizing step")
α) compute the unique root $\xi_n$ in $[0,1[$ of $y_n''$
β) denote $g_n = u_n^3\, y_n''$ and define $\tilde{g}_n$ by
      i)  $\begin{cases} \tilde{g}_n'' = f & \text{in } ]\xi_n, 1[, \\ \tilde{g}_n(\xi_n) = 0, & \tilde{g}_n(1) = 0, \end{cases}$
      ii) $\begin{cases} \tilde{g}_n'' = f & \text{in } ]0, \xi_n[, \\ \tilde{g}_n(\xi_n) = 0, & \tilde{g}_n'(\xi_n -) = \tilde{g}_n'(\xi_n +). \end{cases}$
γ) resize $u_n$ by $u_{n+1}^3 = u_n^3\, \frac{\tilde{g}_n}{g_n}$, and set $n := n + 1$, GO TO step 2.
```

Remark. The resizing rule (γ) is well defined even in ξ_n and in 1 by the Hopf maximum principle and l'Hospital's rule. The sequence $\{u_n\}$ remains continuous in all iterations and $1/u_n^3 \in L^2(0,1)$ if $1/u_0^3 \in L^2(0,1)$.

THEOREM 4.7. *Algorithm* 4.6 *generates extremals u_n for* $(\mathbf{P}_5)$ *in each step $n \geq 1$.*

Proof. If f_n is a minimum for $(\tilde{\mathbf{D}}_n)$, then it is extremal for $(\tilde{\mathbf{D}}_n)$. Otherwise, $y_n(x) \geq -\tau + \varepsilon$ in $[0,1]$ for some positive ε. Consider $f_\delta = f - \delta$, δ positive constant, and y_δ the corresponding solution of (4.11).

Clearly $y_\delta \to y_n$ uniformly in $[0,1]$ for $\delta \to 0$. Then, for some small δ, $[y_\delta,\, f_\delta]$ is an admissible pair for $(\tilde{\mathbf{D}}_n)$ with a lower cost. This is a contradiction to the optimality of f_n. Extremality is obviously preserved by the resizing rule, since (γ) and (β) give

$$\left(u_{n+1}^3\, y_n''\right)'' = (\tilde{g}_n)'' = f \quad \text{in }]0,1[;$$

i.e., y_n is the state associated with u_{n+1} in (4.7), or equivalently in (4.1). □

Remark. The algorithm has a global character since it iterates between extremals of $(\mathbf{P}_5)$. If the cost functional (4.10) is replaced by (2.28), then Algorithm 4.6 has the descent property as well (again by the resizing rule).

We close this section with a numerical example.

Example 4.8. We have made several experiments with Algorithm 4.6 applied to the minimum weight problem $(\mathbf{P}_5)$. The state equation was discretized by usual finite difference approximations for the derivatives, using the grid $x_i = i\,h$, $i = \overline{0,m}$, $h = 1/m$. By the discretization process, the problem $(\mathbf{D}_5)$ is approximated by a linear programming problem (LPP) (this is one of the advantages of an algorithm). The variables of the LPP are given by the discrete values of the pair $[y,f]$. The cost functional is evaluated using Simpson's approximation rule.

The numerical tests have been made with $m = 50$ which allows the LPP to be accurately solved via the simplex algorithm. The root ξ_n in step 4α) of Algorithm 4.6 was found using a cubic spline approximation of y_n. The differential equations corresponding to $\tilde{g}_n$ were solved by integrating first mathematically, using convolution formulae, and approximating next the definite integrals by a sharp numerical integration routine.

Generally, the algorithm stopped by failing to solve the problem $(\tilde{\mathbf{D}}_n)$ when it cannot further decrease the thickness u. The numerical tests have been made on a PC Pentium with floating point arithmetic accuracy of order 10^{-20}. We have fixed the load $\bar{f} \equiv -50$ in $]0,1[$ and $\bar{f} \equiv -1$ in $x = 0$, $x = 1$ or $\hat{f} \equiv -1$ in $[0, 0.5]$, and $\hat{f} \equiv -50$ in $]0.5, 1]$. The obstacle α had the values 0.1 or 0 or -0.1, the state constraint was $\tau = 0.6$ or $\tau = 0.5$, and the initial iteration (thickness) was $\bar{u}_0 = 2 + x\,(x-1)$ or $\hat{u}_0 = 2 - x$. In all these variants a sharp decrease in the thickness was obtained in a maximum of seven iterations, but usually only in three iterations. This information is collected in the following table (each column gives an experiment and $v(u_i)$ is the L^1 norm of u_i):

f	$\bar{f}$	$\bar{f}$	$\bar{f}$	$\hat{f}$	$\hat{f}$
α	0.0	- 0.1	0.1	- 0.1	0.1
τ	0.6	0.6	0.6	0.6	0.5
u_0	$\hat{u}_0$	$\hat{u}_0$	$\bar{u}_0$	$\bar{u}_0$	$\hat{u}_0$
$v\,(u_1)$	1.1369	1.2987	1.5064	1.6727	1.2513
$v\,(u_2)$	0.9194	1.1146	1.2586	1.5916	1.1598
$v\,(u_3)$	0.7583	0.9823	1.2682	1.5693	1.1100
$v\,(u_4)$			0.9280	1.4288	1.0740
$v\,(u_5)$			0.7596	1.3296	1.0468
$v\,(u_6)$			0.6540	1.3760	
$v\,(u_7)$			0.5781	1.2778	

Acknowledgments. The numerical tests were performed with the help of Dr. V. Arnautu and Dr. A. Ignat of the University of Iaşi, Romania.

REFERENCES

[1] V. Arnautu, H. Langmach, J. Sprekels, and D. Tiba, *On the Approximation and the Optimization of Plates*, Preprint 357, Weierstrass Institute for Applied Analysis and Stochastics, Berlin, 1997.

[2] M. P. Bendsoe, *Existence proofs for a class of plate optimization problems*, in System Modelling and Optimization (Copenhagen, 1983), Lecture Notes in Inform. Sci. 59, Springer-Verlag, Heidelberg, 1984, pp. 773–779.

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

[4] E. Casas, *Optimality conditions and numerical approximations for some optimal design problems*, Control Cybernet., 19 (1990), pp. 73–91.

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

[6] S. J. Cox and M. L. Overton, *On the optimal design of columns against buckling*, SIAM J. Math. Anal., 23 (1992), pp. 287–325.

[7] R. J. Duffin, *On a question of Hadamard concerning super-biharmonic functions*, J. Math. Phys., 27 (1949), pp. 253–256.

[8] P. R. Garabedian, *A partial differential equation arising in conformal mapping*, Pacific J. Math., 1 (1951), pp. 485–524.

[9] J. Hadamard, *Sur certains cas intéressants au problème biharmonique*, in Proc. of the Fourth International Mathematical Congress at Rome, 2 (1908), Oeuvres. Vol. 1–4, Editions du Centre National de la Recherche Scientifique, Paris, 1968, pp. 12–14.

[10] R. T. Haftka, Z. Gürdal, and P. M. Kamat, *Elements of Structural Optimization*, Kluwer Academic Press, Boston, 1990.

[11] J. Haslinger and P. Neittaanmäki, *Finite Element Approximation of Optimal Shape Design*, John Wiley, Chichester, 1988.

[12] I. Hlavacek, I. Bock, and J. Lovisek, *Optimal control of a variational inequality with applications to structural analysis.* I. *Optimal design of a beam with unilateral supports*, Appl. Math. Optim., 11 (1984), pp. 111–143.

[13] I. Hlavacek, I. Bock, and J. Lovisek, *Optimal control of a variational inequality with applications to structural analysis.* II. *Local optimization of the stresses in a beam.* III. *Optimal design of an elastic plate*, Appl. Math. Optim., 13 (1985), pp. 117–135.

[14] L. Hörmander, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, 1964.

[15] B. Kawohl, *Personal communication*, 1996.

[16] B. Kawohl and J. Lang, *Are some optimal shape problems convex?*, J. Convex Anal., 4 (1997), pp. 353–361.

[17] A. Langenbach, *Monotone Partialoperatoren in Theorie und Anwendung*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1976.

[18] J. L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.

[19] J. L. Lions and E. Magenes, *Problèmes aux limites non homogénes et applications*, Dunod, Paris, 1968.

[20] F. Murat, *Contre-exemples pour divers problèmes où le contrôle intervient dans les coefficients*, Ann. di Mat., 112 (1977), pp. 49–68.

[21] C. K. Neto and E. Polak, *On the use of consistent approximations for the optimal design of beams*, SIAM J. Control Optim., 34 (1996), pp. 1891–1913.

[22] G. I. Rozvany, N. Olhoff, K. T. Cheng, and J. Taylor, *On the solid plate paradox in structural optimization*, J. Struct. Mechanics, 10 (1982), pp. 1–32.

[23] P. Senatorov, *The stability of the eigenvalues and eigenfunctions of a Sturm–Liouville problem*, Differential Equations, 7 (1971), pp. 1266–1269.

[24] H. S. Shapiro and M. Tegmark, *An elementary proof that the biharmonic Green function of an eccentric ellipse changes sign*, SIAM Rev., 36 (1994), pp. 99–101.

[25] J. Sprekels and D. Tiba, *A duality-type method for the design of beams*, Math. Sci. Appl., to appear.

[26] D. Tiba, *Optimal Control of Nonsmooth Distributed Parameter Systems*, Lecture Notes in Math. 1459, Springer-Verlag, Berlin, 1990.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 2, pp. 502–520

APPROXIMATE SOLUTION OF MARKOV RENEWAL PROGRAMS WITH FINITE TIME HORIZON*

KARL HINDERER† AND KARL-HEINZ WALDMANN‡

Abstract. The present paper investigates the error committed by using an infinite time horizon Markov renewal program as an approximation of the (often more realistic) Markov renewal program with a finite time horizon t_0. Under weak assumptions the error is shown to converge to zero exponentially fast when $t_0 \to \infty$. The convergence is based on explicit error bounds. Improved error bounds hold when the (transformed) transition law has a nontrivial stochastic lower bound. Some bounds use the discounted renewal function. For the latter, monotone upper and lower bounds are obtained by an iterative method combined with an extrapolation. Several examples demonstrate the applicability of the results.

Key words. Markov renewal programs, finite time horizon, approximate solution, discounted renewal function

AMS subject classifications. 90C40, 90C42

PII. S0363012997317207

1. Introduction.

A Markov renewal program (also known as a semi-Markov decision process) with infinite time horizon and infinitely many stages, denoted by MRP$^\infty$, is an important and intensively studied extension of the classical infinite-stage Markovian decision process. The extension concerns the time (points) $0 \equiv T_0 \leq T_1 \leq T_2 \leq \cdots$ where decisions (also called actions) are made. They are no longer fixed and equidistant, but random. In particular, at each time T_ν an action $a_\nu = \pi_\nu(s)$ is taken, depending on the present state $\zeta_\nu = s$ by means of a decision rule π_ν. Associated with a sequence $\pi := (\pi_\nu)_0^\infty$ of decision rules, known as a policy, is a sequence of rewards earned during the decision epochs $[T_\nu, T_{\nu+1})$. These rewards are discounted to time zero with a discount rate $\alpha > 0$. The goal consists of maximizing the expected total discounted reward within the set of all policies. As usual, it is assumed that $(T_{\nu+1} - T_\nu, \zeta_\nu), \nu \geq 1$, form a Markov chain for each policy.

Many applications in Markov renewal programming require a finite time horizon t_0 for a realistic modelling; i.e., the decision process runs for at most t_0 time units. This concerns, e.g., the many models of control of queues in cases where the servers only work during an interval of finite length or during a sequence of disjoint finite time intervals (e.g., during daytime) with identical initial states (e.g., an empty queue).

Despite its importance there is only a small amount of literature on Markov renewal programs with a finite time horizon t_0 (MRP^{t_0} for short). The first paper in this area is Jewell [6]. A thorough investigation of the foundations was made by Schellhaas [11]. Bounds for the value function are derived in Waldmann [13]. An application to asset selling and a survey on papers with applications (including Gertsbakh [4]) is given in Mamer [10].

The scarcity of the literature is probably due to the following reasons:

*Received by the editors February 21, 1997; accepted for publication (in revised form) February 26, 1998; published electronically December 2, 1998.
http://www.siam.org/journals/sicon/37-2/31720.html

†Institut für Mathematische Stochastik, Universität Karlsruhe, Englerstr. 2, D-76128 Karlsruhe, Germany (karl.hinderer@math.uni-karlsruhe.de).

‡Institut für Wirtschaftstheorie und Operations Research, Universität Karlsruhe, Kaiserstr. 12, Geb. 20.21, D-76128 Karlsruhe, Germany (waldmann@wior.uni-karlsruhe.de).

(i) As already observed by Jewell [6], the solution of a Markov renewal program with a finite time horizon can be formally reduced to one with an infinite time horizon with augmented states (s, t), consisting of the "internal" state s and the remaining time horizon t. Thus the action taken at time T_ν usually depends not only on the present internal state s but also on the remaining time horizon t. This enlarges the set of relevant policies considerably, and, as a consequence, much more effort is required for both a qualitative analysis and a numerical computation of MRP^{t_0} compared with MRP^∞. Therefore it is customary (usually without mentioning it) to use the solution of MRP^∞ as approximation for the solution of MRP^{t_0}.

(ii) It seems to be a general belief that the error made in solving MRP^{t_0} approximately by the corresponding infinite time horizon MRP^∞ is "small," at least if t_0 is "large." The goal of the present paper is to confirm this belief by deriving explicit bounds for the errors made.

In particular, we are interested in the following questions:

(a) Denote by $V^{t_0}(s)$ (resp., $V(s)$) the maximal expected total reward in MRP^{t_0} (resp., in MRP^∞) for a given initial state s. Then, using $V(s)$ as approximation for $V^{t_0}(s)$, we want to know "good" upper and lower bounds for the *error* $V^{t_0}(s) - V(s)$ and conditions under which $V^{t_0}(s)$ converges "quickly" to $V(s)$ for $t_0 \to \infty$.

(b) Let π^* be an optimal policy for MRP^∞, and let $V^{t_0}_{\pi^*}(s)$ be the associated expected total reward in MRP^{t_0}. Then we want to know lower bounds for the performance of π^* when applied in MRP^{t_0} as a substitute for an optimal horizon-dependent policy, i.e., lower bounds for the *error* $V^{t_0}_{\pi^*}(s) - V^{t_0}(s)$.

To the best of our knowledge, no results about problems (a) and (b) are known in the literature. (In Jewell [6, p. 745], a proof for the convergence of $V^{t_0}(s)$ to $V(s)$ is indicated for a special model, but it seems that only the convergence of $V^{t_0}_\pi(s)$ toward $V_\pi(s)$ for a fixed stationary policy π is shown; cf. also De Cani [2, p. 730].)

The main purpose of the present paper is to answer the questions in (a) and (b) by deriving under weak assumptions and for general transition law and general reward structure upper and lower bounds for $V^{t_0}(s) - V(s)$ and lower bounds for $V^{t_0}_{\pi^*}(s) - V^{t_0}(s)$ of the form

$$e^{-\delta t_0}\, b(s)\, K(t_0). \tag{1.1}$$

Here $s \mapsto b(s)$ is a bounding function of MRP^∞, which, as usual, ensures the existence and finiteness of $V^{t_0}(s)$ and of $V(s)$. Moreover, K is an explicitly given function, and δ can be chosen in some subinterval of $[0, \alpha]$, determined by integrability conditions; cf. assumptions (A1) and (A2) below. In general K depends on δ and is bounded in t for fixed δ. Then δ is crucial for the quality of the bounds (1.1). Moreover, we obtain exponential convergence of $V^{t_0}(s)$ to $V(s)$ for $t_0 \to \infty$ if upper and lower bounds for $V^{t_0}(s) - V(s)$ of the form (1.1) hold for some $\delta > 0$.

Some of our functions K (e.g., (3.19)) contain the discounted renewal function $H_\infty := \sum_{i=0}^\infty \gamma^i H^{i*}$, determined by a constant γ, defined in (3.7), and a distribution function H on $\mathbb{R}_+$, which is a lower stochastic bound for the "transformed" transition law G, defined in (3.8). As H_∞ is explicitly known in a few cases only, we obtain easily computable upper bounds for it (and hence for K) in the following way: One selects a distribution function $\tilde{H} \geq H$, whose probability distribution is concentrated on $\{0, \varepsilon, 2\,\varepsilon, \ldots, m\,\varepsilon\}$ for some $m \in \mathbb{N}$ and some $\varepsilon > 0$. Then the upper bound $\tilde{H}_\infty(k\,\varepsilon)$ of $H_\infty(k\,\varepsilon)$, $k \in \mathbb{N}$, can be found by a recursion, combined with an extrapolation. The latter is of general interest for approximate solutions of linear difference equations with constant coefficients without solving the characteristic equation.

For technical reasons we mainly study models MRP$^{\infty}$ and MRP^{t_0}, which have a finite number n of decision epochs only, i.e., planning horizon T_n and $\min(t_0, T_n)$, resp. These problems are of independent interest and they allow to derive easily the corresponding results for the case $n = \infty$ by letting n tend to infinity (cf. Theorem 3.7, Proposition 4.2(c) and Theorem 4.3(b)).

Another paper where both stages and a continuous-time horizon play a role is van Dijk [12].

This paper is organized as follows. In section 2 we rigorously introduce the model MRP$^{\infty}$, *for simplicity denoted by MRP from now on,* and the model MRP$'$, which comprises all models MRP^{t_0}, $t_0 > 0$. The flexibility of MRP^{t_0} is increased by including an additional reward $\tilde{h}$ which is obtained when the process passes the horizon t_0. Special cases of the reward structure which are important for many applications are considered in (2.1) and (2.3). In Theorems 3.5 through 3.7 we present upper bounds, which always hold for $H :\equiv 1$, and which yield improved bounds if a nontrivial stochastic lower bound H for the transformed transition law exists. In section 4 lower and absolute bounds are derived, and the performance of optimal policies for MRP, when applied to MRP^{t_0}, is studied. Section 5 is devoted to an efficient (approximate) computation of the discounted renewal function. Section 6 contains four examples.

Notation. $\mathbb{N}_0 := \{0, 1, \ldots\}$; $\mathbb{N} := \{1, 2, \ldots\}$; $\mathbb{R}_+ := [0, \infty)$; $x^{\pm} := \max\{0, \pm x\}$, $x \in \mathbb{R}$; δ_x is the one-point measure in the point $x \in \mathbb{R}$; $0/0 := 0$; $\sigma_n(x) := \sum_{i=0}^{n-1} x^i$ for $n \in \mathbb{N}_0$ and $x \in \mathbb{R}_+$; "measurable" means "measurable with respect to the σ-algebras under consideration." On countable sets we use the power set as σ-algebra. Probability distribution is abbreviated by p.d.

2. The models MRP and MRP$'$. Our infinite time-horizon Markov renewal program MRP with a fixed number $n \in \mathbb{N}$ of decision epochs (stages) consists of a tuple $(S, A, D, \kappa, \tilde{r}, \alpha, V_0)$ of the following meaning. S and A are the (arbitrary) state space and action space, endowed with some σ-algebra $\mathcal{S}$ and $\mathcal{A}$, resp., $D(s) \subset A$ is the set of admissible actions in state s; the constraint set $D := \{(s, a) \in S \times A \mid a \in D(s)\}$ is assumed to be measurable and to contain the graph of a measurable mapping from S into the action space A. The transition p.d. $\kappa(s, a, d(z, u))$ from D into $\mathbb{R}_+ \times S$ defines the joint p.d. of the sojourn time z in s (i.e., the length of the decision epoch) and of the next state u. (It is common to ensure by appropriate assumptions that $T_\nu \to \infty$ almost surely (a.s.) for $\nu \to \infty$. However, our optimization problem makes sense and our results hold without such assumptions.) The one-stage reward $\tilde{r}(s, a, z, u)$ is earned at the beginning of the decision epoch. Further we allow a terminal reward $V_0(s)$, which is realized at the end of the last decision epoch. We suppose $\tilde{r} : D \times \mathbb{R}_+ \times S \to \mathbb{R}$ and $V_0 : S \to \mathbb{R}$ to be measurable. Finally, all rewards are discounted to time zero with rate $\alpha > 0$.

In applications, $\tilde{r}$ often has the form

$$\tilde{r}(s, a, z, u) = \tilde{r}_1(s, a) + \int_0^z e^{-\alpha y}\, \tilde{r}_2(s, a, y)\, dy + e^{-\alpha z}\, \tilde{r}_3(s, a, z, u). \tag{2.1}$$

Here $\tilde{r}_1(s, a)$ and $\tilde{r}_3(s, a, z, u)$ are earned at the beginning and the end, resp., of the decision epoch, while $\tilde{r}_2(s, a, y)$ is a reward rate for continuously earned rewards.

For establishing existence and finiteness of the value functions, we must strengthen the well-known concept of a bounding function. We impose the condition that there exists some parameter $\delta \in [0, \alpha]$, which is kept fixed, for which the subsequent assumptions (A1) and (A2) hold.

(A1) There exists a measurable function b from S into $(0,\infty)$ and a constant $\varrho \geq 0$ such that for all $(s,a) \in D$

(i) $\int \kappa(s,a,d(z,u))e^{\delta z}\,|\tilde{r}(s,a,z,u)| \leq \varrho\, b(s)$,

(ii) $|V_0(s)| \leq \varrho\, b(s)$,

(iii) $0 < \int \kappa(s,a,d(z,u))e^{-(\alpha-\delta)z}\, b(u) \leq \varrho\, b(s)$.

If (A1) holds for some $\delta \in [0,\alpha]$, then it also holds for $\delta = 0$, which means that b is a bounding function of MRP. Let B be the set of measurable functions v on S, which have finite weighted supremum norm

$$\|v\|_b := \sup_{s\in S} [|v(s)|/b(s)].$$

Because of (A1) there exists

$$r(s,a) := \int \kappa(s,a,d(z,u))\,\tilde{r}(s,a,z,u).$$

As usual, a measurable mapping $f : S \to A$ with $f(s) \in D(s), s \in S$, is called a decision rule. A finite sequence $\pi = (f_n, f_{n-1}, \dots, f_1)$ of decision rules, specifying action $f_k(s)$ to be taken in state s at the beginning of the $(n-k+1)$th decision epoch (i.e., having k stages left), is called an n-stage (Markovian) policy. Let Δ_n denote the set of all n-stage policies. For each policy $\pi = (f_n, f_{n-1}, \dots, f_1)$ and initial state s there exists, as b is a bounding function, the expected n-stage reward

$$V_{n\pi}(s) := E_{\pi s}\left[\sum_{\nu=0}^{n-1} e^{-\alpha T_\nu}\cdot \tilde{r}(\zeta_\nu, f_{n-\nu}(\zeta_\nu), Z_{\nu+1}, \zeta_{\nu+1}) + e^{-\alpha T_n}\cdot V_0(\zeta_n)\right].$$

Here $\zeta_0 := s$, ζ_ν for $1 \leq \nu \leq n$ denotes the random state at time T_ν, and $Z_{\nu+1} := T_{\nu+1} - T_\nu$ is the sojourn time in state ζ_ν. $E_{\pi s}$ denotes the expectation determined by the initial state s, the policy π, and the transition law κ in the usual way.

A policy $\pi^* \in \Delta_n$ is called optimal if $V_{n,\pi^*}(s) = V_n(s) := \sup_{\pi\in\Delta_n} V_{n,\pi}(s)$ holds for all $s \in S$. We call V_n, which belongs to B as b is a bounding function, the n-stage value function.

Next we introduce the finite time-horizon version MRP^{t_0} of MRP, which consists of the MRP $(S, A, D, \kappa, \tilde{r}, V_0, \alpha)$, a finite time-horizon (called also initial residual time) $t_0 > 0$, and an additional measurable function $\tilde{h}$ on $\mathbb{R}_+ \times D \times \mathbb{R}_+ \times S$.

To explain the meaning of $\tilde{h}$, let us have $k > 0$ stages left and a residual time $t \geq 0$. Interpret s, a, z, and u as above. Then, if $t - z < 0$, the process stops with a terminal payoff $e^{\alpha t}\,\tilde{h}(t,s,a,z,u)$, which is realized at time t_0. Otherwise, i.e., if $t - z \geq 0$, the process earns immediately $\tilde{r}(s,a,z,u)$ and continues with $k-1$ stages left and residual time $t - z$. Finally, in case of $k = 0$ stages left and a residual time $t \geq 0$, there is a terminal reward $V_0(s)$ again.

All finite time-horizon problems $\mathrm{MRP}^{t_0}, t_0 > 0$, can be modelled by an infinite time-horizon model MRP′, say, with extended state space $X = \mathbb{R} \times S$, where the extended state $x = (t,s)$ consists of two parts, the momentary residual time t, and the momentary internal state s. (The states with $t < 0$ are only needed for a convenient modelling.) The other data are as follows:

- A is the action space.
- $D(s)$ is the set of admissible actions in state $x = (t,s)$.
- A transition occurs from $x = (t,s)$ to $\hat{x} = (t-z,u)$ according to $\kappa(s,a,d(z,u))$.

- The one-stage reward is

$$\tilde{r}'(t,s,a,z,u) := 1_{\mathbb{R}_+}(t)\left[\tilde{r}(s,a,z,u)1_{[0,t]}(z) + \tilde{h}(t,s,a,z,u)1_{(t,\infty)}(z)\right]. \tag{2.2}$$

- $V_0(s)1_{\mathbb{R}_+}(t)$ is the terminal reward.
- α is the discount rate.

If $\tilde{r}$ has the special form (2.1), we use as special form of $\tilde{r}'$ for $t \geq 0$

$$\begin{aligned} \tilde{r}'(t,s,a,z,u) &:= \tilde{r}_1(s,a) + \int_0^{z\wedge t} e^{-\alpha y}\,\tilde{r}_2(s,a,y)\,dy \\ &+ e^{-\alpha z}\,\tilde{r}_3(s,a,z,u)\,1_{[0,t]}(z) + e^{-\alpha t}\,R(t,s,a,z,u)1_{(t,\infty)}(z). \end{aligned} \tag{2.3}$$

Here $R(t,s,a,z,u)$ is an appropriately chosen reward, earned at time t_0. This case is covered by (2.2) by putting for $t \geq 0$

$$\tilde{h}(t,s,a,z,u) := \tilde{r}_1(s,a) + \int_0^t e^{-\alpha y}\,\tilde{r}_2(s,a,y)\,dy + e^{-\alpha t}R(t,s,a,z,u). \tag{2.4}$$

We additionally suppose the following:

(A2) There exists a measurable function $h : D \times \mathbb{R}_+ \times S \to \mathbb{R}_+$ such that there holds for all t,s,a,z,u

$$|\tilde{h}(t,s,a,z,u)| \leq h(s,a,z,u)$$

and

$$\int \kappa\,(s,a,d(z,u))\,e^{\delta z}\,h(s,a,z,u) \leq \varrho\,b(s).$$

(A1) and (A2) hold with $b \equiv 1$ if either V_0, $e^{\delta z}\tilde{r}(s,a,z,u)$ and $e^{\delta z}h(s,a,z,u)$ are bounded or if $V_0, \tilde{r}$, h and $\int \kappa(s,a,d(z,u))e^{\delta z}$ are bounded.

Define r^t by $r^t(s,a) := 0$ for $t < 0$ and for $t \geq 0$ by

$$\begin{aligned} r^t(s,a) :&= \int \kappa(s,a,d(z,u))\tilde{r}'(t,s,a,z,u) \\ &= r(s,a) + \int \kappa(s,a,d(z,u))1_{(t,\infty)}(z)\left[\tilde{h}(t,s,a,z,u) - \tilde{r}(s,a,z,u)\right]. \end{aligned} \tag{2.5}$$

In MRP′ the set Δ_n' of policies $\pi' = (f_n', f_{n-1}', \dots, f_1')$ is defined as in MRP, except that the decision rules f_k' depend on both the internal state s and the residual time t. For $\pi' \in \Delta_n'$ and initial state s there exists, as $(t,s) \mapsto b(s)\,1_{\mathbb{R}_+}(t)$ is a bounding function of MRP′ by (A1) and (A2), the expected n-stage reward

$$\begin{aligned} V_{n\pi'}^t(s) := E_{\pi'(t,s)}\Bigg[\sum_{\nu=0}^{n-1} e^{-\alpha T_\nu} \cdot \tilde{r}'(\tau_\nu, \zeta_\nu, f_{n-\nu}'(\tau_\nu,\zeta_\nu), Z_{\nu+1}, \zeta_{\nu+1}) \\ + e^{-\alpha T_n} \cdot V_0(\zeta_n) \cdot 1_{\mathbb{R}_+}(\tau_n)\Bigg]. \end{aligned}$$

Here τ_ν denotes the random residual time at time T_ν, and $E_{\pi'(t,s)}$ is the expectation determined by the residual time t, the initial state s, the policy π', and the transition law in MRP′ in the usual way.

A policy $\pi'^* \in \Delta'_n$ is called optimal if $V^t_{n,\pi'^*}(s) = V^t_n(s) := \sup_{\pi' \in \Delta'_n} V^t_{n,\pi'}(s)$ holds for all $(t,s) \in X$. The function $(t,s) \mapsto V^t_n(s)$, which is the n-stage value function in MRP$'$, belongs to the set B' of measurable functions v' on X with $v'(t,s) = 0$ for $t < 0$ and which have finite weighted supremum norm

$$\|v'\|_b := \sup_{(t,s)\in X} [|v'(t,s)|/b(s)].$$

Our bounds for $V^t_n - V_n$ are based on the value iteration in MRP and in MRP$'$, i.e., on the recursions

$$V_n(s) = \sup_{a\in D(s)} \left[r(s,a) + \int \kappa\,(s,a,d(z,u))\; e^{-\alpha z}\, V_{n-1}(u) \right] \tag{2.6}$$

and for $t \in \mathbb{R}$

$$V^t_n(s) = \sup_{a\in D(s)} \left[r^t(s,a) + \int \kappa\,(s,a,d(z,u))\; e^{-\alpha z}\, V^{t-z}_{n-1}(u) \right]. \tag{2.7}$$

As there exist a variety of approaches to ensure the validity of the value iterations, some of which are quite elementary, we do not make any topological assumptions on S and/or A (such as being Borel spaces), but instead make the following assumption:

(A3) The functions $s \to V_n(s)$ and $(t,s) \to V^t_n(s)$ are measurable and satisfy (2.6) and (2.7), resp.

Clearly, if S and A are finite, (A3) holds trivially. Note that (2.7) implies that $V^t_n(s) = 0$ for $t < 0$.

3. Upper bounds for $V^t_n - V_n$. We first make some preparations. For $v \in B$ and $(s,a) \in D$ put

$$\psi v(s,a) := \int \kappa\,(s,a,d(z,u))\, e^{-\alpha z} v(u),$$
$$\psi_t\, v(s,a) := \int \kappa(s,a,d(z,u)) 1_{(t,\infty)}(z) e^{-\alpha z} v(u),\ t \geq 0.$$

Application to b yields the constants

$$\underline{\beta} := \inf_{(s,a)\in D} \psi b(s,a)/b(s),$$
$$\beta := \sup_{(s,a)\in D} \psi b(s,a)/b(s),$$

and a transition measure μ from D into $\mathbb{R}_+$, defined by

$$\mu(s,a,(t,\infty)) := \psi_t\, b(s,a)/b(s), \quad t \geq 0.$$

Finally, let

$$d := \sup_{s\in S} (V_0(s) - V_1(s))/b(s),$$

and for $n \in \mathbb{N}_0$,

$$e_n := d\,\sigma_n(\beta) \text{ if } d \geq 0, \text{ and } e_n := d\,\sigma_n(\underline{\beta}) \text{ otherwise.}$$

Note that d is finite as V_0, $V_1 \in B$.

LEMMA 3.1. *Assume* (A1)–(A3). *Then*

$$V_0(s) - V_n(s) \le e_n b(s) \text{ for all } s \in S, n \in \mathbb{N}.$$

Proof. The assertion follows by an inductive argument based on

$$\begin{aligned} V_0(s) - V_n(s) &= \Big(V_0(s) - V_1(s)\Big) + \Big(V_1(s) - V_n(s)\Big) \\ &\le d\, b(s) + \sup_a \psi\,(V_0 - V_{n-1})\,(s,a),\ n \ge 1. \qquad \square \end{aligned}$$

For

$$\Delta^t(s,a) := e^{\delta t}\Big[r^t(s,a) - r(s,a) - \psi_t V_0(s,a)\Big], \quad t \ge 0, \tag{3.1}$$

we obtain the next result.

LEMMA 3.2. *Assume* (A1) *and* (A2). *Then*

$$0 \le w := \sup\{\Delta^t(s,a)/b(s) : (s,a) \in D,\ t \ge 0\} < \infty.$$

Proof. From (2.5) we see that $\Delta^t(s,a) = \int \kappa(s,a,d(z,u))\, g_t(s,a,z,u)$, where

$$g_t(s,a,z,u) := e^{\delta t}\, 1_{(t,\infty)}(z)\Big[\tilde{h}(t,s,a,z,u) - \tilde{r}(s,a,z,u) - e^{-\alpha z} \cdot V_0(u)\Big].$$

Because of $g_t(s,a,z,u) = 0$ for $t \ge z$, $g_t(s,a,z,u)$ trivially converges for $t \to \infty$ to zero. Moreover,

$$\begin{aligned} |g_t(s,a,z,u)| &\le \Lambda(s,a,z,u) \\ := e^{\delta z}\Big(h(s,a,z,u) + |\tilde{r}(s,a,z,u)|\Big) &+ e^{-(\alpha-\delta)z} \cdot |V_0(u)|. \end{aligned}$$

Utilizing (A1) and (A2) we obtain

$$\begin{aligned} \int \kappa(s,a,d(z,u))\,\Lambda(s,a,z,u) &\le \int \kappa(s,a,d(z,u))\Big[e^{\delta z}\, h(s,a,z,u) \\ &\qquad + e^{\delta z} \cdot |\tilde{r}(s,a,z,u)| + e^{-(\alpha-\delta)z} \cdot |V_0(u)|\Big] \\ &\le \varrho\,(2+\varrho)\, b(s). \end{aligned}$$

By the dominated convergence theorem $\Delta^t(s,a)$ converges for $t \to \infty$ to zero, from which the assertion follows. $\square$

We define operators L on B and L' on B', resp., by

$$\begin{aligned} Lv(s,a) &:= r(s,a) + \int \kappa(s,a,d(z,u)) e^{-\alpha z} v(u),\ (s,a) \in D, \\ L'v'(t,s,a) &:= r^t(s,a) + \int \kappa(s,a,d(z,u)) e^{-\alpha z} v'(t-z,u),\ (t,s,a) \in \mathbb{R} \times D. \end{aligned}$$

The next two results are our main tools for deriving bounds.

LEMMA 3.3. *Assume* (A1) *and* (A2). *Let* $v \in B, v' \in B'$, *and*

$$v'(t,s) - v(s) \le e^{-\delta t}\, b(s)\, K(t),\ t \ge 0,\ s \in S, \tag{3.2}$$

for some lower bounded measurable function K *on* $\mathbb{R}_+$. *Then*

$$L'v'(t,s,a) - Lv(s,a) \le e^{-\delta t} b(s) \hat{K}_v(t),\ t \ge 0,\ (s,a) \in D, \tag{3.3}$$

where

$$\hat{K}_v(t) := w + \sup_{(s,a)\in D}\left[\hat{d}_v\, e^{\delta t}\mu(s,a,(t,\infty)) + \int_{[0,t]} \mu(s,a,dz)e^{\delta z}K(t-z)\right]$$

and $\hat{d}_v := \sup_{s\in S}(V_0(s) - v(s))/b(s)$ *(which is finite as* $v, V_0 \in B$*).*

Proof. Fix $t \geq 0$, $(s,a) \in D$. Then

$$\begin{aligned} &L'v'(t,s,a) - Lv(s,a)\\ &= r^t(s,a) + \int \kappa(s,a,d(z,u))1_{[0,t]}(z)\, e^{-\alpha z}\, v'(t-z,u) - r(s,a) - \psi v(s,a)\\ &= e^{-\delta t}\Bigg[\ \Delta^t(s,a) + e^{\delta t}\psi_t(V_0 - v)(s,a)\\ &\qquad + \int \kappa(s,a,d(z,u))1_{[0,t]}(z)e^{\delta t-\alpha z}(v'(t-z,u) - v(u))\Bigg]. \end{aligned} \tag{3.4}$$

As $V_0(s) - v(s) \leq \hat{d}_v\, b(s)$, we have

$$\psi_t(V_0 - v)(s,a)/b(s) \leq \hat{d}_v\, \psi_t b(s,a)/b(s) = \hat{d}_v\, \mu(s,a,(t,\infty)).$$

Now the assertion follows, using equation (3.2), by taking the supremum over (s,a) in equation (3.4). □

LEMMA 3.4. *Assume* (A1)–(A3). *Then there holds for all* $n \in \mathbb{N}$ *and* $t \geq 0$

$$V_n^t(s) - V_n(s) \leq e^{-\delta t} b(s) K_n(t), \quad s \in S, \tag{3.5}$$

for each sequence $(K_n)_0^\infty$ *of lower bounded measurable functions on* $\mathbb{R}_+$*, fulfilling* $K_0 \equiv 0$ *and for* $n \in \mathbb{N}_0$, $t \geq 0$ *and* $(s,a) \in D$ *the condition*

$$K_{n+1}(t) \geq w + e_n\, e^{\delta t}\mu(s,a,(t,\infty)) + \int_{[0,t]} \mu(s,a,dz)e^{\delta z}K_n(t-z). \tag{3.6}$$

Proof. For $n = 1$, (3.5) follows from Lemma 3.3 with $v'(t,s) := V_0(s)\, 1_{\mathbb{R}_+}(t)$ and $v(s) := V_0(s)$. Thus assume (3.5) to hold for some $n \geq 0$. From (2.6) and (2.7) we know that

$$\begin{aligned} V'_{n+1}(t,s) - V_{n+1}(s) &= \sup_{a\in D(s)} L'V'_n(t,s,a) - \sup_{a\in D(s)} LV_n(s,a)\\ &\leq \sup_{a\in D(s)} (L'V'_n(t,s,a) - LV_n(s,a)), \end{aligned}$$

where $V'_n(t,s) = V_n^t(s)$. Now (3.5) follows for $n+1$ from Lemma 3.3 with $v' := V'_n$ and $v := V_n$, as $\hat{d}_v \leq e_n$ by Lemma 3.1. □

Now we construct explicitly given sequences $(K_n)_1^\infty$ of bounded functions satisfying (3.5). They lead to our main results, Theorems 3.5–3.7 and the supplementary Remark 3.3. Define

$$\gamma := \sup_{(s,a)\in D}\left\{b(s)^{-1} \cdot \int \kappa(s,a,d(z,u))e^{-(\alpha-\delta)z}b(u)\right\}, \tag{3.7}$$

and for fixed $(s,a) \in D$ a distribution function $G(s,a,\cdot)$ with $G(s,a,t) := 0$ for $t < 0$ by

$$\begin{aligned}(3.8)\quad & G(s,a,t) \\ & := \int \kappa(s,a,d(z,u)) 1_{[0,t]}(z) e^{-(\alpha-\delta)z} b(u) \Big/ \int \kappa(s,a,d(z,u)) e^{-(\alpha-\delta)z} b(u), \quad t \geq 0.\end{aligned}$$

We call $G(s,a,\cdot)$ the transformed transition law. Because of (A1) we have $0 < \beta \leq \gamma \leq \rho$, and $\gamma = \beta$ if and only if $\delta = 0$.

We additionally suppose the following:

(A4) There is some distribution function H with $H(t) = 0$ for $t < 0$ such that

$$H(t) \geq G(s,a,t), \quad (s,a) \in D, \quad t \geq 0.$$

Assumption (A4) is equivalent to saying that H is stochastically smaller than $G(s,a,\cdot)$ for all $(s,a) \in D$, or equivalently, that H is a stochastic lower bound of the transformed transition law. Clearly (A4) is always fulfilled for $H \equiv 1$. Therefore the subsequent bounds hold already under (A1)–(A3) when H^{i*}, the i-fold convolution of H with itself, is replaced everywhere by 1. If both S and A are finite, (A4) holds with $H(t) := \max_{(s,a)} G(s,a,t)$. In applications, often κ has the form $\kappa(s,a,d(z,u)) = Q(dz) \times P(s,a,z,du)$. Then, if either $P(s,a,z,du)$ does not depend on z or if (A1) and (A2) hold for $b \equiv 1$, $G(s,a,\cdot) = G(\cdot)$ is independent of $(s,a) \in D$, and then (A4) holds with $H := G$. Some of our bounds use

$$H_n := \sum_{i=0}^{n-1} \gamma^i H^{i*}.$$

As $H^{i*} \leq H^i$, we have the simple bound $H_n \leq \sigma_n(\gamma H)$, with equality when $H \equiv 1$. Observe that the subsequent bounds depend on the constants d, w, γ, and β or $\underline{\beta}$.

THEOREM 3.5. *Assume* (A1)–(A4). *Then* (3.5) *holds for the bounded functions*

$$K_n = \max\{w,d\} \cdot \left[\sigma_n(\beta) + (\gamma - \beta) \cdot \sum_{i=0}^{n-2} \gamma^i H^{i*} \sigma_{n-i-1}(\beta)\right] \tag{3.9}$$

$$\leq \max\{w,d\} \cdot \{1 + \sigma_{n-1}(\beta)\,[\beta + (\gamma - \beta) H_{n-1}]\} \tag{3.10}$$

Remark 3.1. Let $\hat{e}_n := \max\{w,d\}\,\sigma_n(\beta)$. Then $(K_n)_1^\infty$ as defined in (3.9) can be computed recursively via

$$K_{n+1} - \hat{e}_{n+1} = (\gamma - \beta)\hat{e}_n + \gamma H * (K_n - \hat{e}_n), \quad n \geq 0. \tag{3.11}$$

Proof of Theorem 3.5. For use later in the proof of Proposition 4.2, we prove a bit more than needed for the moment. (a) Consider K_n as defined in (3.9), but with arbitrary numbers $w \in \mathbb{R}_+$, $d \in \mathbb{R}$. Then $(K_n)_0^\infty$ still satisfies (3.11). We show that $(K_n)_0^\infty$ satisfies (3.6). First, (3.6) holds for $n = 0$. Thus assume (3.6) to hold for some $n \geq 0$. Since $\hat{e}_n \geq 0$ ensures $e^{\delta z}\hat{e}_n \geq e^{\delta t}\hat{e}_n$ for $z \geq t$, since $\hat{e}_1 \geq w$ and since $\gamma \geq \int \mu(s,a,dz) e^{\delta z}$ we get

$$\hat{e}_{n+1} + (\gamma - \beta)\hat{e}_n = \hat{e}_1 + \gamma \hat{e}_n \geq w + \hat{e}_n \int \mu(s,a,dz) \left(1_{[0,t]}(z) e^{\delta z} + 1_{(t,\infty)}(z) e^{\delta t}\right).$$

Further, since $H \leq_{st} G(s,a,\cdot), (s,a) \in D$, and since $K_n(t-z) - \hat{e}_n$ is nonnegative and decreasing in z by (3.9),

$$\begin{aligned}\gamma H * (K_n - \hat{e}_n)(t) &\geq \gamma \int_{[0,t]} G(s,a,dz)\left[K_n(t-z) - \hat{e}_n\right] \\ &\geq \int_{[0,t]} \mu(s,a,dz)e^{\delta z}\left[K_n(t-z) - \hat{e}_n\right].\end{aligned}$$

By combining (3.11) with both inequalities, (3.6) is obtained for $n+1$.

(b) Using (a) for the original w and d, the assertion follows immediately from Lemma 3.4. □

Our second type of bounds uses e_n instead of $\hat{e}_n$.

THEOREM 3.6. *Assume* (A1)–(A4). *Then* (3.5) *holds for the bounded function*

$$\begin{aligned}&(3.12) \qquad K_n = e_n + \sum_{i=0}^{n-1} \gamma^i H^{i*}\left[w - d + (\gamma-\beta)e_{n-i-1}^+\right]^+ \\ &(3.13) \qquad \leq e_n + \left[w - d + (\gamma-\beta)e_{n-1}^+\right]^+ \cdot H_n.\end{aligned}$$

Remark 3.2. $(K_n)_1^\infty$ as defined in (3.12) can be computed recursively via

$$(3.14) \qquad K_{n+1} - e_{n+1} = \left[w - d + (\gamma-\beta)e_n^+\right]^+ + \gamma H * (K_n - e_n), \quad n \in \mathbb{N}_0.$$

Proof of Theorem 3.6. (a) Consider K_n as defined in (3.12), but with arbitrary numbers $w \in \mathbb{R}_+$, $d \in \mathbb{R}$. Then $(K_n)_1^\infty$ still satisfies (3.14). We show that $(K_n)_0^\infty$ satisfies (3.6). First, it follows from (3.14) by induction that $K_n - e_n \geq 0$. Then, using

$$\beta(x) := \beta \text{ if } x \geq 0, \text{ and } \beta(x) := \underline{\beta} \text{ else,}$$

we obtain from (3.14)

$$\begin{aligned}K_{n+1} &= d + \beta(d)e_n + \left[w - d + (\gamma-\beta)e_n^+\right]^+ + \gamma H * (K_n - e_n) \\ &= \max\{d + \beta(d)e_n, w + \beta(d)e_n + (\gamma-\beta)e_n^+\} + \gamma H * (K_n - e_n).\end{aligned}$$

As $\beta(d)\, e_n - \beta\, e_n^+ = -\underline{\beta}e_n^-$ we get

$$(3.15) \qquad K_{n+1} \geq w + \gamma e_n^+ - \underline{\beta}e_n^- + \gamma H * (K_n - e_n).$$

Obviously (3.6) holds for $n = 0$. Assume it to hold for $n \geq 0$. Then from (3.15), as $K_n(t-z) - e_n$ is nonnegative and decreasing in z by (3.12),

$$\begin{aligned}K_{n+1} - w \geq\ & e_n^+ \int_{[0,t]} \mu(s,a,dz)e^{\delta z} + e_n^+ e^{\delta t}\mu(s,a,(t,\infty)) \\ & - e_n^- \int_{[0,t]} \mu(s,a,dz)e^{\delta z} - e_n^- e^{\delta t}\mu(s,a,(t,\infty)) \\ & + \int_{[0,t]} \mu(s,a,dz)e^{\delta z}\left(K_n(t-z) - e_n\right) \\ =\ & e_n e^{\delta t}\mu(s,a,(t,\infty)) + \int_{[0,t]} \mu(s,a,dz)e^{\delta z} K_n(t-z).\end{aligned}$$

Thus (3.6) holds for $n+1$.

(b) Now the assertion follows from Lemma 3.4, using (a) for the original w and d. $\square$

In general neither of the two bounds in (3.9) and (3.12) are better than the other. For example, if $w = 0$ and $d < 0$, then K_n from (3.9) equals zero while K_n from (3.12) equals $d\sum_{i=0}^{n-1}(\beta^i - \gamma^i\, H^{i*})$, which may be positive or negative.

Simpler, yet weaker bounds than those in Theorems 3.5 and 3.6 can easily be obtained in case $\gamma \leq 1$ from (3.9) and (3.13) by incorporating the discounted renewal function

$$H_\infty := \sum_{i=0}^{\infty} \gamma^i H^{i*}.$$

Note that H_∞ is bounded for $\gamma < 1$ by $1/(1-\gamma\, H) \leq 1/(1-\gamma)$ and (by a well-known result in renewal theory) affinely upper bounded for $\gamma = 1$. The condition $\gamma \leq 1$ in Remark 3.3 and Theorem 3.7 below is fulfilled if (A1) holds with $b \equiv 1$. If in addition S and A are finite, then $\beta < 1$.

Remark 3.3. Assume (A1)–(A4) and $\gamma \leq 1$. The upper bound (3.5) holds for

$$K_n := e_n + \left[w - d + (\gamma - \beta)e^+_{n-1}\right]^+ \cdot H_\infty \tag{3.16}$$

and also for

$$K_n := e_n + (w + d^-)H_n \tag{3.17}$$

$$\leq e_n + (w + d^-)H_\infty. \tag{3.18}$$

Note that (3.17) follows from (3.16) as $(\gamma - \beta)e^+_{n-1} \leq (1 - \beta^n)d^+ \leq d^+$.

It is well known that under $\beta < 1$ the value functions V_n and V^t_n converge toward the value functions V and V^t of the infinite-stage Markov renewal programs MRP and MRP$'$, resp. (This holds also if T_∞, the limit of the sequence of decision epochs T_n for $n \to \infty$, is finite with positive probability. Therefore we did not include the usual condition which ensures that $T_\infty = \infty$ a.s.)

From (3.10), (3.16), and (3.18) we immediately get the following theorem.

THEOREM 3.7. *Assume* (A1)–(A4), $\gamma \leq 1$, *and* $\beta < 1$. *Then there hold the upper bounds*

$$V^t(s) - V(s) \leq e^{-\delta t}\, b(s)\, K(t),$$

where

$$K := \frac{\max\{w, d\}}{1 - \beta}\Big(1 + (\gamma - \beta)H_\infty\Big) \tag{3.19}$$

or

$$K := \frac{d}{1 - \beta(d)} + \left[w - d + \frac{(\gamma - \beta)d^+}{1 - \beta}\right]^+ H_\infty \tag{3.20}$$

$$\leq \frac{d}{1 - \beta(d)} + (w + d^-)H_\infty. \tag{3.21}$$

Note that d and w depend on V_0, and each V_0 with bounded V_0/b is admissible.

4. Further bounds and the performance of optimal policies for MRP when applied to MRP′. Assume (A1)–(A3). Define $\underline{d}$ and $\underline{w}$ like d and w, resp., with "sup" replaced by "inf." Set

$$\underline{e}_n := \underline{d} \cdot \sigma_n(\underline{\beta}) \text{ if } \underline{d} \geq 0, \text{ and } \underline{e}_n := \underline{d} \cdot \sigma_n(\beta) \text{ otherwise.}$$

We now investigate the performance of an optimal policy $\pi \in \Delta_n$ for MRP, when applied to MRP′. Note that policies for MRP are also policies for MRP′, but not vice versa. We need as preparation a counterpart to Lemma 3.4 which as by-product yields lower bounds for $V_n^t - V_n$.

Lemma 4.1. *Assume* (A1)–(A2).

(a) *Let* $\pi \in \Delta_n$ *be any policy for MRP. Let* $(\underline{K}_n)_0^\infty$ *be any sequence of upper bounded measurable functions on* $\mathbb{R}_+$*, fulfilling* $K_0 \equiv 0$ *and for* $n \in \mathbb{N}_0$*,* $t \geq 0$*, and* $(s,a) \in D$ *the condition*

$$\underline{K}_{n+1}(t) \leq \underline{w} + \underline{e}_n\, e^{\delta t} \mu(s,a,(t,\infty)) + \int_{[0,t]} \mu(s,a,dz) e^{\delta z} \underline{K}_n(t-z). \tag{4.1}$$

Then

$$V_{n\pi}^t(s) - V_{n\pi}(s) \geq e^{-\delta t} b(s) \underline{K}_n(t), \quad t \geq 0. \tag{4.2}$$

(b) *Any sequence* $(\underline{K}_n)_0^\infty$ *which satisfies* (4.1) *and is independent of* π *yields the lower bound*

$$V_n^t(s) - V_n(s) \geq e^{-\delta t} b(s) \underline{K}_n(t), \quad t \geq 0. \tag{4.3}$$

Proof. (a1) For $v \in B$ put $\hat{\underline{d}}_v := \inf_{s \in S}(V_0(s) - v(s))/b(s)$. Let F be the set of decision rules in MRP. Exactly as in the proof of Lemma 3.3 one shows the following: If

$$v'(t,s) - v(s) \geq e^{-\delta t}\, b(s)\, \underline{K}(t),\ t \geq 0,\ s \in S, \tag{4.4}$$

for some upper bounded measurable function $\underline{K}$ on $\mathbb{R}_+$, then for $f \in F$, $t \geq 0$, $s \in S$

$$U_f' v'(t,s) - U_f v(s) := L'v'(t,s,f(s)) - Lv(s,f(s)) \geq e^{-\delta t} b(s) \hat{\underline{K}}_v(t),$$

where

$$\hat{\underline{K}}_v(t) :=\underline{w} + \inf_{(s,a) \in D} \Big[\hat{\underline{d}}_v\, e^{\delta t} \mu(s,a,(t,\infty)) \\ + \int_{[0,t]} \mu(s,a,dz) e^{\delta z} \underline{K}(t-z) \Big].$$

(a2) Assertion (a) for $n = 1$ follows from (a1) with $v'(t,s) := V_0(s)\, 1_{\mathbb{R}_+}(t)$, $v(s) := V_0(s)$. Assume it to hold for some $n \geq 1$. It is well known that for any $(n+1)$-stage policy $\pi := (f_{n+1}, f_n, \ldots, f_1) =: (f, \sigma)$ there holds $V_{n+1,\pi}' = U_{f_{n+1}}' V_{n\sigma}'$ with $V_{n\sigma}'(t,s) := V_{n\sigma}^t(s)$ and $V_{n+1,\pi} = U_{f_{n+1}} V_{n\sigma}$. Now (a) follows for $n+1$ from (a1) with $v' := V_{n\sigma}'$ and $v := V_{n\sigma}$, as

$$\hat{\underline{d}}_v = \inf_s (V_0(s) - V_{n\sigma}(s))/b(s) \geq \inf_s (V_0(s) - V_n(s))/b(s) = \underline{e}_n.$$

(b) The assertion follows from (a) as

$$V_n^t(s) \geq \sup_{\pi \in \Delta_n} V_{n\pi}^t(s) \geq \sup_{\pi \in \Delta_n} V_{n\pi}(s) + e^{-\delta t} b(s) \underline{K}_n(t). \qquad \square$$

We now list a few consequences of Lemma 4.1.

Proposition 4.2. *Assume* (A1)–(A4).

(a) *The inequalities* (4.2) *and* (4.3) *hold for both*

$$\underline{K}_n := \min\{\underline{w}, \underline{d}\} \cdot [1 + \sigma_{n-1}(\beta)(\beta + (\gamma - \beta)\, H_{n-1})] \tag{4.5}$$

and

$$\underline{K}_n := \underline{e}_n - [\underline{w} - \underline{d} - (\gamma - \beta)\underline{e}_{n-1}^{-}]^{-} \cdot H_n. \tag{4.6}$$

(b) *Put*

$$d_0 := \sup_{s \in S} (|V_0(s) - V_1(s)|/b(s)) = \max(d, -\underline{d})$$

and

$$w_0 := \sup\{|\Delta^t(s,a)|/b(s) : (s,a) \in D,\, t \geq 0\} = \max(w, -\underline{w}).$$

There holds

$$|V_n^t(s) - V_n(s)| \leq e^{-\delta t} b(s) K_n(t), \quad t \geq 0, \tag{4.7}$$

for

$$K_n := \max\{w_0, d_0\} \cdot [1 + \sigma_{n-1}(\beta)(\beta + (\gamma - \beta)\, H_{n-1})]\,. \tag{4.8}$$

(c) *If* $\gamma \leq 1$, *then* (4.7) *holds for*

$$K_n := d_0\, \sigma_n(\beta) + (w_0 + d_0) H_n.$$

If in addition $\beta < 1$, *there holds*

$$|V^t(s) - V(s)| \leq e^{-\delta t}\, b(s)\, K(t),$$

both for

$$K := \frac{\max\{w_0, d_0\}}{1 - \beta}(1 + (\gamma - \beta) H_\infty) \tag{4.9}$$

and for

$$K := \frac{d_0}{1 - \beta} + (w_0 + d_0)\, H_\infty. \tag{4.10}$$

Proof. (a) Instead of (4.5) we verify the stronger assertion that (4.1) holds for

$$\begin{aligned}\underline{K}_n &:= -\max\{-\underline{w}, -\underline{d}\} \cdot \left[\sigma_n(\beta) + (\gamma - \beta) \cdot \sum_{i=0}^{n-2} \gamma^i H^{i*} \sigma_{n-i-1}(\beta)\right] \\ &\geq \min\{\underline{w}, \underline{d}\} \cdot \{1 + \sigma_{n-1}(\beta)\,[\beta + (\gamma - \beta) H_{n-1}]\}\,.\end{aligned}$$

In fact, part (a1) of the proof of Theorem 3.5 with w and d replaced by $-\underline{w}$ and by $-\underline{d}$, resp., shows that $(\underline{K}_n)_0^\infty$ satisfies (4.1).

(b) The proof of (4.6) is similar, and the other bounds follow easily. □

Combining (4.2) with $\underline{K}_n$ from (4.5) and (3.5) with K_n from (3.10), we obtain part (a) of our subsequent main result of this section. Part (b) follows by letting n tend to ∞. Several similar bounds may be derived from the preceding results.

THEOREM 4.3. *Assume* (A1)–(A4). *Put* $\Theta := \max(w, d) - \min(\underline{w}, \underline{d})$.

(a) *If* $\pi \in \Delta_n$ *is optimal for MRP, then*

$$V^t_{n\pi}(s) \geq V^t_n(s) - e^{-\delta t} b(s) K_n(t), \tag{4.11}$$

where

$$K_n := \Theta \cdot [1 + \sigma_{n-1}(\beta)(\beta + (\gamma - \beta) H_{n-1})] . \tag{4.12}$$

(b) *If* $\beta < 1, \gamma \leq 1$ *and if* $(f, f, \ldots)$ *is a stationary optimal policy for the infinite-stage MRP, then*

$$V^t_f(s) \geq V^t(s) - e^{-\delta t} b(s) K(t), \tag{4.13}$$

where

$$K := \Theta \cdot [1 + (\gamma - \beta) H_\infty] / (1 - \beta). \tag{4.14}$$

5. Bounds for the discounted renewal function H_∞.

Throughout this section we assume $\gamma < 1$. In what follows we are interested in deriving easy-to-calculate bounds for $H_\infty = \sum_{i=0}^\infty \gamma^i H^{i*}$. First, $H_n \leq \sigma_n(\gamma H)$ implies $H_\infty \leq 1/(1 - \gamma H)$. Our further bounds for H_∞ are obtained by means of bounds for

$$W(t) := (1-\gamma) H_\infty(t) = 1 - \gamma + \gamma H * W(t), \quad t \in \mathbb{R}_+,$$

(and $W(t) := 0$ for $t < 0$). W can be thought of as the distribution function of a random sum $Z = Y_1 + \cdots + Y_\tau$ of independently and identically distributed (i.i.d.) nonnegative random variables $Y_1, Y_2, \ldots$ with distribution function H such that $\tau + 1 \sim$ Geo$(1-\gamma)$. In general, W has to be determined numerically, as explicit representations of W are only known for some special cases. Denote by *Geo*(p), $0 < p \leq 1$, the p.d. with discrete density $p(1-p)^{k-1}$, $k \in \mathbb{N}$. Using, e.g., Laplace transforms one obtains

(a) If $Y_1 \sim$ Bi$(1, p)$, $0 \leq p \leq 1$, then $Z + 1 \sim$ Geo(p'), where $p' := (1-\gamma)/(1 - \gamma(1-p))$. Hence

$$W(k) = 1 - (1 - p')^{k+1}, \quad k \in \mathbb{N}_0. \tag{5.1}$$

(b) If $Y_1 + 1 \sim$ Geo(p), then $Z \sim (1-\gamma)\delta_0 + \gamma\,Geo(p')$, where $1 - p' = (1-p)/(1-\gamma p)$, i.e.,

$$W(k) = 1 - \gamma(1 - p')^{k+1}, \quad k \in \mathbb{N}_0.$$

(c) If $Y_1 \sim$ Exp(λ), then $Z \sim (1-\gamma)\delta_0 + \gamma\,Exp((1-\gamma)\lambda)$, i.e.,

$$W(t) = 1 - \gamma e^{-(1-\gamma)\lambda t}, \quad t \geq 0. \tag{5.2}$$

Our numerical approach for calculating W is based on the following:

(A5) The p.d. corresponding to H is concentrated on $\{0, 1, \ldots, m\}$ for some $m \in \mathbb{N}$, and the discrete density g of H satisfies $g(m) > 0$.

Then the relation $W = 1 - \gamma + \gamma H * W$ yields the following recursive method for computing W, starting with $W(0) = (1-\gamma)\,/\,(1 - \gamma g(0))$:

$$W(k) = \frac{1}{1 - \gamma g(0)} \cdot \left\{ 1 - \gamma + \gamma \cdot \sum_{i=1}^{k \wedge m} g(i)\, W(k-i) \right\}, \quad k \in \mathbb{N}, \tag{5.3}$$

(It is obvious how (5.3) must be modified for the more general case where the probability distribution corresponding to H is concentrated on $\{0, \varepsilon, \ldots, m\varepsilon\}$ for some $m \in \mathbb{N}$ and some $\varepsilon > 0$.)

Having computed $W(k)$ by means of (5.3) up to some $k_0 \geq m$, we can find bounds for $W(k), k > k_0$, by extrapolation. As the method may be of independent interest, we state it for more general difference equations than (5.3), as follows.

PROPOSITION 5.1. *Let $(b_k)_0^\infty$ be the solution of the homogeneous linear difference equation of order $m \in \mathbb{N}$*

$$b_k = \sum_{i=1}^{m} d_i \, b_{k-i}, \quad k \geq m,$$

with nonnegative coefficients $d_i, 1 \leq i \leq m$, with $d_m > 0$ and positive initial values $b_j, 0 \leq j \leq m-1$. Put

$$c_k := \min_{1 \leq i \leq m} \frac{b_{k-i+1}}{b_{k-i}}, \quad k \geq m.$$

Then $(c_k)_m^\infty$ is increasing, and for $k \geq m$ and $j \in \mathbb{N}$ we have

$$b_{k+j} \geq c_{k+1}^{j-1} \cdot b_{k+1} \geq c_k^j \cdot b_k.$$

Analogously, for the decreasing sequence of numbers

$$\underline{c}_k := \max_{1 \leq i \leq m} \frac{b_{k-i+1}}{b_{k-i}}, \quad k \geq m,$$

we obtain

$$b_{k+j} \leq \underline{c}_{k+1}^{j-1} \cdot b_{k+1} \leq \underline{c}_k^j \cdot b_k.$$

Proof. First, $d_m > 0$ and $b_j > 0$ for $0 \leq j \leq m-1$ implies $b_i > 0$ for $i \geq m$ and hence $c_k > 0$ for $k \geq m$. Fix $k \geq m$. The definition of c_k implies $b_{k-i+1} \geq c_k b_{k-i}$ for $1 \leq i \leq m$. Therefore

$$b_{k+1} \geq c_k \sum_{i=1}^{m} d_i b_{k-i} = c_k b_k, \quad k \geq m. \tag{5.4}$$

Thus $\frac{b_{k+1}}{b_k} \geq c_k$, which implies

$$c_{k+1} = \min \left\{ \frac{b_{k+1}}{b_k}, \ldots, \frac{b_{k-m+2}}{b_{k-m+1}} \right\} \geq \min \left\{ \frac{b_{k+1}}{b_k}, c_k \right\} \geq c_k. \tag{5.5}$$

Finally the first assertion follows from (5.4) and (5.5) by induction on j, while the second one can be obtained analogously. □

For $k \geq m$ the recursion (5.3) may be written in terms of $\overline{W} = 1 - W$, yielding

$$\overline{W}(k) = \frac{\gamma}{1 - \gamma g(0)} \sum_{i=1}^{m} g(i) \overline{W}(k-i), \quad k \geq m. \tag{5.6}$$

Equation (5.3) implies $W(k) \leq W(0) + \overline{W}(0) \, W(k-1)$, from which we infer both

$$\overline{W}(k)/\overline{W}(k-1) \geq \overline{W}(0) = \gamma(1 - g(0))/(1 - \gamma g(0)) > 0$$

and $\overline{W}(k) \geq \overline{W}(0)^k > 0$, $k \in \mathbb{N}$.

Applying Proposition 5.1 to the difference equation (5.3) yields the announced upper bounds for W, as follows.

THEOREM 5.2. *Assume* (A5). *For all* $k \geq m, j \in \mathbb{N}$

$$W(k+j) \leq 1 - c_{k+1}^{j-1} \cdot \overline{W}(k+1) \leq 1 - c_k^j \cdot \overline{W}(k), \tag{5.7}$$

where

$$c_k := \min_{1 \leq i \leq m} \frac{\overline{W}(k-i+1)}{\overline{W}(k-i)} \in [\overline{W}(0), 1]$$

is increasing in k.

Note that in the special case $m = 1$, which corresponds to $Y_1 \sim \mathrm{Bi}(1, 1-g(0))$, the bounds for W coincide with the exact values given in (5.1). Concerning the asymptotic behavior of the numbers c_k, the following can be said. Assume that the different roots of the characteristic equation of the difference equation (5.6), i.e., of the equation

$$B(\lambda) := (1 - \gamma g(0))\lambda^m - \gamma g(1)\lambda^{m-1} - \cdots - \gamma g(m) = 0, \tag{5.8}$$

have distinct moduli. By Descartes's rule of signs, (5.8) has a unique positive root λ^*, and $\lambda^* < 1$, as $B(0) = -\gamma g(m) < 0 < B(1) = 1 - \gamma$. Now a well-known result (cf., e.g., Elaydi [3, p. 310]) implies that $\lim_{k\to\infty} \overline{W}(k+1)/\overline{W}(k) = \lambda^*$, and hence $\lim_{k\to\infty} c_k = \lambda^*$. Similar monotone lower bounds for W can be obtained, using the numbers $\underline{c}_k$, defined like c_k with "min" replaced by "max." Also the asymptotic behavior of $\underline{c}_k$ is the same.

We finally remark that our monotone upper (and lower) bounds for W are of independent interest, e.g., they may be exploited for evaluating the equilibrium waiting time distribution for an M/G/1 queue. Further, there is an application in the collective risk theory, where one refers to W as the distribution function of the total claim amounts. These fields of active research have led to various (nonadaptive) exponential bounds on the tails of W. We only refer to Lin [8] and the references given there.

A second type of upper bounds for W is due to Kalashnikov [7], who utilizes Jensen's inequality to prove

$$W(t) \leq 1 - \gamma^{R(t)}, \quad t \geq 0, \tag{5.9}$$

where $R(t) := \sum_{i=0}^{\infty} H^{i*}$ denotes the standard renewal function. Some elementary upper bounds for $R(t)$ (e.g., the well-known ones of an affine type $\nu_1 t + \nu_2$ proved by Lorden [9] and Brown [1]) may be used to obtain from (5.9) simple upper bounds for W, too.

6. Examples. As a test for the quality of our bounds we first derive from our bounds the classical ones for the standard Markovian decision model. Then we look at three special examples: a stopping problem, a renewal problem, and an allocation problem. In all examples $\tilde{r}$ and $\tilde{r}'$ have the special form (2.1) and (2.3) with $\tilde{r}_3 \equiv 0$, resp. The computation of w is facilitated by the formula

$$\begin{aligned} r^t(s,a) - r(s,a) = \int \kappa(s,a,d(z,u))\, 1_{(t,\infty)}(z) \Bigg[& e^{-\alpha t} R(t,s,a,z,u) \\ & - \int_t^z e^{-\alpha y} \tilde{r}_2(s,a,y) dy \Bigg]. \end{aligned} \tag{6.1}$$

Example 6.1 (the standard Markovian decision model). *Consider the familiar infinite-stage Markovian decision model with transition law $P(s,a,du)$, bounded one-stage rewards $r(s,a)$, and discount factor $\beta < 1$. Being interested in the approximation of the infinite-stage model by a finite-stage one without terminal reward (or the other way around), the results of sections* 3–5 *can be applied to the MRP and MRP′ with $\kappa(s,a,d(z,u)) = \delta_1(dz) \times P(s,a,du)$, $\tilde{r}_1 = r$, $\tilde{r}_2 = R \equiv 0$. Then V^t and V are the value function of the t-stage ($t \in \mathbb{N}$) and the infinite-stage model, resp.*

(A1) and (A2) hold for $0 \leq \delta \leq \alpha$ with $b \equiv 1$, and (A3) is assumed to be true. (A4) holds with $H = 1_{[1,\infty]}$, hence $H_\infty(t) = \sigma_{t+1}(\gamma)$. For $\delta \in [0,\alpha]$ we have $r^t - r \equiv 0$, hence $w = 0$; $\beta = \underline{\beta} = e^{-\alpha} < 1$; $\gamma = e^{-(\alpha-\delta)} \leq 1$. From (3.21) we obtain in case $d \geq 0$

$$V^t(s) - V(s) \leq d\frac{(\beta/\gamma)^t}{1-\beta}, \tag{6.2}$$

and in case $d < 0$

$$V^t(s) - V(s) \leq d\,(\beta/\gamma)^t \left(\frac{1}{1-\beta} - \sigma_{t+1}(\gamma)\right). \tag{6.3}$$

The bounds in (6.2) and (6.3) are minimal for $\delta = \alpha$ (hence $\gamma = 1$) and $\delta = 0$, resp. Then (6.2) becomes a classical bound while (6.3) yields the slightly weaker bound $V^t(s) - V(s) \leq d\beta^{t+1}/(1-\beta)$. Similar arguments also work for lower bounds for $V^t - V$.

Example 6.2 (an optimal stopping problem). *Consider a homogenous Markov chain in continuous time with finite state space $S' \subset \mathbb{R}$. Interpret the states of the Markov chain as offers, which, at the time points the jumps occur, can be accepted (action $a = 1$) or not (action $a = 0$). If an offer is accepted, there is a utility $g(s)$ with $\min_s g(s) = 0$ and the process stops. As long as the process is not stopped there is a cost $c > 0$ per unit time. If the process is not stopped during the time interval $[0,t_0]$ under consideration, there is an additional terminal reward $R_0 \in \mathbb{R}$. Future payoffs are discounted with rate $\alpha > 0$.*

Denote by $\lambda > 0$ the parameter of the exponentially distributed sojourn time. We say that the process goes to ∞ when the stop action is chosen. Then the above is a two action MRP with $S = S' \cup \{\infty\}$, $\kappa(s,a,[0,z] \times \{u\}) = (1-e^{-\lambda z})\,P_{s,u}(a)$, where $P_{s,\infty}(a) = a$ for $s < \infty$ and $P_{\infty,\infty}(a) = 1$. For $s < \infty$ we have $\tilde{r}_1(s,a) = a\,g(s)$; $\tilde{r}_2(s,a,y) = -(1-a)c$; $R(t,s,a,z,u) = (1-a)R_0$, while for $s = \infty$ these quantities vanish.

(A1) and (A2) hold for $b \equiv 1$ if $\delta < \lambda$, (A3) holds as A is finite, and (A4) holds with $H \sim \mathrm{Exp}(\alpha + \lambda - \delta)$. Now we get—using $V_0 :\equiv 0$, $d = 0$, $\beta = \lambda/(\alpha+\lambda) < 1$—$\gamma = \lambda/(\alpha + \lambda - \delta) \leq 1$ and

$$w = \left[R_0 + \frac{c}{\lambda+\alpha}\right]^+.$$

From (3.21) we obtain for $\delta < \alpha$, hence $\gamma < 1$, using (5.2)

$$V^t(s) - V(s) \leq e^{-\delta t} w \frac{1 - \gamma e^{-(\alpha+\lambda-\delta)(1-\gamma)t}}{1-\gamma},$$

and for $\delta = \alpha$

$$V^t(s) - V(s) \leq e^{-\alpha t} w(\lambda t + 1).$$

Example 6.3 (an optimal renewal problem). *Consider a system (or unit) which is subject to failure. Upon failure the system is renewed instantaneously at a cost* $c(a) \geq 0$, *depending on the type* a *of renewal. Associated with a renewal of type* a *(action* $a \in \{1, \dots, m\}$ *with* $m \in \mathbb{N}$*) is an exponentially distributed lifetime with parameter* $\lambda_a > 0$ *fulfilling* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0$. *Further we assume* $0 \leq c(1) \leq c(2) \leq \cdots \leq c(m)$.

We want to find the successive renewals for which the expected total cost guaranteeing a total lifetime of at least t_0 time units becomes minimal.

Note that S consists of one element only. Letting $\tilde{r}_1(s,a) = -c(a)$ and $\tilde{r}_2 = R \equiv 0$, we see, that (A1) and (A2) hold for $b \equiv 1$ if $\delta < \lambda_m$. (A3) holds as A is finite and (A4) holds with $H \sim \operatorname{Exp}(\alpha + \lambda - \delta)$. Now we get from (3.21), using $\beta = \lambda_1/(\lambda_1 + \alpha)$ and $V_0 :\equiv 0$, $w = 0$ and $d = c(1) \geq 0$ for all $\delta \leq \alpha, \delta < \lambda_m$,

$$V^t(s) - V(s) \leq e^{-\delta t} c(1) \left(\lambda_1 + \alpha\right) / \alpha.$$

Example 6.4 (optimal N-stage allocation with decision times according to a Poisson process). *Initially a capital of amount* $\overline{s} > 0$ *is available. The times between successive investment decisions are i.i.d. random variables distributed as* $\operatorname{Exp}(\lambda)$. *Let* s *be the momentarily available capital. Then each investment of amount* $a \in [0, s]$ *has utility* $\sqrt{a}$. *The terminal utilities are* $R(t,s,a,z,u) = \sqrt{s-a}$ *and* $V_0(s) = \sqrt{s}$*, resp. The discount rate is* $\alpha > 0$. *It follows that* $\kappa(s,a,d(z,u)) = \operatorname{Exp}(\lambda) \times \delta_{s-a}$, $\tilde{r}_1(s,a) = \sqrt{a}$*, and* $\tilde{r}_2 \equiv 0$. *From* (2.4) *we get*

$$\tilde{h}(t,s,a,z,u) = \sqrt{a} + \sqrt{s-a}\, e^{-\alpha t} \leq 2\sqrt{s} =: h(s,a,z,u).$$

We assume $0 \leq \delta \leq \alpha$ and $\delta < \lambda$. Then (A1) and (A2) hold with $b(s) = \varepsilon + \sqrt{s}$ for any $\varepsilon > 0$, while (A3) can be verified sequentially using Lemma 4.2 in Hinderer and Stieglitz [5]. Furthermore, $\beta = \lambda/(\alpha + \lambda)$. Using

$$\sup_{0 \leq a \leq s} \left(\sqrt{a} + c\sqrt{s-a}\right) = \sqrt{1+c^2}\,\sqrt{s},\ c \geq 0,$$

we get from (2.6) by induction the well-known result that

$$V_n(s) = \sqrt{\sigma_{n+1}(\beta^2)}\,\sqrt{s},$$

and in the same way from (2.7)

$$V_n^t(s) = d_n(t)\sqrt{s} \text{ for } t \geq 0,$$

where $d_0(\cdot) :\equiv 1$, and

$$d_{n+1}^2(t) = 1 + \left[e^{-(\alpha+\lambda)t} + \frac{\lambda}{\alpha+\lambda} \cdot \operatorname{Exp}(\alpha+\lambda) * d_n(t)\right]^2, \quad n \geq 0.$$

With $\eta := \lambda + \alpha - \delta$ we have $\gamma = \lambda/\eta \leq 1$, $d = 0$, $w \leq \alpha/(\alpha+\lambda)$. By applying (3.17) with $H(t) = 1 - e^{-\eta t}$ and letting ϵ tend to zero, we obtain the following upper bound for the complicated function $d_n(\cdot)$:

$$d_n(t) \leq \sqrt{\sigma_{n+1}(\beta^2)} + \alpha\, e^{-\delta t} \sigma_n(\gamma\, H(t))/(\alpha+\lambda). \tag{6.4}$$

If $\delta < \alpha$ it follows, using (5.2), that $\sigma_n(\gamma\, H(t))$ in (6.4) can be replaced by

$$H_\infty(t) = \left(\eta - \lambda e^{-(\alpha-\delta)t}\right) / \left(\alpha - \delta\right).$$

Note added in proof. Recursion formula (5.6) has already been used by F. Dufresne and H. Gerber, *Three methods to calculate the probability of ruin*, Astin Bulletin, 19 (1989), pp. 71–90.

REFERENCES

[1] M. Brown, *Bounds, inequalities and monotonicity properties for some specialized renewal processes*, Ann. Probab., 8 (1980), pp. 227–240.

[2] J.S. De Cani, *A dynamic programming algorithm for embedded Markov chains when the planning horizon is at infinity*, Man. Sci., 10 (1964), pp. 716–733.

[3] S.N. Elaydi, *An Introduction to Difference Equations*, Springer, Berlin, 1996.

[4] I.B. Gertsbakh, *Models of Preventive Maintenance*, North–Holland, Amsterdam, 1977.

[5] K. Hinderer and M. Stieglitz, *Increasing and Lipschitz continuous minimizers in one-dimensional linear-convex systems without constraints: The continuous and the discrete case*, Math. Methods Oper. Res., 44 (1996), pp. 189–204.

[6] W.S. Jewell, *Markov–Renewal Programming.* I: *Formulation, finite return models*, Oper. Res., 11 (1963), pp. 938–948.

[7] V.V. Kalashnikov, *Two-side estimates of geometric convolutions*, in Stability Problems for Stochastic Models, V.V. Kalashnikov and V.M. Zolotarev, eds., Springer, Berlin, 1993, pp. 76–88.

[8] X. Lin, *Tail of compound distributions and excess time*, J. Appl. Probab., 33 (1996), pp. 184–195.

[9] G. Lorden, *On excess over the boundary*, Ann. Math. Statist., 41 (1970), pp. 520–527.

[10] J.W. Mamer, *Successive approximations for finite-horizon, semi-Markov decision processes with application to asset liquidation*, Oper. Res., 34 (1986), pp. 638–644.

[11] H. Schellhaas, *Markov renewal decision processes with finite time horizon*, OR Spektrum, 2 (1980), pp. 33–40.

[12] N.M. van Dijk, *Transient error bound analysis for continuous–time Markov reward structures*, Performance Evaluation, 13 (1991), pp. 147–158.

[13] K.-H. Waldmann, *On bounds for dynamic programs*, Math. Oper. Res., 10 (1985), pp. 220–232.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 2, pp. 521–542

BOUNDARY OBSERVABILITY, CONTROLLABILITY, AND STABILIZATION OF LINEAR ELASTODYNAMIC SYSTEMS*

FATIHA ALABAU† AND VILMOS KOMORNIK‡

Abstract. In 1988 Lions obtained observability and exact controllability results for linear homogeneous isotropic elastodynamic systems [*SIAM Rev.*, 30 (1988), pp. 1–68]. Applying some new identities we extend his theorems to nonisotropic systems. In 1991 Lagnese obtained uniform stabilizability results for two-dimensional linear homogeneous isotropic systems by applying a somewhat artificial feedback [*Nonlinear Anal.*, 16 (1991), pp. 35–54]. Then he asked whether analogous results hold for a natural and physically implementable boundary feedback. Using some new identities and applying a method introduced in 1987 by Zuazua and the second author [*J. Math. Pures. Appl.*, 69 (1990), pp. 33–54], we give an affirmative answer to this question in all dimensions and also for nonisotropic systems. Moreover, we obtain good decay estimates. Finally, applying a recent general method of uniform stabilization, we construct boundary feedbacks leading to arbitrarily large energy decay rates.

Key words. observability, controllability, stabilizability by feedback, partial differential equation, elasticity

AMS subject classifications. 35L55, 35Q72, 93B05, 93B07, 93C20, 93D15

PII. S0363012996313835

1. Introduction and formulation of the main results.

Let n be a positive integer, and let (a_{ijkl}) be a tensor such that

$$a_{ijkl} = a_{jikl} = a_{klij}$$

(all indices run over the integers $1, \dots, n$), satisfying for some $\alpha > 0$ the ellipticity condition

$$a_{ijkl}\varepsilon_{ij}\varepsilon_{kl} \geq \alpha\varepsilon_{ij}\varepsilon_{ij} \tag{1.1}$$

for every *symmetric* tensor ε_{ij}. (Here and in what follows we shall use the summation convention for repeated indices.)

Let Ω be a nonempty bounded open set in $\mathbb{R}^n$ having a boundary Γ of class C^2. Given a function $\xi = (\xi_1, \dots, \xi_n) : \Omega \to \mathbb{R}^n$, we shall use the notations

$$\varepsilon_{ij} = \tfrac{1}{2}(\xi_{i,j} + \xi_{j,i}), \qquad \sigma_{ij} = a_{ijkl}\varepsilon_{kl},$$

where $\xi_{i,j} = \partial\xi_i/\partial x_j$ and $\xi_{j,i} = \partial\xi_j/\partial x_i$. If it is necessary to be more precise, we shall write $\varepsilon_{ij}(\xi)$ and $\sigma_{ij}(\xi)$ instead of ε_{ij}, σ_{ij}.

*Received by the editors December 18, 1996; accepted for publication (in revised form) October 10, 1997; published electronically December 11, 1998.
http://www.siam.org/journals/sicon/37-2/31383.html

†UFR de Mathématiques et d'Informatique, Université de Bordeaux I, 351, cours de la Libération, 33405 Talence Cédex, France (alabau@math.u-bordeaux.fr).

‡Institut de Recherche Mathématique Avancée, Université Louis Pasteur et CNRS, 7, rue René Descartes, 67084 Strasbourg Cédex, France (komornik@math.u-strasbg.fr).

Consider the problem

$$\begin{cases} \xi_i'' - \sigma_{ij,j} = 0 \quad \text{in} \quad \Omega \times \mathbb{R}, \\ \xi_i = 0 \quad \text{on} \quad \Gamma \times \mathbb{R}, \\ \xi_i(0) = \xi_i^0 \quad \text{and} \quad \xi_i'(0) = \xi_i^1 \quad \text{in} \quad \Omega, \\ i = 1, \dots, n, \end{cases} \tag{1.2}$$

where $\xi_i' = \partial \xi_i / \partial t$ and $\xi_i'' = \partial^2 \xi_i / \partial t^2$. For $n = 1, 2, 3$ this is the linear model of homogeneous elastodynamic systems.

We recall (see, e.g., [7]) that this problem is well-posed in the following sense:

- Given $(\xi_0, \xi_1) \in H_0^1(\Omega)^n \times L^2(\Omega)^n$ arbitrarily, the problem (1.2) has a unique (so-called *weak*) solution

$$\xi \in C(\mathbb{R}; H_0^1(\Omega)^n) \cap C^1(\mathbb{R}; L^2(\Omega)^n).$$

- If $(\xi_0, \xi_1) \in (H^2(\Omega) \cap H_0^1(\Omega))^n \times H_0^1(\Omega)^n$, then the solution is more regular:

$$\xi \in C(\mathbb{R}; (H^2(\Omega) \cap H_0^1(\Omega))^n) \cap C^1(\mathbb{R}; H_0^1(\Omega)^n) \cap C^2(\mathbb{R}; L^2(\Omega)^n);$$

it is called a *strong* solution.

- The energy of the (weak) solutions, defined by the formula

$$E = \tfrac{1}{2} \textstyle\int_\Omega \xi_i' \xi_i' + \sigma_{ij} \varepsilon_{ij} \; dx, \tag{1.3}$$

is independent of the time $t \in \mathbb{R}$.

In order to simplify we shall only consider in this paper time-independent homogeneous systems: the coefficients a_{ijkl} do not depend on $x \in \Omega$ or on $t \in \mathbb{R}$.

Fix a point $x_0 \in \mathbb{R}^n$ arbitrarily and fix a measurable partition Γ_0, Γ_1 of Γ such that

$$(x - x_0) \cdot \nu(x) \leq 0 \quad \text{for all} \quad x \in \Gamma_0, \tag{1.4}$$

where $\nu = (\nu_1, \dots, \nu_n)$ denotes the outward unit normal vector to Γ. In other words, Γ_1 contains all points $x \in \Gamma$ for which $(x - x_0) \cdot \nu(x) > 0$. (The simplest way to satisfy (1.4) is to choose $\Gamma_0 = \emptyset$ and $\Gamma_1 = \Gamma$.) Set

$$R = \sup\{|x - x_0| \ : \ x \in \Omega\}. \tag{1.5}$$

Then we have the following theorem.

THEOREM 1.1. *Assume* (1.1), (1.4), *and let* $T > 2\sqrt{2/\alpha}R$. *Then there exist two positive constants* c_1 *and* c_2 *such that every strong solution of* (1.2) *satisfies the inequalities*

$$c_1 E \leq \int_0^T \int_{\Gamma_1} \sigma_{ij} \varepsilon_{ij} \; d\Gamma \; dt \leq c_2 E. \tag{1.6}$$

Remarks.

- In the isotropic case Theorem 1.1 reduces to a former result of Lions [21].

- Theorem 1.1 means that in some sense the observation of the solution in a neighborhood of the boundary during a sufficiently large time allows one to determine the initial data. Indeed, if two solutions coincide in this set, then the boundary integral in (1.6) for their difference vanishes, and therefore the energy of their difference is equal to zero by the first inequality in (1.6). This implies that the two solutions correspond in fact to the same initial data, and hence they are identical.
- By a simple density argument, the second estimate in (1.6) allows us to *define* the trace of $\sigma_{ij}\varepsilon_{ij}$ on $\Gamma_1 \times \mathbb{R}$ as an element of $L^2_{loc}(\mathbb{R}; L^2(\Gamma_1))$ for every *weak solution* of (1.2). Then the estimates (1.6) remain valid for all weak solutions, too.
- Due to the finite propagation property of the system of elasticity, the first inequality in (1.6) cannot hold for arbitrarily small T. The proof of Theorem 2 in [13] shows that the condition $T > 2\sqrt{2/\alpha}R$ is the best possible if the system is isotropic, i.e., when

$$a_{ijkl} = \lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}),$$

where λ and μ are the (positive) Lamé constants.

Applying the Hilbert uniqueness method (HUM) we shall deduce from Theorem 1.1 an exact controllability result for the nonhomogeneous problem

$$\begin{cases} y_i'' - \sigma_{ij,j}(y) = 0 \quad \text{in} \quad \Omega \times \mathbb{R}, \\ y_i = v_i \quad \text{on} \quad \Gamma \times \mathbb{R}, \\ y_i(0) = y_i^0 \quad \text{and} \quad y_i'(0) = y_i^1 \quad \text{in} \quad \Omega, \\ i = 1, \dots, n. \end{cases} \tag{1.7}$$

Theorem 1.2. *Assume* (1.1), (1.4) *and fix* $T > 2\sqrt{2/\alpha}R$ *arbitrarily. Then for any given* $y^0, \tilde{y}^0 \in L^2(\Omega)^n$ *and* $y^1, \tilde{y}^1 \in H^{-1}(\Omega)^n$ *there exists* $v \in L^2_{loc}(\mathbb{R}; L^2(\Gamma)^n)$ *such that the solution of* (1.7) *satisfies*

$$y(T) = \tilde{y}^0 \quad \textit{and} \quad y'(T) = \tilde{y}^1 \quad \textit{in} \quad \Omega.$$

Moreover, we may assume that v *vanishes outside of* $\Gamma_1 \times (0, T)$.

This extends an earlier theorem of Lions [21] to nonisotropic systems.

In the second half of the paper we shall study the uniform stabilizability of elasticity systems by applying suitable dissipative boundary feedbacks. A natural and physically implementable system of this type was proposed by Lagnese [17], [18]:

$$\begin{cases} \xi_i'' - \sigma_{ij,j} = 0 \quad \text{in} \quad \Omega \times (0, \infty), \\ \xi_i = 0 \quad \text{on} \quad \Gamma_0 \times (0, \infty), \\ \sigma_{ij}\nu_j + A\xi_i + B\xi_i' = 0 \quad \text{on} \quad \Gamma_1 \times (0, \infty), \\ \xi_i(0) = \xi_i^0 \quad \text{and} \quad \xi_i'(0) = \xi_i^1 \quad \text{in} \quad \Omega, \\ i = 1, \dots, n. \end{cases} \tag{1.8}$$

Here A, B are given nonnegative numbers for simplicity. It is easy to generalize our results to the case where A and B are nonnegative functions of class C^1 on Γ_1. Indeed, defining the energy of the solutions of (1.8) by

$$E = \frac{1}{2}\int_\Omega \xi_i'\xi_i' + \sigma_{ij}\varepsilon_{ij}\ dx + \frac{1}{2}\int_{\Gamma_1} A\xi_i\xi_i\ d\Gamma, \tag{1.9}$$

an easy computation shows that the energy is a nonincreasing function of $t \in [0, +\infty)$.

We shall consider this problem under a rather strict geometric assumption on Ω and Γ_0, Γ_1: we assume that

$$\Omega = \Omega_1 \backslash \overline{\Omega_0}, \tag{1.10}$$

where Ω_1 is an open ball, say $\Omega_1 = B(x_0; R)$, Ω_0 is a star-shaped domain with respect to x_0 whose closure belongs to Ω_1, and

$$\Gamma_0 = \partial\Omega_0, \quad \Gamma_1 = \partial\Omega_1.$$

We do not exclude the case $\Omega_0 = \emptyset$; then $\Omega = B(x_0; R)$, $\Gamma_0 = \emptyset$, and $\Gamma_1 = \Gamma$. Another typical case is

$$\Omega = \{x \in \mathbb{R}^n \ : \ r < |x - x_0| < R\}$$

for some $0 < r < R$, and

$$\Gamma_0 = \{x \in \mathbb{R}^n \ : \ |x - x_0| = r\},$$
$$\Gamma_1 = \{x \in \mathbb{R}^n \ : \ |x - x_0| = R\}.$$

We shall prove the following theorem.

Theorem 1.3. *Assume* (1.1)*, and let* Ω*,* Γ_0*, and* Γ_1 *be as described above. Given two positive constants* A *and* B *with* $A < \alpha/(4R)$*, there exists a positive number* ω *such that all (weak) solutions of* (1.8) *satisfy the energy estimate*

$$E(t) \le E(0)e^{1-\omega t} \tag{1.11}$$

for all $t \ge 0$.

If $\Gamma_0 \ne \emptyset$*, then the result holds also for* $A = 0$.

Remarks.

- This result seems to be new even in the isotropic case.
- Our proof will be based on a Liapunov-type method introduced in [15] and then modified in [10]. We shall also need a crucial new identity (see (4.4)) which will allow us to estimate some boundary integrals.
- The interest of this result can be questioned because of the very restrictive geometric assumptions. We recall, however, that the validity of this result was not at all obvious, and some researchers even conjectured that this result does not hold for any domain. See also a remark in [18, p. 37] in this respect.
- In fact, our proof given below can be easily adapted for slightly more general domains such that Ω_1 is *close* to a ball. We shall outline the suitable modifications of the proof at the end of section 4 for a special case (see Theorem 4.4), and we refer to [22] for more general results. It is plausible that the theorem remains valid if we allow Ω_1 to be a general star-shaped domain with respect to x_0. However, our "elementary" method may not be sufficiently powerful for this. A similar problem for the wave equation was solved earlier by Lasiecka and Triggiani [19] by using microlocal estimates: they relaxed a geometric condition of an earlier result of Lagnese [16]. Adapting their approach, recently Horn [8] obtained very interesting stabilization results for the system of elasticity: she established similar results to our Theorem 1.3, without geometric conditions on Ω, in the particular case of *isotropic* elasticity systems. Unlike our result, due to an indirect compactness-uniqueness argument, she did not obtain explicit constants in the estimates.

- Theorem 1.3 probably remains valid without the condition $A < \alpha/(4R)$, but we could not prove it.

The proof of Theorem 1.3 will provide an explicit constant ω. If ω is bigger, then the energy decay is faster. However, some results of Koch and Tataru [9] indicate that we cannot achieve arbitrarily large decay rates ω by using feedbacks of the type as in (1.8). On the other hand, in [14] a general method was developed for this purpose. This method, similar in its spirit to HUM, allows us to construct boundary feedbacks for observable systems which lead to arbitrarily large decay rates. Applying this approach we shall deduce from Theorem 1.1 the following theorem.

THEOREM 1.4. *Assume* (1.1), (1.4), *and assume that* Γ_0 *is closed. Fix* $\omega > 0$ *arbitrarily. Then there exist two bounded linear maps*

$$P : H^{-1}(\Omega)^n \to H_0^1(\Omega)^n \quad \textit{and} \quad Q : L^2(\Omega)^n \to H_0^1(\Omega)^n$$

and a constant $M > 0$ *such that the problem*

$$\begin{cases} \xi_i'' - \sigma_{ij,j} = 0 \quad \textit{in} \quad \Omega \times \mathbb{R}, \\ \xi_i = 0 \quad \textit{on} \quad \Gamma_0 \times \mathbb{R}, \\ \xi_i = \sigma_{ij}(P\xi' + Q\xi)\nu_j \quad \textit{on} \quad \Gamma_1 \times \mathbb{R}, \\ \xi_i(0) = \xi_i^0 \quad \textit{and} \quad \xi_i'(0) = \xi_i^1 \quad \textit{in} \quad \Omega, \\ i = 1, \dots, n \end{cases} \tag{1.12}$$

is well-posed in $\mathcal{H} = L^2(\Omega)^n \times H^{-1}(\Omega)^n$ *and its solutions satisfy the estimate*

$$\|(\xi, \xi')(t)\|_{\mathcal{H}} \le M \|(\xi, \xi')(0)\|_{\mathcal{H}} e^{-\omega t} \tag{1.13}$$

for all $t \ge 0$.

Remarks.

- Note that (unlike in the preceding theorem) we do not have any geometric assumption on Ω. Thus Theorem 1.4 applies to all bounded domains of class C^2, choosing, for example, $\Gamma_0 = \emptyset$ and $\Gamma_1 = \Gamma$.
- Although the feedback has a more complicated structure than in Theorem 1.3, it can be constructed explicitly. Let us note that Bourquin and Briffaut [3] tested numerically the method of [14]; this required the use of the explicit form of the operators P and Q.
- Our proof allows us to estimate the constant M in (1.13).

Some results of this paper were announced without proof in [1].

2. Observability: Proof of Theorem 1.1. We proceed in three steps.

Step 1. Fix an arbitrary function $h \in W^{1,\infty}(\Omega)^n$ and a number $T > 0$. Integrating by parts it follows from the equation in (1.2) that

$$\begin{aligned} 0 = \int_0^T \int_\Omega (h_m \xi_{i,m})(\xi_i'' - \sigma_{ij,j}) \, dx \, dt \\ = \left[\int_\Omega h_m \xi_{i,m} \xi_i' \, dx \right]_0^T - \int_0^T \int_\Gamma h_m \xi_{i,m} \sigma_{ij} \nu_j \, d\Gamma \, dt \\ + \int_0^T \int_\Omega h_{m,j} \sigma_{ij} \xi_{i,m} + h_m \sigma_{ij} \xi_{i,jm} - \frac{1}{2} h_m (\xi_i' \xi_i')_{,m} \, dx \, dt. \end{aligned}$$

Since

$$\sigma_{ij}\xi_{i,jm} = \sigma_{ij}\varepsilon_{ij,m} = \frac{1}{2}(\sigma_{ij}\varepsilon_{ij})_{,m},$$

integrating by parts the last two terms in the last integral and then multiplying by two the whole expression we obtain the following identity:

$$\begin{aligned}\text{(2.1)}\quad \int_0^T \int_\Gamma 2h_m\xi_{i,m}\sigma_{ij}\nu_j + (h\cdot\nu)(\xi_i'\xi_i' - \sigma_{ij}\varepsilon_{ij})\ d\Gamma\ dt &\\ = \left[\int_\Omega 2h_m\xi_{i,m}\xi_i'\ dx\right]_0^T &\\ + \int_0^T \int_\Omega 2h_{m,j}\sigma_{ij}\xi_{i,m} + (\operatorname{div}\ h)(\xi_i'\xi_i' - \sigma_{ij}\varepsilon_{ij})\ dx\ dt.&\end{aligned}$$

Note that in the proof of identity (2.1) we did not use the boundary conditions in (1.2).

Step 2. Using the Korn inequality and the assumption $h \in W^{1,\infty}(\Omega)^n$, the right-hand side of (2.1) can be easily majorized by a constant multiple cE of the energy E. (Let us recall that in this case the Korn equality is easy to establish; see the identity (2.4) below.)

Furthermore, we deduce from the homogeneous Dirichlet boundary conditions (1.2) that $\xi_i' = 0$ and

$$\xi_{i,m}\nu_j = \xi_{i,\nu}\nu_m\nu_j = \xi_{i,j}\nu_m$$

on Γ, and hence

$$h_m\xi_{i,m}\sigma_{ij}\nu_j = (h\cdot\nu)\sigma_{ij}\xi_{i,j} = (h\cdot\nu)\sigma_{ij}\varepsilon_{ij}.$$

Therefore the left-hand side of (2.1) reduces to

$$\int_0^T \int_\Gamma (h\cdot\nu)\sigma_{ij}\varepsilon_{ij}\ d\Gamma\ dt.$$

Choosing h such that $h = \nu$ on Γ, the second inequality in (1.6) follows with $c_2 = c$.

Step 3. Choosing now $h(x) = x - x_0$, the identity (2.1) reduces to

$$\begin{aligned}\int_0^T \int_\Gamma (h\cdot\nu)\sigma_{ij}\varepsilon_{ij}\ d\Gamma\ dt &\\ = \left[\int_\Omega 2h_m\xi_{i,m}\xi_i'\ dx\right]_0^T + \int_0^T \int_\Omega (2-n)\sigma_{ij}\varepsilon_{ij} + n\xi_i'\xi_i'\ dx\ dt.&\end{aligned}$$

Furthermore, we also deduce from (1.2) that

$$\begin{aligned}0 &= \int_0^T \int_\Omega \xi_i(\xi_i'' - \sigma_{ij,j})\ dx\ dt\\ &= \left[\int_\Omega \xi_i\xi_i'\ dx\right]_0^T - \int_0^T \int_\Gamma \xi_i\sigma_{ij}\nu_j\ d\Gamma\ dt + \int_0^T \int_\Omega \sigma_{ij}\varepsilon_{ij} - \xi_i'\xi_i'\ dx\ dt\\ &= \left[\int_\Omega \xi_i\xi_i'\ dx\right]_0^T + \int_0^T \int_\Omega \sigma_{ij}\varepsilon_{ij} - \xi_i'\xi_i'\ dx\ dt.\end{aligned}$$

Combining with the preceding identity we obtain that

$$\int_0^T \int_\Gamma (h\cdot\nu)\sigma_{ij}\varepsilon_{ij}\ d\Gamma\ dt = \left[\int_\Omega (2h_m\xi_{i,m} + (n-1)\xi_i)\xi_i'\ dx\right]_0^T + \int_0^T \int_\Omega \sigma_{ij}\varepsilon_{ij} + \xi_i'\xi_i'\ dx\ dt$$

whence

$$R\int_0^T \int_{\Gamma_1} \sigma_{ij}\varepsilon_{ij}\ d\Gamma\ dt \geq 2TE - \left|\left[\int_\Omega (2h_m\xi_{i,m} + (n-1)\xi_i)\xi_i'\ dx\right]_0^T\right|. \tag{2.2}$$

Let us majorize the last integral. Thanks to the boundary condition in (1.2) we have

$$\|2h_m\xi_{i,m} + (n-1)\xi_i\|_{L^2(\Omega)} \leq \|2h_m\xi_{i,m}\|_{L^2(\Omega)}$$

for each fixed i. Indeed, as it was shown in [10], we have

$$\begin{aligned}\|2h_m\xi_{i,m} + (n-1)\xi_i\|^2_{L^2(\Omega)} - \|2h_m\xi_{i,m}\|^2_{L^2(\Omega)} &= \int_\Omega (n-1)^2\xi_i^2 + 4(n-1)h_m\xi_i\xi_{i,m}\ dx \\ &= \int_\Omega (n-1)^2\xi_i^2 - 2n(n-1)\xi_i^2\ dx + \int_\Gamma (2n-2)h_m\nu_m\xi_i^2\ d\Gamma \\ &= (1-n^2)\int_\Omega \xi_i^2\ dx \leq 0.\end{aligned}$$

Therefore for any fixed $\delta > 0$ we have

$$\begin{aligned}\left|\int_\Omega (2h_m\xi_{i,m} + (n-1)\xi_i)\xi_i'\ dx\right| &\leq 2R\sum_{i=1}^n \left(\int_\Omega |\nabla\xi_i|^2\ dx\right)^{1/2} \left(\int_\Omega (\xi_i')^2\ dx\right)^{1/2} \\ &\leq R\delta\int_\Omega \xi_{i,m}\xi_{i,m}\ dx + R\delta^{-1}\int_\Omega \xi_i'\xi_i'\ dx.\end{aligned} \tag{2.3}$$

Furthermore, applying the Green formula and using the boundary condition in (1.2), we have

$$\int_\Omega \xi_{i,m}\xi_{i,m}\ dx = \int_\Omega 2\varepsilon_{im}\xi_{i,m}\ dx - \int_\Omega \xi_{m,i}\xi_{i,m}\ dx = \int_\Omega 2\varepsilon_{im}\varepsilon_{im} - \xi_{m,m}\xi_{i,i}\ dx,$$

i.e.,

$$\int_\Omega 2\varepsilon_{im}\varepsilon_{im}\ dx = \int_\Omega \xi_{i,m}\xi_{i,m}\ dx + \int_\Omega |\mathrm{div}\ \xi|^2\ dx. \tag{2.4}$$

It follows from (1.1) and (2.4) that

$$\int_\Omega \xi_{i,m}\xi_{i,m}\ dx \leq (2/\alpha)\int_\Omega \sigma_{im}\varepsilon_{im}\ dx.$$

Substituting into (2.3), choosing $\delta = \sqrt{\alpha/2}$, and using the definition (1.3) of the energy we obtain that

$$\left|\int_\Omega (2h_m \xi_{i,m} + (n-1)\xi_i)\xi_i' \, dx\right| \le 2\sqrt{2/\alpha} RE.$$

Therefore we deduce from (2.2) the inequality

$$R\int_0^T \int_{\Gamma_1} \sigma_{ij}\varepsilon_{ij} \, d\Gamma \, dt \ge (2T - 4\sqrt{2/\alpha_1}R)E,$$

and the first estimate of (1.6) follows with $c_1 = (2T - 4\sqrt{2/\alpha}R)/R$. The proof of Theorem 1.1 is completed.

In the next section we shall also need the following result, which gives an equivalent form of the integral in (1.6).

Lemma 2.1. *Assume* (1.1) *and put*

$$\beta = \sum_{i,j,k,l} |a_{ijkl}|^2.$$

Then every strong solution of (1.2) *satisfies on* $\Gamma \times \mathbb{R}$ *the inequalities*

$$(\alpha/2)\sigma_{ij}\varepsilon_{ij} \le \sum_{i=1}^n |\sigma_{ij}\nu_j|^2 \le (\beta/\alpha)\sigma_{ij}\varepsilon_{ij}. \tag{2.5}$$

Proof. The proof of the second inequality does not use the boundary conditions

$$\sum_i |\sigma_{ij}\nu_j|^2 \le \sum_{i,j} |\sigma_{ij}|^2 = \sum_{i,j} |a_{ijkl}\varepsilon_{kl}|^2$$
$$\le \sum_{i,j}\left(\sum_{k,l}|a_{ijkl}|^2 \sum_{k,l}|\varepsilon_{kl}|^2\right) = \beta\varepsilon_{kl}\varepsilon_{kl} \le (\beta/\alpha)\sigma_{ij}\varepsilon_{ij}.$$

For the proof of the reverse inequality first we note that, thanks to the boundary conditions in (1.2), we have

$$\xi_{i,j}\xi_{i,j} = \frac{1}{2}\sum_{i,j}(\xi_{i,j} + \xi_{j,i})^2 - \xi_{i,j}\xi_{j,i} = 2\varepsilon_{ij}\varepsilon_{ij} - \xi_{i,\nu}\nu_j\xi_{j,\nu}\nu_i = 2\varepsilon_{ij}\varepsilon_{ij} - |\text{div}\ \xi|^2,$$

i.e.,

$$2\varepsilon_{ij}\varepsilon_{ij} = \xi_{i,j}\xi_{i,j} + |\text{div}\ \xi|^2 \qquad \text{on} \qquad \Gamma. \tag{2.6}$$

(Compare with (2.4).) Therefore

$$\sigma_{ij}\varepsilon_{ij} = \sigma_{ij}\xi_{i,j} = \sigma_{ij}\nu_j\xi_{i,\nu} \le \left(\sum_i |\sigma_{ij}\nu_j|^2\right)^{1/2} (\xi_{i,\nu}\xi_{i,\nu})^{1/2}$$
$$\le \left(\sum_i |\sigma_{ij}\nu_j|^2\right)^{1/2} (2\varepsilon_{ij}\varepsilon_{ij})^{1/2} \le \left(\sum_i |\sigma_{ij}\nu_j|^2\right)^{1/2} (2\alpha^{-1}\sigma_{ij}\varepsilon_{ij})^{1/2},$$

and the first inequality in (2.5) follows. □

3. Controllability: Proof of Theorem 1.2. Let us first study the well-posedness of the nonhomogeneous problem (1.7). In order to find a reasonable *definition* of the solution, let us multiply the equation in (1.7) by an arbitrary solution of the problem (1.2) and integrate by parts formally. We obtain that

$$\begin{aligned}\int_0^T \int_\Omega \xi_i(y_i'' - \sigma_{ij,j}(y))\ dx\ dt \\ = \int_0^T \int_\Omega (\xi_i'' - \sigma_{ij,j}(\xi))y_i\ dx\ dt + \left[\int_\Omega \xi_i y_i' - \xi_i' y_i\ dx\right]_0^T \\ + \int_0^T \int_\Gamma -\xi_i\sigma_{ij}(y)\nu_j + \sigma_{ij}(\xi)\nu_j y_i\ d\Gamma\ dt \\ = \left[\int_\Omega \xi_i y_i' - \xi_i' y_i\ dx\right]_0^T + \int_0^T \int_\Gamma \sigma_{ij}(\xi)\nu_j v_i\ d\Gamma\ dt.\end{aligned}$$

Hence, putting

$$H = H^{-1}(\Omega)^n \times L^2(\Omega)^n, \quad \text{and} \quad H' = H_0^1(\Omega)^n \times L^2(\Omega)^n$$

for brevity, we have

$$\begin{aligned}(3.1) \quad \langle (y'(T), -y(T)), (\xi(T), \xi'(T)\rangle_{H,H'} \\ = \langle (y_i^1, -y_i^0), (\xi_i^0, \xi_i^1)\rangle_{H,H'} - \int_0^T \int_\Gamma \sigma_{ij}(\xi)\nu_j v_i\ d\Gamma\ dt.\end{aligned}$$

This leads to the following definition.

Definition. *A solution of* (1.7) *is a* continuous *function* $(y', -y) : \mathbb{R} \to H$ *satisfying the identity* (3.1) *for all* $T \in \mathbb{R}$ *and for all (weak) solutions of the problem* (1.2).

This definition is justified by the following theorem.

Theorem 3.1. *Assume* (1.1). *Then for any given*

$$y^0 \in L^2(\Omega)^n, \quad y^1 \in H^{-1}(\Omega)^n, \quad \textit{and} \quad v \in L^2_{loc}(\mathbb{R}; L^2(\Gamma)^n)$$

the problem (1.7) *has a unique solution satisfying*

$$y \in C(\mathbb{R}; L^2(\Omega)^n) \cap C^1(\mathbb{R}; H^{-1}(\Omega)^n).$$

Furthermore, the linear map $(y^0, y^1, v) \mapsto y$ *is continuous with respect to these topologies.*

Proof. We apply Theorem 1.1. Thanks to the second estimate in (1.6), the right-hand side of the equality (3.1) defines a bounded linear form of $(\xi^0, \xi^1) \in H'$. Since the linear map $(\xi^0, \xi^1) \mapsto (\xi(T), \xi'(T))$ is an automorphism of H' (because the problem (1.2) is reversible), the right-hand side of the equality (3.1) can also be considered as a bounded linear form of $(\xi(T), \xi'(T)) \in H'$. Since $H'' = H$, we conclude the existence of a unique $(y'(T), -y(T)) \in H$ satisfying (3.1).

Since the bounded linear form used in this proof depends continuously on $t \in \mathbb{R}$, the solution y has the regularity required in the theorem. Finally, the bounded linear form clearly depends continuously on the initial data and on the boundary value, hence y also has this property. □

Turning to the proof of Theorem 1.2, let us first note that it is sufficient to consider the case where $\tilde{y}^0 = \tilde{y}^1 = 0$ in Ω. The general case then follows by a standard argument, valid for every linear reversible problem; see, e.g., [12], [14].

The main idea is to seek a control in the form $v_i = \sigma_{ij}(\xi)\nu_j$, where ξ is the solution of (1.2) for some suitable initial data. (Thanks to Theorem 1.1 and to Lemma 2.1, these controls have the required regularity for the well-posedness of (1.7).)

Given $(\xi^0, \xi^1) \in H'$ arbitrarily, first solve the problem (1.2), then solve the problem

$$\begin{cases} y_i'' - \sigma_{ij,j}(y) = 0 \quad \text{in} \quad \Omega \times \mathbb{R}, \quad 1 \le i \le n, \\ y_i = 0 \quad \text{on} \quad \Gamma_0 \times \mathbb{R}, \quad 1 \le i \le n, \\ y_i = \sigma_{ij}(\xi)\nu_j \quad \text{on} \quad \Gamma_1 \times \mathbb{R}, \quad 1 \le i \le n, \\ y_i(T) = y_i'(T) = 0 \quad \text{in} \quad \Omega, \quad 1 \le i \le n, \end{cases} \tag{3.2}$$

and set

$$\Lambda(\xi^0, \xi^1) = (y'(0), -y(0)).$$

(The problem (3.2) is well-posed in an analogous sense as (1.7) because the different choice of the initial time does not change the character of a time-reversible problem.) Obviously, $\Lambda : H' \to H$ is a bounded linear map. Applying HUM, it is sufficient to show that Λ is onto. Indeed, then for any given $(y^0, y^1) \in L^2(\Omega)^n \times H^{-1}(\Omega)^n$ it will suffice to choose the control v defined by $v_i = 0$ on $\Gamma_0 \times (0, \infty)$ and $v_i = \sigma_{ij}(\xi)\nu_j$ on $\Gamma_1 \times (0, \infty)$, where ξ is the solution of (1.2) corresponding to $(\xi^0, \xi^1) = \Lambda^{-1}(y^1, -y^0)$.

We shall prove that Λ is in fact an isomorphism. For this first we prove the identity

$$\langle \Lambda(\xi^0, \xi^1), (\xi^0, \xi^1) \rangle_{H,H'} = \int_0^T \int_{\Gamma_1} \sum_i |\sigma_{ij}(\xi)\nu_j|^2 \, d\Gamma \, dt. \tag{3.3}$$

This follows from the equality

$$\begin{aligned} &\int_0^T \int_\Omega \xi_i (y_i'' - \sigma_{ij,j}(y)) \, dx \, dt \\ &\qquad = \left[\int_\Omega \xi_i y_i' - \xi_i' y_i \, dx \right]_0^T + \int_0^T \int_\Omega (\xi_i'' - \sigma_{ij,j}(\xi)) y_i \, dx \, dt \\ &\qquad\qquad + \int_0^T \int_\Gamma -\xi_i \sigma_{ij}(y)\nu_j + \sigma_{ij}(\xi)\nu_j y_i \, d\Gamma \, dt, \end{aligned}$$

using the definition of Λ and the equations of (1.2) and (3.2).

Applying the first estimate of (1.6) in Theorem 1.1 and Lemma 2.1, we conclude from the identity (3.3) that Λ is coercive. Applying the Lax–Milgram theorem we conclude that Λ is an isomorphism.

4. Stabilizability: Proof of Theorem 1.3. We assume throughout this section that Ω, Γ_0, and Γ_1 satisfy the conditions of Theorem 1.3. The well-posedness of the problem (1.8) can be established by standard methods as, e.g., in [17], [18]; we omit the details. Setting

$$V = \{ v \in H_0^1(\Omega) \ : \ v = 0 \ \text{ on } \ \Gamma_0 \}$$

we have the following proposition.

PROPOSITION 4.1. *Assume* (1.1). *Then for every given* $\xi^0 \in V^n$ *and* $\xi^1 \in L^2(\Omega)^n$ *the problem* (1.8) *has a unique (weak) solution satisfying*

$$\xi \in C([0,+\infty); V^n) \cap C^1([0,+\infty); L^2(\Omega)^n).$$

Now assume also that $\xi^0 \in (H^2(\Omega) \cap V)^n$, $\xi^1 \in V^n$ *and* $\sigma_{ij}(\xi^0)\nu_j + A\xi_i^0 + B\xi_i^1 = 0$ *on* Γ_1, $i = 1, \ldots, n$. *Then the corresponding (strong) solution is more regular:*

$$\xi \in C([0,+\infty); (H^2(\Omega) \cap V)^n) \cap C^1([0,+\infty); V^n) \cap C^2([0,+\infty); L^2(\Omega)^n).$$

Let us turn to the proof of Theorem 1.3. All computations which follow will be justified for strong solutions. Since the constant ω in (1.11) will not depend on the choice of the initial data, once the estimates (1.11) will be established for regular solutions, they will also be satisfied for all weak solutions by an easy density argument. So in the rest of this section we shall only consider strong solutions.

Furthermore, we shall assume that $n \geq 2$. The case $n = 1$ is similar and even simpler, but we have to choose some constants differently. In any case, for $n = 1$, our problem reduces to the usual wave equation which has been studied extensively before; see, e.g., [15], [23], [11].

First we show the dissipativity of the problem (1.8).

LEMMA 4.2. *The energy of the strong solutions of* (1.8) *is a nonincreasing function of the time* $t \geq 0$. *More precisely, we have*

$$E(S) - E(T) = \int_S^T \int_{\Gamma_1} B\xi_i'\xi_i' \, d\Gamma \, dt, \quad 0 \leq S < T < +\infty. \tag{4.1}$$

Proof. We have

$$\begin{aligned} E' &= \int_\Omega \xi_i'\xi_i'' + \sigma_{ij}\varepsilon_{ij}' \, dx + \int_{\Gamma_1} A\xi_i\xi_i' \, d\Gamma \\ &= \int_\Omega \xi_i'\sigma_{ij,j} + \sigma_{ij}\xi_{i,j}' \, dx + \int_{\Gamma_1} A\xi_i\xi_i' \, d\Gamma \\ &= \int_\Gamma \xi_i'\sigma_{ij}\nu_j \, d\Gamma + \int_{\Gamma_1} A\xi_i\xi_i' \, d\Gamma = -\int_{\Gamma_1} B\xi_i'\xi_i' \, d\Gamma \leq 0; \end{aligned}$$

integrating between S and T the lemma follows. □

Now we are going to establish a basic identity for the solutions of (1.8). Given $0 \leq S < T < +\infty$ arbitrarily, we have

$$\begin{aligned} 0 &= \int_S^T \int_\Omega \xi_i(\xi_i'' - \sigma_{ij,j}) \, dx \, dt \\ &= \left[\int_\Omega \xi_i\xi_i' \, dx\right]_S^T - \int_S^T \int_\Gamma \xi_i\sigma_{ij}\nu_j \, d\Gamma \, dt + \int_S^T \int_\Omega \sigma_{ij}\varepsilon_{ij} - \xi_i'\xi_i' \, dx \, dt \end{aligned}$$

whence

$$\int_S^T \int_\Gamma \xi_i\sigma_{ij}\nu_j \, d\Gamma \, dt = \left[\int_\Omega \xi_i\xi_i' \, dx\right]_S^T + \int_S^T \int_\Omega \sigma_{ij}\varepsilon_{ij} - \xi_i'\xi_i' \, dx \, dt. \tag{4.2}$$

Note that the identity (2.1) of section 2 remains valid if we replace the lower integral limit 0 by S; just take the difference of this identity taken for T and for S instead of T. Apply this identity with $h(x) \equiv x - x_0$ and combine it with (4.2). Writing

$$M\xi_i = 2h_m\xi_{i,m} + (n-1)\xi_i$$

for brevity, we have

$$\int_S^T \int_\Gamma (M\xi_i)\sigma_{ij}\nu_j + (h\cdot\nu)(\xi_i'\xi_i' - \sigma_{ij}\varepsilon_{ij})\ d\Gamma\ dt \\ = \left[\int_\Omega (M\xi_i)\xi_i'\ dx\right]_S^T + \int_S^T \int_\Omega \sigma_{ij}\varepsilon_{ij} + \xi_i'\xi_i'\ dx\ dt.$$

Taking into account the definition (1.9) of the energy, we can rewrite it in the following form:

$$2\int_S^T E\ dt + \left[\int_\Omega (M\xi_i)\xi_i'\ dx\right]_S^T \\ = \int_S^T \int_\Gamma (M\xi_i)\sigma_{ij}\nu_j + (h\cdot\nu)(\xi_i'\xi_i' - \sigma_{ij}\varepsilon_{ij})\ d\Gamma\ dt \\ + \int_S^T \int_{\Gamma_1} A\xi_i\xi_i\ d\Gamma\ dt.$$

Now using the boundary conditions in (1.8) we obtain that

$$(4.3)\quad 2\int_S^T E\ dt + \left[\int_\Omega (M\xi_i)\xi_i'\ dx\right]_S^T \\ = \int_S^T \int_{\Gamma_0} (h\cdot\nu)\sigma_{ij}\varepsilon_{ij}\ d\Gamma\ dt \\ + \int_S^T \int_{\Gamma_1} A\xi_i\xi_i - (M\xi_i)(A\xi_i + B\xi_i') + (h\cdot\nu)(\xi_i'\xi_i' - \sigma_{ij}\varepsilon_{ij})\ d\Gamma\ dt.$$

(The term on Γ_0 is obtained in the same way as in Step 2 of the proof of Theorem 1.1.)

Next we transform the integral over Γ_1. Applying the Green formula twice and using the boundary condition on Γ_0 we have

$$\int_\Omega \xi_{m,i}\xi_{i,m}\ dx = \int_\Gamma \xi_{m,i}\xi_i\nu_m\ d\Gamma - \int_\Omega \xi_{m,im}\xi_i\ dx \\ = \int_\Gamma \xi_{m,i}\xi_i\nu_m - \xi_{m,m}\nu_i\xi_i\ d\Gamma + \int_\Omega \xi_{m,m}\xi_{i,i}\ dx \\ = \int_{\Gamma_1} 2\varepsilon_{mi}\xi_i\nu_m - \xi_{i,m}\xi_i\nu_m - \varepsilon_{mm}\nu_i\xi_i\ d\Gamma + \int_\Omega \varepsilon_{mm}\varepsilon_{ii}\ dx.$$

On the other hand,

$$\int_\Omega \xi_{m,i}\xi_{i,m}\ dx = \int_\Omega 2\varepsilon_{mi}\xi_{i,m} - \xi_{i,m}\xi_{i,m}\ dx = \int_\Omega 2\varepsilon_{mi}\varepsilon_{mi} - \xi_{i,m}\xi_{i,m}\ dx,$$

and therefore

$$\int_\Omega 2\varepsilon_{mi}\varepsilon_{mi} - \xi_{i,m}\xi_{i,m} - \varepsilon_{mm}\varepsilon_{ii}\ dx = \int_{\Gamma_1} 2\varepsilon_{mi}\xi_i\nu_m - \xi_{i,m}\xi_i\nu_m - \varepsilon_{mm}\nu_i\xi_i\ d\Gamma. \tag{4.4}$$

This is a crucial new identity. Since $h = R\nu$ on Γ_1, we deduce the following equality:

$$\begin{aligned}\int_S^T\int_{\Gamma_1} -2Ah_m\xi_{i,m}\xi_i\ d\Gamma\ dt &= \int_S^T\int_\Omega 4AR\varepsilon_{mi}\varepsilon_{mi} - 2AR\xi_{i,m}\xi_{i,m} - 2AR|\text{div }\xi|^2\ dx\ dt\\ &\quad + \int_S^T\int_{\Gamma_1} 2AR(\text{div }\xi)(\nu\cdot\xi) - 4AR\varepsilon_{mi}\xi_i\nu_m\ d\Gamma\ dt.\end{aligned} \tag{4.5}$$

Next we obtain by a similar computation that

$$\begin{aligned}\int_S^T\int_\Omega \xi_{m,i}\xi'_{i,m}\ dx\ dt &= \int_S^T\int_\Gamma \xi_{m,i}\xi'_i\nu_m\ d\Gamma\ dt - \int_S^T\int_\Omega \xi_{m,im}\xi'_i\ dx\ dt\\ &= \int_S^T\int_\Gamma \xi_{m,i}\xi'_i\nu_m - \xi_{m,m}\nu_i\xi'_i\ d\Gamma\ dt + \int_S^T\int_\Omega \xi_{m,m}\xi'_{i,i}\ dx\ dt\\ &= \int_S^T\int_{\Gamma_1} 2\varepsilon_{mi}\xi'_i\nu_m - \xi_{i,m}\xi'_i\nu_m - \varepsilon_{mm}\nu_i\xi'_i\ d\Gamma\ dt + \left[\frac{1}{2}\int_\Omega |\text{div }\xi|^2\ dx\right]_S^T.\end{aligned}$$

Furthermore,

$$\begin{aligned}\int_S^T\int_\Omega \xi_{m,i}\xi'_{i,m}\ dx\ dt = \int_S^T\int_\Omega 2\varepsilon_{mi}\varepsilon'_{mi} - \xi_{i,m}\xi'_{i,m}\ dx\ dt\\ = \left[\int_\Omega \varepsilon_{mi}\varepsilon_{mi} - \frac{1}{2}\xi_{i,m}\xi_{i,m}\ dx\right]_S^T.\end{aligned}$$

Using again the relation $h = R\nu$ on Γ_1, it follows that

$$\begin{aligned}\int_S^T\int_{\Gamma_1} -2Bh_m\xi_{i,m}\xi'_i\ d\Gamma\ dt &= \int_S^T\int_\Omega 2BR\xi_{m,i}\xi'_{i,m}\ dx\ dt - \left[\int_\Omega BR|\text{div }\xi|^2\ dx\right]_S^T\\ &\quad + \int_S^T\int_{\Gamma_1} 2BR\varepsilon_{mm}\nu_i\xi'_i - 4BR\varepsilon_{mi}\xi'_i\nu_m\ d\Gamma\ dt\\ &= \left[\int_\Omega 2BR\varepsilon_{mi}\varepsilon_{mi} - BR\xi_{i,m}\xi_{i,m} - BR|\text{div }\xi|^2\ dx\right]_S^T\\ &\quad + \int_S^T\int_{\Gamma_1} 2BR(\text{div }\xi)(\nu\cdot\xi') - 4BR\varepsilon_{mi}\xi'_i\nu_m\ d\Gamma\ dt.\end{aligned}$$

Substituting the equalities (4.5) and (4.6) into the identity (4.3) and using the equality $h \cdot \nu = R$ on Γ_1 we obtain that

$$
\begin{aligned}
(4.6)\quad & 2\int_S^T E\,dt \\
&= \left[\int_\Omega -(M\xi_i)\xi_i' + 2BR\varepsilon_{mi}\varepsilon_{mi} - BR\xi_{i,m}\xi_{i,m} - BR|\mathrm{div}\ \xi|^2\,dx\right]_S^T \\
&\quad + \int_S^T\int_\Omega 4AR\varepsilon_{mi}\varepsilon_{mi} - 2AR\xi_{i,m}\xi_{i,m} - 2AR|\mathrm{div}\ \xi|^2\,dx\,dt \\
&\quad + \int_S^T\int_{\Gamma_0}(h\cdot\nu)\sigma_{ij}\varepsilon_{ij}\,d\Gamma\,dt \\
&\quad + \int_S^T\int_{\Gamma_1} A\xi_i\xi_i - (n-1)\xi_i(A\xi_i + B\xi_i') + R(\xi_i'\xi_i' - \sigma_{ij}\varepsilon_{ij}) + 2AR(\mathrm{div}\ \xi)(\nu\cdot\xi) \\
&\qquad - 4AR\varepsilon_{mi}\xi_i\nu_m + 2BR(\mathrm{div}\ \xi)(\nu\cdot\xi') - 4BR\varepsilon_{mi}\xi_i'\nu_m\,d\Gamma\,dt.
\end{aligned}
$$

Let us rewrite it in the following form:

$$
\begin{aligned}
(4.7)\quad & 2\int_S^T E\,dt \\
&= \left[\int_\Omega -(M\xi_i)\xi_i' + 2BR\varepsilon_{mi}\varepsilon_{mi} - BR\xi_{i,m}\xi_{i,m} - BR|\mathrm{div}\ \xi|^2\,dx\right]_S^T \\
&\quad + \left[\frac{1}{2}\int_{\Gamma_1}(1-n)B\xi_i\xi_i\,d\Gamma\right]_S^T \\
&\quad + \int_S^T\int_\Omega 4AR\varepsilon_{mi}\varepsilon_{mi} - 2AR\xi_{i,m}\xi_{i,m} - 2AR|\mathrm{div}\ \xi|^2\,dx\,dt \\
&\quad + \int_S^T\int_{\Gamma_0}(h\cdot\nu)\sigma_{ij}\varepsilon_{ij}\,d\Gamma\,dt \\
&\quad + \int_S^T\int_{\Gamma_1}(2-n)A\xi_i\xi_i + R(\xi_i'\xi_i' - \sigma_{ij}\varepsilon_{ij}) + 2AR(\mathrm{div}\ \xi)(\nu\cdot\xi) \\
&\qquad - 4AR\varepsilon_{mi}\xi_i\nu_m + 2BR(\mathrm{div}\ \xi)(\nu\cdot\xi') - 4BR\varepsilon_{mi}\xi_i'\nu_m\,d\Gamma\,dt.
\end{aligned}
$$

This is our main identity.

Let us majorize the right-hand side of this identity. First of all, using the definition of the energy and the Korn inequality,

$$\left|\int_\Omega -(M\xi_i)\xi_i' + 2BR\varepsilon_{mi}\varepsilon_{mi} - BR\xi_{i,m}\xi_{i,m} - BR|\mathrm{div}\ \xi|^2\,dx\right| \le c_1E$$

with some constant c_1, depending on A, B but independent of S, T and of the particular choice of the initial data in (1.8). We have similarly

$$\left|\frac{1}{2}\int_{\Gamma_1}(1-n)B\xi_i\xi_i\,d\Gamma\right| \le c_2E.$$

For the estimate of the third term on the right-hand side note that

$$\int_\Omega 4AR\varepsilon_{mi}\varepsilon_{mi} - 2AR\xi_{i,m}\xi_{i,m} - 2AR|\text{div }\xi|^2\ dx \le 8AR\alpha^{-1}E$$

with α defined by the ellipticity condition (1.1):

$$\sigma_{ij}\varepsilon_{ij} \ge \alpha\varepsilon_{ij}\varepsilon_{ij}.$$

Since $h\cdot\nu = -r \le 0$ on Γ_0, the integral on Γ_0 is ≤ 0. In the integral over Γ_1 we have $(2-n)A\xi_i\xi_i \le 0$ because $n \ge 2$, and we have

$$-R\sigma_{ij}\varepsilon_{ij} \le -R\alpha\varepsilon_{ij}\varepsilon_{ij}$$

by using (1.1) again. Finally, applying Lemma 4.2 we deduce from the identity (4.7) the following inequality:

$$\begin{aligned}(4.8)\quad 2\int_S^T E\ dt \le c_3E(S) + 8AR\alpha^{-1}\int_S^T E\ dt\\ + \int_S^T\int_{\Gamma_1} R\xi_i'\xi_i' - R\alpha\varepsilon_{ij}\varepsilon_{ij} + 2AR(\text{div }\xi)(\nu\cdot\xi)\\ - 4AR\varepsilon_{mi}\xi_i\nu_m + 2BR(\text{div }\xi)(\nu\cdot\xi') - 4BR\varepsilon_{mi}\xi_i'\nu_m\ d\Gamma\ dt.\end{aligned}$$

Here $c_3 = 2c_1 + 2c_2$ is a constant depending on A and B but not on the choice of the initial data.

From now on we shall distinguish three cases. Consider first the case $A = 0$; then $\Gamma_0 \ne \emptyset$ by assumption. For any fixed $\delta > 0$ we have

$$2BR(\text{div }\xi)(\nu\cdot\xi') \le \delta|\text{div }\xi|^2 + B^2R^2\delta^{-1}|\xi'|^2,$$

$$-4BR\varepsilon_{mi}\xi_i'\nu_m \le \delta\varepsilon_{mi}\varepsilon_{mi} + 4B^2R^2\delta^{-1}|\xi'|^2.$$

Using these inequalities and the obvious inequality $|\text{div }\xi|^2 \le \varepsilon_{mi}\varepsilon_{mi}$, we deduce from (4.8) that

$$2\int_S^T E\ dt \le c_3E(S) + \int_S^T\int_{\Gamma_1}(R + 5B^2R^2\delta^{-1})|\xi'|^2 + (2\delta - R\alpha)\varepsilon_{ij}\varepsilon_{ij}\ d\Gamma\ dt.$$

Choosing $\delta = R\alpha/2$ and using (4.1), it follows that

$$2\int_S^T E\ dt \le c_4E(S)$$

with

$$c_4 = c_3 + RB^{-1} + 10BR\alpha^{-1}.$$

Hence there exists a constant $\omega > 0$, independent of the choice of the initial data, such that

$$(4.9)\qquad \omega\int_S^T E\ dt \le E(S)$$

for all $0 \leq S < T < +\infty$. Since the function $E(t)$ is nonnegative and nonincreasing and applying a well-known Gronwall-type inequality (see, e.g., [12, Theorem 8.1]), hence we conclude that the inequality (1.11) is satisfied.

Now let us consider the case where $0 < A < \alpha/(24R)$. Applying the Young inequality, for any fixed $\delta > 0$ we have

$$2AR(\text{div }\xi)(\nu \cdot \xi) \leq A\delta|\xi|^2 + AR^2\delta^{-1}|\text{div }\xi|^2,$$

$$2BR(\text{div }\xi)(\nu \cdot \xi') \leq B^2A^{-1}\delta|\xi'|^2 + AR^2\delta^{-1}|\text{div }\xi|^2,$$

$$-4AR\varepsilon_{mi}\xi_i\nu_m \leq A\delta|\xi|^2 + 4AR^2\delta^{-1}\varepsilon_{mi}\varepsilon_{mi},$$

$$-4BR\varepsilon_{mi}\xi_i'\nu_m \leq B^2A^{-1}\delta|\xi'|^2 + 4AR^2\delta^{-1}\varepsilon_{mi}\varepsilon_{mi}.$$

Substituting them into (4.8) and using the inequality $|\text{div }\xi|^2 \leq \varepsilon_{mi}\varepsilon_{mi}$ again, we obtain that

$$\begin{aligned}2\int_S^T E\ dt &\leq c_3E(S) + 8AR\alpha_1^{-1}\int_S^T E\ dt \\ &+ \int_S^T\int_{\Gamma_1}(R + 2B^2A^{-1}\delta)|\xi'|^2 + 2A\delta|\xi|^2 + (10AR^2\delta^{-1} - R\alpha_1)\varepsilon_{ij}\varepsilon_{ij}\ d\Gamma\ dt.\end{aligned}$$

Using (4.1) we have

$$\begin{aligned}&\int_S^T\int_{\Gamma_1}(R + 2B^2A^{-1}\delta)|\xi'|^2\ d\Gamma\ dt \\ &\qquad = (RB^{-1} + 2BA^{-1}\delta)(E(S) - E(T)) \leq (RB^{-1} + 2BA^{-1}\delta)E(S).\end{aligned}$$

Substituting into the preceding inequality and choosing $\delta = 10AR/\alpha$, we conclude that

$$2\int_S^T E\ dt \leq c_4E(S) + 8AR\alpha^{-1}\int_S^T E\ dt + \int_S^T\int_{\Gamma_1}20A^2R\alpha^{-1}|\xi|^2\ d\Gamma\ dt \tag{4.10}$$

with

$$c_4 = c_3 + RB^{-1} + 20BR\alpha^{-1}.$$

Now observe that

$$\int_S^T\int_{\Gamma_1}20A^2R\alpha^{-1}|\xi|^2\ d\Gamma\ dt \leq 40AR\alpha^{-1}\int_S^T E\ dt$$

by the definition of the energy; therefore we conclude from (4.10) the estimate (4.9) with

$$\omega = (2 - 48AR\alpha^{-1})/c_4.$$

If $0 < A < \alpha/(24R)$, then $\omega > 0$ and the estimate (1.11) follows as in the first case.

Now let us turn to the case $\alpha/(24R) \leq A < \alpha/(4R)$. In this case we have to estimate the last integral of (4.10) in a different way. Applying a method of [5] we shall prove the following lemma.

LEMMA 4.3. *There exists a constant* $c_5 > 0$ *such that*

$$\int_S^T \int_{\Gamma_1} |\xi|^2 \, d\Gamma \, dt \leq c_5 \varepsilon^{-1} E(S) + \varepsilon \int_S^T E \, dt$$

for all $0 < \varepsilon < 1$ *and for all* $0 \leq S < T < +\infty$.

Assuming this lemma for the moment, we deduce from (4.10) the inequality

$$(2 - 8AR\alpha^{-1} - \varepsilon) \int_S^T E \, dt \leq (c_4 + 20A^2 R\alpha^{-1} c_5 \varepsilon^{-1}) E(S).$$

Choosing $\varepsilon < 2 - 8AR\alpha^{-1}$, the integral inequality (4.9) follows again and the proof can be completed as before.

It remains to prove the lemma. For every $t \geq 0$ let us denote by $\eta(t)$ the solution of the problem

$$\begin{cases} -\sigma_{ij,j} = 0 & \text{in} \quad \Omega, \\ \eta = \xi & \text{on} \quad \Gamma. \end{cases}$$

Then we have

$$\|\eta\|_{L^2(\Omega)^n} \leq c\|\xi\|_{L^2(\Gamma)^n} \leq c\sqrt{E} \tag{4.11}$$

by the elliptic regularity theory. (See, e.g., [4] or [6] for the proof of the ellipticity of the above system.) Applying this result with ξ' instead of ξ we also obtain that

$$\|\eta'\|_{L^2(\Omega)^n} \leq c\|\xi'\|_{L^2(\Gamma)^n} \leq c\sqrt{|E'|}. \tag{4.12}$$

Let us also observe that

$$\begin{aligned} &\int_\Omega \sigma_{ij}(\eta)\varepsilon_{ij}(\xi - \eta) \, dx \\ &\qquad = -\int_\Omega \sigma_{ij,j}(\eta)(\xi_i - \eta_i) \, dx + \int_\Gamma \sigma_{ij}(\eta)\nu_j(\xi_i - \eta_i) \, d\Gamma = 0; \end{aligned} \tag{4.13}$$

hence

$$\int_\Omega \sigma_{ij}(\eta)\varepsilon_{ij}(\xi) \, dx = \int_\Omega \sigma_{ij}(\eta)\varepsilon_{ij}(\eta) \, dx \geq 0. \tag{4.14}$$

Now consider the following equality:

$$\begin{aligned} 0 &= \int_S^T \int_\Omega \eta_i(\xi_i'' - \sigma_{ij,j}(\xi)) \, dx \, dt \\ &= \left[\int_\Omega \eta_i \xi_i' \, dx\right]_S^T + \int_S^T \int_\Omega -\eta_i'\xi_i' + \sigma_{ij}(\xi)\varepsilon_{ij}(\eta) \, dx \, dt - \int_S^T \int_\Gamma \eta_i \sigma_{ij}(\xi)\nu_j \, d\Gamma \, dt \\ &= \left[\int_\Omega \eta_i \xi_i' \, dx\right]_S^T + \int_S^T \int_\Omega -\eta_i'\xi_i' + \sigma_{ij}(\xi)\varepsilon_{ij}(\eta) \, dx \, dt + \int_S^T \int_\Gamma \xi_i(A\xi_i + B\xi_i') \, d\Gamma \, dt. \end{aligned}$$

(In the last step we used the boundary conditions on ξ and on η.) Using (4.13) we deduce the following inequality:

$$\begin{aligned}(4.15)\quad A\int_S^T\int_\Gamma|\xi|^2 d\Gamma\ dt \\ \le \left|\left[\int_\Omega \eta_i\xi_i'\ dx\right]\right|_S^T + \int_S^T\int_\Omega \eta_i'\xi_i'\ dx\ dt - B\int_S^T\int_\Gamma \xi_i\xi_i'\ d\Gamma\ dt.\end{aligned}$$

Using (4.11), (4.12) and applying the Poincaré inequality we can estimate the right-hand side of this inequality as follows:

$$\left|\left[\int_\Omega \eta_i\xi_i'\ dx\right]\right| \le \|\eta\|_{L^2(\Omega)^n}\|\xi'\|_{L^2(\Omega)^n} \le cE;$$

$$\begin{aligned}\int_S^T\int_\Omega \eta_i'\xi_i'\ dx\ dt &\le \int_S^T \|\eta'\|_{L^2(\Omega)^n}\|\xi'\|_{L^2(\Omega)^n}\ dt \\ &\le c\int_S^T \sqrt{|E'|}\sqrt{E}\ dt \le \int_S^T (A\varepsilon/2)E + (c/\varepsilon)|E'|\ dt \\ &= \int_S^T (A\varepsilon/2)E\ dt + (c/\varepsilon)E(S) - (c/\varepsilon)E(T) \\ &\le \int_S^T (A\varepsilon/2)E\ dt + (c/\varepsilon)E(S),\end{aligned}$$

and finally

$$\begin{aligned}-B\int_S^T\int_\Gamma \xi_i\xi_i'\ d\Gamma\ dt &\le \int_S^T \|\xi\|_{L^2(\Gamma)^n}\|\xi'\|_{L^2(\Gamma)^n}\ dt \\ &\le c\int_S^T \sqrt{E}\sqrt{|E'|}\ dt \le \int_S^T (A\varepsilon/2)E + (c/\varepsilon)|E'|\ dt \\ &\le \int_S^T (A\varepsilon/2)E\ dt + (c/\varepsilon)E(S).\end{aligned}$$

Substituting these inequalities into (4.15) the lemma follows.

We end this section by showing that Theorem 1.3 can be extended to cases where the outer boundary of Ω is not exactly (but is close to) a sphere. Fix n numbers $\beta_1, \beta_2, \dots, \beta_n \in (0,1)$, and set

$$\Omega = \left\{x \in \mathbb{R}^n\ :\ 1 < \sum_{i=1}^n \beta_i x_i^2 < 4\right\},$$

$$\Gamma_0 = \left\{x \in \mathbb{R}^n\ :\ \sum_{i=1}^n \beta_i x_i^2 = 1\right\},$$

$$\Gamma_1 = \left\{x \in \mathbb{R}^n\ :\ \sum_{i=1}^n \beta_i x_i^2 = 4\right\}.$$

Theorem 4.4. *Assume* (1.1) *and fix two positive constants A and B with $A < \alpha/48$. If $\beta_1, \beta_2, \dots, \beta_n$ are sufficiently close to* 1*, then there exists a positive number ω such that all (weak) solutions of* (1.8) *satisfy the energy estimate*

$$E(t) \le E(0)e^{1-\omega t}$$

for all $t \ge 0$.

The idea of the proof is to modify the *multiplier* so as to preserve the property $h = R\nu$ on Γ_1. The price to pay will be to have some extra terms in the modified identity (4.7). Fortunately these extra terms can be easily estimated by applying the Korn inequality.

More precisely, let us define a function $h : \overline{\Omega} \to \mathbb{R}^n$ by the formula

$$h(x) = \left(\frac{\beta_1 x_1^2 + \cdots + \beta_n x_n^2}{\beta_1^2 x_1^2 + \cdots + \beta_n^2 x_n^2} \right)^{1/2} (\beta_1 x_1, \dots, \beta_n x_n).$$

Then h is of class C^∞ and $h = -\nu$ on Γ_0, $h = 2\nu$ on Γ_1.

Applying the identity (2.1) with this h and repeating the proof of Theorem 1.3 we obtain instead of (4.3) the identity

$$\begin{aligned} 2\int_S^T E\,dt + \left[\int_\Omega (M\xi_i)\xi_i'\,dx \right]_S^T &= \int_S^T \int_{\Gamma_0} (h\cdot\nu)\sigma_{ij}\varepsilon_{ij}\,d\Gamma\,dt \\ &+ \int_S^T \int_{\Gamma_1} A\xi_i\xi_i - (M\xi_i)(A\xi_i + B\xi_i') + (h\cdot\nu)(\xi_i'\xi_i' - \sigma_{ij}\varepsilon_{ij})\,d\Gamma\,dt \\ &- \int_S^T \int_\Omega 2f_{m,j}\sigma_{ij}\xi_{i,m} + (\mathrm{d}iv\ f)(\xi_i'\xi_i' - \sigma_{ij}\varepsilon_{ij})\,dx\,dt, \end{aligned}$$

where f is defined by the formula $f(x) = h(x) - x$.

Leaving the last two new terms unchanged and transforming the boundary integrals as before (now $R = 2$), we obtain instead of (4.10) the identity

$$\begin{aligned} 2\int_S^T E\,dt \le c_4E(S) + 16A\alpha^{-1}\int_S^T E\,dt + \int_S^T \int_{\Gamma_1} 40A^2\alpha^{-1}|\xi|^2\,d\Gamma\,dt \\ - \int_S^T \int_\Omega 2f_{m,j}\sigma_{ij}\xi_{i,m} + (\mathrm{d}iv\ f)(\xi_i'\xi_i' - \sigma_{ij}\varepsilon_{ij})\,dx\,dt. \end{aligned}$$

Majorizing the boundary integral by using the definition of the energy as before, we deduce the inequality

$$\begin{aligned} 2\int_S^T E\,dt \le c_4E(S) + 96A\alpha^{-1}\int_S^T E\,dt \\ - \int_S^T \int_\Omega 2f_{m,j}\sigma_{ij}\xi_{i,m} + (\mathrm{d}iv\ f)(\xi_i'\xi_i' - \sigma_{ij}\varepsilon_{ij})\,dx\,dt. \end{aligned}$$

It remains to show that for any given $\varepsilon > 0$ the last integral can be majorized by

$$\varepsilon \int_S^T E\,dt$$

if $\beta_1, \beta_2, \dots, \beta_n$ are sufficiently close to 1.

First we note that

$$\left| \int_S^T \int_\Omega 2f_{m,j}\sigma_{ij}\xi_{i,m} + (\mathrm{d}iv\ f)(\xi_i'\xi_i' - \sigma_{ij}\varepsilon_{ij})\ dx\ dt \right| \\ \leq c \sum_{i,j=1}^n \|f_{i,j}\|_{L^\infty(\Omega)} \int_S^T E\ dt$$

with a constant $c = c(\beta_1, \beta_2, \dots, \beta_n)$ coming from the Korn inequality. Looking at the proof of this inequality in [7] we see that c remains uniformly bounded as β_1, β_2, … , β_n all tend to 1. (The constants appear at the moment of using a partition of unity and the approximated domain, a ball, is very smooth.) It remains to show that the functions $f_{i,j}$ converge to 0 uniformly on $\overline{\Omega}$ as β_1, β_2, … , β_n all tend to 1. Now a direct easy computation shows that $h_{i,j}$ converges to $\delta_{i,j}$ uniformly on $\overline{\Omega}$, and hence the claim follows. (At this point we use the fact that Ω has a hole inside, in order to remove a possible singularity of h at the origin.)

5. Stabilizability: Proof of Theorem 1.4. Let us first consider the homogeneous problem (1.2) of Theorem 1.1 with the boundary observation $\psi_i = \sigma_{ij,j}(\xi)\nu_j|_{\Gamma_1}$:

$$\begin{cases} \xi_i'' - \sigma_{ij,j} = 0 \quad \text{in} \quad \Omega \times \mathbb{R}, \\ \xi_i = 0 \quad \text{on} \quad \Gamma \times \mathbb{R}, \\ \xi_i(0) = \xi_i^0 \quad \text{and} \quad \xi_i'(0) = \xi_i^1 \quad \text{in} \quad \Omega, \\ \psi_i = \sigma_{ij,j}(\xi)\nu_j|_{\Gamma_1} \quad \text{in} \quad \mathbb{R}, \\ i = 1, \dots, n. \end{cases} \tag{5.1}$$

Putting $\varphi = (\xi, \xi')$, $\varphi_0 = (\xi^0, \xi^1)$ and introducing two linear operators A^* and B^* by the formulas

$$D(A^*) = D(B^*) = (H^2(\Omega) \cap H_0^1(\Omega))^n \times H_0^1(\Omega)^n, \\ A^*(\eta^0, \eta^1) = -(\eta^1, \sigma_{ij,j}(\eta^0)), \\ B^*(\eta^0, \eta^1) = \sigma_{ij,j}(\xi)\nu_j|_{\Gamma_1},$$

we may rewrite (5.1) in the following operational form:

$$\varphi' = -A^*\varphi, \quad \varphi(0) = \varphi_0, \quad \psi = B^*\varphi. \tag{5.2}$$

Let us introduce the Hilbert spaces $H = H^{-1}(\Omega)^n \times (L^2(\Omega)^n)$ and $G = L^2(\Gamma_1)^n$. Identifying $L^2(\Omega)$ and $L^2(\Gamma_1)$ with their duals, we have $H' = H_0^1(\Omega)^n \times L^2(\Omega)^n$ and $G' = L^2(\Gamma_1)^n$.

We claim that the following four conditions are satisfied:

(H1) The operator A^* generates a *group* e^{sA^*} in H'.

(H2) We have $D(A^*) = D(B^*)$ and there exist two numbers $\lambda \in \mathbb{C}$ and $c \in \mathbb{R}$ such that

$$\|B^*\phi\| \leq c\|(A + \lambda I)^*\phi\|$$

for all $\phi \in D(A^*)$.

(H3) There exist two positive numbers T' and c' such that

$$\|\psi\|_{L^2(0,T';G')} \leq c'\|\varphi_0\|_{H'}$$

for all $\varphi_0 \in D(A^*)$.

(H4) There exist two positive numbers T and c such that

$$\|\psi\|_{L^2(0,T;G')} \geq c\|\varphi_0\|_{H'}$$

for all $\varphi_0 \in D(A^*)$.

Indeed, property (H1) is satisfied by the well-posedness and time-reversibility of the problem (1.2) studied in section 1. Property (H2) (with $\lambda = 0$) follows from the elliptic regularity of the system of elasticity; see, e.g., [4] or [6]. The properties (H3) and (H4) are equivalent to the two inequalities (1.6) in Theorem 1.1; we can choose any $T', T > 2\sqrt{2/\alpha}R$.

We may now apply the following theorem proved in [14].

THEOREM 5.1. *Assume* (H1) *to* (H4) *for some* $T > 0$. *Then for every fixed* $\omega > 0$ *there exists an isomorphism* Λ_ω *of* H' *onto* H *such that putting* $F = -B^*\Lambda_\omega^{-1}$ *the operator* $A + BF$ *generates (in some sense) a strongly continuous* group *in* H, *and there exists a constant* M *such that the solutions of the closed-loop problem*

$$x' = Ax + BFx, \qquad x(0) = x_0 \tag{5.3}$$

satisfy the estimate

$$\|x(t)\|_H \leq M\|x_0\|_H e^{-\omega t}, \qquad \textit{for all } t > 0,$$

for all $x_0 \in H$, *and for all* $t \geq 0$.

Let us remark that the operator Λ_ω is constructed explicitly and that the constant M can also be estimated explicitly. See [14] for details.

Let us write the problem (5.3) in a more explicit form. First of all, it follows from the proof of Theorem 3.1 in section 3 that the open-loop problem

$$x' = Ax + Bu, \qquad x(0) = x_0,$$

where $x = (-y', y)$ and $x_0 = (-y^1, y^0)$, can be written in the following form:

$$\begin{cases} y_i'' - \sigma_{ij,j}(y) = 0 \quad \text{in} \quad \Omega \times \mathbb{R}, \\ y_i = 0 \quad \text{on} \quad \Gamma_0 \times \mathbb{R}, \\ y_i = u_i \quad \text{on} \quad \Gamma_1 \times \mathbb{R}, \\ y_i(0) = y_i^0 \quad \text{and} \quad y_i'(0) = y_i^1 \quad \text{in} \quad \Omega, \\ i = 1, \ldots, n. \end{cases}$$

Therefore it only remains to show that $u = Fx$ can be written as

$$u_i = \sigma_{ij}(Py' + Qy)\nu_j \quad \text{on} \quad \Gamma_1 \times \mathbb{R}$$

with suitable bounded linear maps P, Q as given in Theorem 1.4. This follows easily from the properties of Λ_ω and from the definition of B^*. Indeed, writing the operator

$$\Lambda_\omega^{-1} : H^{-1}(\Omega)^n \times L^2(\Omega)^n \to H_0^1(\Omega)^n \times L^2(\Omega)^n$$

in the matrix form

$$\Lambda_\omega^{-1} = \begin{pmatrix} P & -Q \\ -R & S \end{pmatrix},$$

we have

$$u = -B^* \Lambda_\omega^{-1} x = \sigma_{ij}(Py' + Qy)$$

on Γ_1, and the proof of Theorem 1.4 is completed.

REFERENCES

[1] F. Alabau and V. Komornik, *Observabilité, contrôlabilité et stabilisation frontière du système d'élasticité linéaire*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 519–524.

[2] F. Bourquin, *Stabilisation rapide selon Komornik et méthode de Faedo–Galerkin*, in preparation.

[3] F. Bourquin and J.-S. Briffaut, *Numerical implementation of the feedback of Komornik*, to appear.

[4] P. G. Ciarlet, *Mathematical Elasticity, Vol.* I: *Three-Dimensional Elasticity*, North-Holland, Amsterdam, 1988.

[5] F. Conrad and B. Rao, *Decay of solutions of wave equations in a star-shaped domain with nonlinear boundary feedback*, Asymptotic Anal., 7 (1993), pp. 159–177.

[6] R. Dautray and J.-L. Lions, *Analyse mathématique et calcul scientifique pour les sciences et les techniques*, Tome 1, Chapitre VII, Masson, Paris, 1988.

[7] G. Duvaut and J.-L. Lions, *Les inéquations en mécanique et en physique*, Dunod, Paris, 1972.

[8] M. A. Horn, *Implications of sharp trace regularity results on boundary stabilization of the system of linear elasticity*, J. Math. Anal. Appl., to appear.

[9] H. Koch and D. Tataru, *On the spectrum of hyperbolic semigroups*, Comm. Partial Differential Equations, 20 (1995), pp. 901–937.

[10] V. Komornik, *Contrôlabilité exacte en un temps minimal*, C. R. Acad. Sci. Paris Sér I Math., 304 (1987), pp. 223–225.

[11] V. Komornik, *Rapid boundary stabilization of the wave equation*, SIAM J. Control Optim., 29 (1991), pp. 197–208.

[12] V. Komornik, *Exact Controllability and Stabilization: The Multiplier Method,* Masson, Paris, and John Wiley & Sons, Chicester, 1994.

[13] V. Komornik, *Boundary stabilization of linear elasticity systems*, Lecture Notes in Pure and Appl. Math. 174, Dekker, New York, 1995, pp. 135–146.

[14] V. Komornik, *Rapid boundary stabilization of linear distributed systems*, SIAM J. Control Optim., 35 (1997), pp. 1591–1613.

[15] V. Komornik and E. Zuazua, *A direct method for the boundary stabilization of the wave equation*, J. Math. Pures Appl., 69 (1990), pp. 33–54.

[16] J. E. Lagnese, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163–182.

[17] J. E. Lagnese, *Boundary stabilization of linear elastodynamic systems*, SIAM J. Control Optim., 21 (1983), pp. 968–984.

[18] J. E. Lagnese, *Uniform asymptotic energy estimates for solutions of the equations of dynamic plane elasticity with nonlinear dissipation at the boundary*, Nonlinear Anal., 16 (1991), pp. 35–54.

[19] I. Lasiecka and R. Triggiani, *Uniform stabilization of the wave equation with Dirichlet or Neumann feedback control without geometric conditions*, Appl. Math. Optim., 25 (1992), pp. 189–224.

[20] J.-L. Lions, *Exact controllability, stabilizability, and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[21] J.-L. Lions, *Contrôlabilité exacte et stabilisation de systèmes distribués*, Vol. 1, Masson, Paris, 1988.

[22] P. Martinez, *Uniform boundary stabilization of elasticity systems of cubic crystals by nonlinear feedbacks*, Nonlinear Anal., to appear.

[23] R. Triggiani, *Wave equation on a bounded domain with boundary dissipation: An operator approach*, J. Math. Anal. Appl., 137 (1989), pp. 438–461.

SIAM J. Control Optim.
Vol. 37, No. 2, pp. 543–565

EXACT BOUNDARY CONTROLLABILITY OF THE KORTEWEG–DE VRIES EQUATION*

BING-YU ZHANG†

Abstract. We consider boundary control of the distributed parameter system described by the Korteweg–de Vries (KdV) equation posed on a finite interval $\alpha \leq x \leq \beta$:

$$\begin{cases} u_t + u_x + uu_x + u_{xxx} = 0 \\ u(\alpha,t) = h_1(t), \qquad u(\beta,t) = h_2(t), \qquad u_x(\beta,t) = h_3(t)u \end{cases} \tag{$*$}$$

for $t \geq 0$. It is shown that by choosing appropriate control inputs $(h_j(t),\ j = 1,2,3)$, one can always guide the system $(*)$ from a given initial state $\phi \in H^s(\alpha,\beta)$ to a given terminal state $\psi \in H^s(\alpha,\beta)$ in the time period $[0,T]$ so long as ϕ and ψ satisfy

$$\|\phi(\cdot) - w(\cdot,0)\|_{H^s(\alpha,\beta)} \leq \delta \quad \text{and} \quad \|\psi(\cdot) - w(\cdot,T)\|_{H^s(\alpha,\beta)} \leq \delta$$

for some $\delta > 0$ independent of ϕ and ψ, where $s \geq 0$ and $w \equiv w(x,t)$ is a given smooth solution of the KdV equation. This exact boundary controllability is established by considering a related initial value control problem of the KdV equation posed on the whole line R. Various recently discovered smoothing properties of the KdV equation have played important roles in our approach.

Key words. boundary control, exact controllability, the KdV equation, smoothing properties

AMS subject classifications. 35K60, 93C20

PII. S0363012997327501

1. Introduction.

In this paper we continue our earlier work [27] to study control of the system described by the Korteweg–de Vries (KdV) equation

$$u_t + uu_x + u_{xxx} = 0, \tag{1.1}$$

where $u \equiv u(x,t)$ is a real-valued function of the two real variables x and t and the subscripts denote the corresponding partial derivatives. The equation was first derived by Korteweg and de Vries in 1895 as a model for propagation of some surface water waves along a channel [16]. Its original form is

$$\eta_t = \frac{3}{2}\sqrt{\frac{g}{l}}\left(\frac{1}{2}\eta^2 + \frac{2}{3}\alpha\eta + \frac{1}{3}\sigma\eta_{xx}\right)_x, \tag{1.2}$$

where η is the surface elevation above the equilibrium level l, α is a small constant related to the uniform motion of the liquid, g is the gravitational constant, and $\sigma = l^3/3 - Tl/\rho g$ with surface capillary tension T and density ρ. When posed on the whole line R or on a periodic domain, (1.2) can always be reduced by certain variable transformations to its standard form (1.1). The KdV equation has been intensively studied from various aspects of both mathematics and physics since the 1960s, [2], [7], [8], [28], [29] when *solitons* were discovered through solving the KdV equation and the *inverse scattering method*, a so-called nonlinear Fourier transform, was invented to seek solitons [1], [9], [6]. It turns out that the equation is not only a good model

*Received by the editors September 19, 1997; accepted for publication (in revised form) April 1, 1998; published electronically January 28, 1998.

http://www.siam.org/journals/sicon/37-2/32750.html

†Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221-0025 (bzhang@math.uc.edu).

for some water waves but also a very useful approximation model in nonlinear studies whenever one wishes to include and balance a weak nonlinearity and weak dispersive effects [19]. In particular, the equation is now commonly accepted as a mathematical model for the unidirectional propagation of small-amplitude long waves in nonlinear dispersive systems. In such applications, u is typically an amplitude or velocity, x is often proportional to distance in the direction of propagation, and t is proportional to elapsed time. In [27], we considered distributed control of the KdV equation

$$u_t + uu_x + u_{xxx} = f \tag{1.3}$$

on the domain $0 \leq x \leq 2\pi$, $t \geq 0$ with periodic conditions

$$u(0,t) = u(2\pi,t), \quad u_x(0,t) = u_x(2\pi,t), \quad u_{xx}(0,t) = u_{xx}(2\pi,t). \tag{1.4}$$

The distributed control $f \equiv f(x,t)$ is restricted so that the "volume" $\int_0^{2\pi} u(x,t)dx$ of the solution is conserved. If f is allowed to act on the whole spatial domain $(0,2\pi)$, the system is shown to be *globally* exactly controllable, i.e., for given $T > 0$ and functions $\phi(x)$, $\psi(x)$ with the same "volume," one can always find a control f so that the system (1.3)–(1.4) has a solution $u(x,t)$ satisfying

$$u(x,0) = \phi(x) \quad \text{and} \quad u(x,T) = \psi(x).$$

If the control f is allowed to act on only a small subset of the domain $(0,2\pi)$, then the system is *locally* exactly controllable in the sense that the initial and terminal states, ϕ and ψ, are required to have small "amplitude" in a certain sense (cf. [27] for the details).

In this paper we consider boundary control of the KdV equation

$$u_t + u_x + uu_x + u_{xxx} = 0 \tag{1.5}$$

on the domain (α,β), $t \geq 0$ with the boundary conditions

$$u(\alpha,t) = h_1(t), \qquad u(\beta,t) = h_2(t), \qquad u_x(\beta,t) = h_3(t). \tag{1.6}$$

The boundary value functions $h_j(t)$, $j = 1,2,3$, are considered as control inputs. Note that there is an extra term u_x in (1.5) as compared with the standard KdV equation (1.1). This occurs because, when posed on a finite interval, one cannot reduce the original KdV equation (1.2) to its standard form; the term u_x does not go away. On the other hand, the existence of the term u_x does make some differences for the equation when posed on a finite interval.[1]

Our main concern for the system (1.5)–(1.6) is the following exact control problem.

[1] For instance, while the boundary value problem

$$\begin{cases} u_t + u_{xxx} = 0, \quad x \in (0,2\pi),\ t > 0, \\ u(0,t) = u(2\pi,t) = u_x(2\pi,t) = 0 \end{cases}$$

has no nontrivial steady solutions, the problem

$$\begin{cases} u_t + u_x + u_{xxx} = 0, \quad x \in (0,2\pi),\ t > 0, \\ u(0,t) = u(2\pi,t) = u_x(2\pi,t) = 0 \end{cases}$$

does have some nontrivial steady solutions.

Let $T > 0$ and $s \geq 0$ be given. For any $\phi \in H^s(\alpha,\beta)$ and $\psi \in H^s(\alpha,\beta)$, can one find appropriate boundary control inputs h_j, $j = 1,2,3$, such that the system (1.5)–(1.6) *has a solution $u \in C([0,T]; H^s(\alpha,\beta))$ satisfying*

$$u(x,0) = \phi(x) \quad \textit{and} \quad u(x,T) = \psi(x)$$

on the interval (α,β)?

Here, by saying that $f(x) = g(x)$ on the interval (α,β), we mean that

$$\int_\alpha^\beta f(x)\xi(x)dx = \int_\alpha^\beta g(x)\xi(x)dx \qquad \text{for any } \xi \in C_0^\infty(\alpha,\beta).$$

In other words, f and g are equal in the interval (α,β) in the sense of distributions. In the case of $s > 1/2$, this is equivalent to saying that $\phi(x) = \psi(x)$ in the pointwise sense for any $x \in (\alpha,\beta)$.

Recently this exact boundary control problem was considered by Rosier [21] and the following result was obtained.

Let $T > 0$ be given. Let

$$\mathcal{N} = \left\{ 2\pi\sqrt{\frac{k^2+kl+l^2}{3}}; \quad \textit{k and l are any positive integers} \right\}.$$

If $\alpha < \beta$ satisfying $\beta - \alpha \notin \mathcal{N}$, then there exists a $\delta > 0$ such that for ϕ, $\psi \in L^2(\alpha,\beta)$ satisfying

$$\|\phi\|_{|L^2(\alpha,\beta)} \leq \delta \quad \textit{and} \quad \|\psi\|_{L^2(\alpha,\beta)} \leq \delta,$$

one can choose $h_1 = h_2 \equiv 0$ and $h_3 \in L^2(0,T)$ so that (1.5)–(1.6) *has a solution $u \in C([0,T]; L^2(\alpha,\beta)) \cap L^2(0,T; H^1(\alpha,\beta))$ satisfying*

$$u(x,0) = \phi(x) \quad \textit{and} \quad u(x,T) = \psi(x)$$

on the interval (α,β).

In this paper we present the following theorem.

THEOREM 1.1. *Let $T > 0$ and $s \geq 0$ be given and $[\alpha,\beta] \subset (\alpha_1,\beta_1)$. Suppose that*

$$w \equiv w(x,t) \in C^\infty(\alpha_1,\beta_1) \times (-\epsilon, T+\epsilon)$$

for some $\epsilon > 0$ satisfies

$$w_t + w_x + ww_x + w_{xxx} = 0, \qquad (x,t) \in (\alpha_1,\beta_1) \times (-\epsilon, T+\epsilon).$$

Then there exists a $\delta > 0$ such that for any ϕ, $\psi \in H^s(\alpha,\beta)$ satisfying

$$\|\phi(\cdot) - w(\cdot,0)\|_{H^s(\alpha,\beta)} \leq \delta \quad \textit{and} \quad \|\psi(\cdot) - w(\cdot,T)\|_{H^s(\alpha,\beta)} \leq \delta,$$

one can find control inputs h_1, h_2, and $h_3 \in L^2(0,T)$ ($h_j \in C[0,T]$, $j = 1,2,3$, if $s > 3/2$) such that the system (1.5)–(1.6) *has a solution*

$$u \in C([0,T]; H^s(\alpha,\beta)) \cap L^2(0,T; H^{s+1}(\alpha,\beta))$$

satisfying

$$u(x,0) = \phi(x) \quad \textit{and} \quad u(x,T) = \psi(x)$$

on the interval (α,β).

Several remarks are now in order.

(i) The local controllability result presented in Theorem 1.1 is stronger than the usual sense of local controllability. In general, we say a nonlinear system has exact local controllability if one can find control inputs to guide the system from a given initial state to a given terminal state so long as the "amplitude" of the initial and terminal states are small in a certain sense. The exact controllability we established for the system (1.3)–(1.4) is of this type and so is the result of Rosier cited earlier. In Theorem 1.1, the "amplitude" of the given initial and terminal states does not have to be small. In fact their "amplitudes" can be as large as one wishes so long as they locate in a δ-neighborhood of the initial state and terminal state, respectively, of a smooth solution of the KdV equation.

(ii) The solution u in the theorem is usually called the "path" to connect the given initial state ϕ and the terminal state ψ. Theorem 1.1 shows that the "smoother" ϕ and ψ are, the "smoother" the path—as are the control inputs one needs. In particular, if $s > 7/2$, then the solution

$$u \in C^1([0,T]; C[\alpha,\beta]) \cap C([0,T]; C^3[\alpha,\beta])$$

and the KdV equation is satisfied in the classical sense.

(iii) In the linear case (dropping the nonlinear term uu_x from the equation), Theorem 1.1 holds without any size restriction on the initial and terminal states. Moreover, the control inputs h_j, $j = 1, 2, 3$, are C^∞-smooth functions in the interval $(0,T)$ and the corresponding solution $u \equiv u(x,t)$ is also a C^∞-smooth function in the domain $(\alpha,\beta) \times (0,T)$, although its initial state ϕ and the terminal state ψ may belong only to the space $L^2(\alpha,\beta)$. In other words, the interior regularity of the control inputs and the corresponding solution are all C^∞ no matter what the regularity of the initial state and the terminal state is. Consequently, both (1.5) (dropping uu_x) and the boundary conditions (1.6) are satisfied in the pointwise sense for any $x \in (\alpha,\beta)$ and $t \in (0,T)$, rather than in the sense of distributions. We believe that this should also be the case for the nonlinear system.

(iv) The function $w(x,t)$ in Theorem 1.1 is assumed to be C^∞-smooth only for simplicity. In fact it will be sufficient to require $w \in C(-\epsilon, T+\epsilon, H^{s+1}(\alpha_1, \beta_1))$.

Consider a control system in a Hilbert space X described by the following abstract semilinear evolution equation:

$$\frac{dy}{dt} = Ay + F(y) + Bh, \tag{1.7}$$

where A is the infinitesimal generator of a C^0-semigroup $W(t)$ in the space X, F is a nonlinear mapping from X to X, B is a linear operator from another Hilbert space Y to the space X, and $h \in L^2(0,T;Y)$ is considered as a control input. A standard approach to establish exact controllability of the system (1.7) can be summarized as the following three steps.

Step 1. Establish exact controllability of the corresponding linear control system

$$\frac{dy}{dt} = Ay + Bh$$

by showing that there exists a bounded linear operator G from $X \times X$ to the space $L^2(0,T;Y)$ such that for any ϕ, $\psi \in X$, the unique solution $y(t)$ of the initial value problem

$$\frac{dy}{dt} = Ay + BG(\phi, \psi), \qquad y(0) = \phi$$

satisfies $y(T) = \psi$.

Step 2. Rewrite the nonlinear system (1.7) in its integral form:

$$y(t) = W(t)y(0) + \int_0^t W(t-\tau)F(y(\tau))d\tau + \int_0^t W(t-\tau)Bh(\tau)d\tau.$$

For given ϕ, $\psi \in X$, let

$$\omega(T, u) \equiv \int_0^T W(t-\tau)F(u(\tau))d\tau$$

and define

$$\Gamma(u) = W(t)\phi + \int_0^t W(t-\tau)F(u(\tau))d\tau + \int_0^t W(t-\tau)BG(\phi, \psi - \omega(T, u))(\tau)d\tau.$$

Note that

$$\Gamma(u)(0) = \phi \quad \text{and} \quad \Gamma(u)(T) = \omega(T, u) + \psi - \omega(T, u) = \psi$$

by virtue of the operator G.

Step 3. Show that the map Γ has a fixed point in a complete metric space $\mathcal{X}$ which is contained in the space $C([0, T], X)$.

However, there are some difficulties when applying this approach for the distributed control system (1.3)–(1.4). In particular, Step 3 is hard to carry out. Partially this is caused by the nonlinear term uu_x in the equation whose regularity is less than that of u because of differentiation. In order to show that the map Γ has a fixed point, it is necessary that the corresponding linear equation

$$(1.8) \qquad \begin{cases} u_t + u_{xxx} = f, \\ u(x, 0) = 0, \qquad 0 < x < 2\pi, \ t > 0, \\ u(0, t) = u(2\pi, t), \quad u_x(0, t) = u_x(2\pi, t), \quad u_{xx}(0, t) = u_{xx}(2\pi, t) \end{cases}$$

has certain smoothing properties to recover the lost regularity. But it is well known that, in contrast to the heat equation and the wave equation, $f \in L^1(0, T; H^s_p)$ can only lead to a solution u of (1.8) in the space $C([0, T]; H^s_p)$. Here H^s_p denotes the set of all periodic functions in the space $H^s(0, 2\pi)$. Thus the spatial regularity of the solution u of (1.8) is the same as that of the forcing term f. For many years people thought that it might be impossible for the solution u of (1.8) to possess higher spatial regularity than that of f. It came as a surprise when Bourgain [4], [5] discovered a rather subtle smoothing property for periodic solutions of the KdV equation in 1993. This type of smoothing property of Bourgain played a key role in establishing exact controllability of the system (1.3)–(1.4) in [27].

As for the boundary control system (1.5)–(1.6), there are even more difficulties using this approach. The spatial regularity is not only lost through the term uu_x but also through the boundary value functions h_1, h_2, and h_3 when they are considered as the trace of the solution u.

To obtain his result for the exact boundary control problem of the KdV equation, Rosier used the following type of smoothing property to carry out Step 3. For the initial-boundary value problem

$$\begin{cases} u_t + u_x + u_{xxx} = f, & x \in (\alpha, \beta),\ t > 0, \\ u(x,0) = \phi(x), & \\ u(\alpha, t) = u(\beta, t) = u_x(\beta, t) = 0, & \end{cases} \tag{1.9}$$

$\phi \in L^2(\alpha, \beta)$ and $f \in L^2(0,T; L^2(\alpha,\beta))$ implies that the solution u of (1.9) belongs to the space $C([0,T]; L^2(\alpha,\beta)) \cap L^2(0,T; H^1(\alpha,\beta))$ and $u_x(\alpha,t) \in L^2(0,T)$ with

$$\sup_{t\in(0,T)} \|u(\cdot,t)\|_{L^2(0,T)} + \|u\|_{L^2(0,T;H^1(\alpha,\beta))} + \|u_x(\alpha,t)\|_{L^2(0,T)}$$
$$\leq c\left(\|\phi\|_{L^2(0,T)} + \|f\|_{L^2(0,T;L^2(\alpha,\beta))}\right),$$

where $c > 0$ is independent of f and ϕ. To establish the needed exact controllability of the associated linear system, he used the Hilbert uniqueness method together with a compactness argument.

Our approach is different from Rosier's. Our strategy to overcome these difficulties caused by losing regularity is to bypass them rather than look for some stronger smoothing properties. More precisely, instead of considering boundary control of the KdV equation posed on the finite interval (α, β), we consider initial value control of the KdV equation posed on the whole real line R:

$$\begin{cases} u_t + u_x + uu_x + u_{xxx} = 0, & x,\ t \in R, \\ u(x,0) = h(x). & \end{cases} \tag{1.10}$$

Initial value control problem. *Let $T > 0$ and $s \geq 0$ be given. For any ϕ, $\psi \in H^s(R)$, can we find an external modification of $\phi(x)$, $h \in H^s(R)$, such that the corresponding solution u of the initial value problem (IVP)* (1.10) *satisfies*

$$u(x,0) = \phi(x) \quad \textit{and} \quad u(x,T) = \psi(x)$$

on the interval (α, β)?

An affirmative answer to this initial value control problem leads to a positive answer to the boundary control problem asked earlier, but not vice versa. Indeed, for any given ϕ, $\psi \in H^s(\alpha,\beta)$, we can extend ϕ and ψ to be functions in $H^s(R)$ (let us still write their extensions as ϕ and ψ). If we could find an $h \in H^s(R)$ such that the corresponding unique solution u of the IVP (1.10) satisfies

$$u(x,0) = \phi(x) \quad \text{and} \quad u(x,T) = \psi(x)$$

on the interval (α,β), then we could simply choose $h_1(t) = u(\alpha,t)$, $h_2(T) = u(\beta,t)$, and $h_3(t) = u_x(\beta,t)$. The restriction of u to the domain $[\alpha,\beta] \times [0,T]$ would be the desired solution of the boundary control system (1.5)–(1.6).

We point out that this kind of problem is of interest in its own right. Let us take the linearized KdV equation

$$u_t + u_x + (a(x,t)u)_x + u_{xxx} = 0, \quad x, t \in R, \tag{1.11}$$

as an example. For given ϕ, $\psi \in H^s(R)$, to find a solution u of (1.11) satisfying

$$u(x,0) = \phi(x) \quad \text{and} \quad u(x,T) = \psi(x) \tag{1.12}$$

on the interval $(-\infty, \infty)$ is an overdetermined problem; there is no solution in general. A further question then is the following.

Is there a solution of (1.11) *satisfying* (1.12) *on a given finite interval*?

As we will see later, there are infinitely many such solutions! We will prove the following result for the initial value control problem.

THEOREM 1.2. *Let $T > 0$ and $s \geq 0$ be given. Suppose that $w \in C(R; H^\infty(R))$ is a given solution of*

$$w_t + w_x + ww_x + w_{xxx} = 0, \quad x, t \in R.$$

Then there exists a $\delta > 0$ such that for any ϕ, $\psi \in H^s(\alpha, \beta)$ satisfying

$$\|\phi(\cdot) - w(\cdot, 0)\|_{H^s(\alpha,\beta)} \leq \delta, \quad \|\psi(\cdot) - w(\cdot, T)\|_{H^s(\alpha,\beta)} \leq \delta,$$

one can find $h \in H^s(R)$ such that the corresponding solution u of the IVP (1.10) *satisfies*

$$u(x,0) = \phi(x) \quad \textit{and} \quad u(x,T) = \psi(x)$$

on the interval (α, β).

This theorem is proved using the approach outlined earlier for the abstract system. In contrast with the boundary control problem of the KdV equation, the initial value control problem is easier to handle although it is more general. First of all the needed smoothing properties of the KdV equation have been well established in the literature. Second, we do not need to handle some difficult regularity problems associated with the boundary value problem.

The idea of dealing with a problem posed on a finite interval by extending it to a related problem on the whole line R is not new in the literature. To our knowledge, it is Russell [22], [23] who first used this idea in the field of control theory to establish exact boundary controllability of a class of linear hyperbolic equations. Later this idea was explored by Littman [17] and Littman and Taylor [18]. A systematic method has been developed to handle exact boundary control problems of partial differential equations; the method is summarized by Littman and Taylor [18] as follows:

reversibility + smoothing property + uniqueness → controllability.

We will use this method to establish exact controllability of the associated linear KdV equation which is one of the key steps leading to the exact controllability of the nonlinear KdV equation.

The remainder of the paper is organized as follows.

In section 2, we recall various smoothing properties of the linear KdV equation posed on the whole line R. As we have explained earlier, those smoothing properties will play a major role in the proof of our main theorems.

In section 3, the exact controllability of the associated linear system is established using a method due to Littman [17] with some modification.

The proofs of our main results, Theorems 1.1 and 1.2, are provided in section 4.

2. Smoothing properties. In this section we recall several smoothing properties of the KdV equation. Those smoothing properties not only play important roles in establishing exact controllability of the KdV equation but also are very useful in studying other aspects of the KdV equation. We refer to [3], [4], [5], [11], [12], [13], [14], [31], [32], [33] for other applications of those smoothing properties.

To begin, let $\{W(t)\}_{-\infty}^{+\infty}$ denote the unitary group generated by the linear third order operator A from $L^2(R)$ to $L^2(R)$:

$$Af = -f' - f'''$$

with $\mathcal{D}(A) = H^3(R)$. Then the solution of the IVP associated with the linear KdV equation

$$\begin{cases} u_t + u_x + u_{xxx} = 0, \quad x, t \in R, \\ u(x,0) = \phi(x) \end{cases} \tag{2.1}$$

is given by

$$u(t) = W(t)\phi,$$

and the solution of the inhomogeneous equation

$$\begin{cases} u_t + u_x + u_{xxx} = f, \quad x, t \in R\ , \\ u(x,0) = 0 \end{cases} \tag{2.2}$$

is represented by

$$u(t) = \int_0^t W(t-\tau)f(\cdot,\tau)d\tau.$$

For given $b > 0$ let L_b^2 denote the weighted Sobolev space $L^2(e^{2bx}dx)$. The operator A also generates a semigroup $W_b(t)$ in the space L_b^2 which is formally equivalent to the semigroup

$$W_b(t) = \exp[-t(D-b)^3 - t(D-b)] \tag{2.3}$$

in the space $L^2(R)$ (cf. [10]).

The following two smoothing properties of the linear KdV equation are due to Kato [10]. Lemma 2.1 asserts that the solution u of the IVP (2.1) is smoother than its initial data ϕ if ϕ belongs to a certain weighted Sobolev space. Lemma 2.2 shows that the solutions of the inhomogeneous problem (2.2) possess a similar smoothing property.

LEMMA 2.1. *$\{W_b(t),\ t > 0\}$ is an infinitely differentiable semigroup on $H^s(R)$ for each real s. Moreover, for any $s \le s'$, $W_b(t)$ is bounded as an operator from $H^s(R)$ to $H^{s'}(R)$,*

$$\|W_b(t)\|_{\mathcal{L}(H^s,\,H^{s'})} \le ct^{-(s'-s)/2}\exp(b^3t), \tag{2.4}$$

where $\|\cdot\|_{\mathcal{L}(H^s,\,H^{s'})}$ denotes the operator norm for bounded linear operators from $H^s(R)$ to $H^{s'}(R)$.

As a result, if $\phi \in H^s(R)$ has compact support, then $W(t)\phi$ is infinitely smooth everywhere except at $t = 0$.

LEMMA 2.2. *For any real s, let*

$$e^{bx}u \in L^\infty([0,T]; H^s(R)), \qquad e^{bx}f \in L^\infty([0,T]; H^{s-1}(R)),$$

and

$$u_t + u_{xxx} + u_x = f, \quad 0 < t < T.$$

Then

$$e^{bx}u \in C([0,T]; H^0) \cap C([0,T]; H^s) \quad \textit{for every } s' < s+1,$$

and

$$e^{bx}u(t) = W_b(t)u(0) + \int_0^t W_b(t-\tau)e^{bx}f(\tau)d\tau.$$

Next we recall several more subtle smoothing properties of the linear KdV equation due to Bourgain [4], [5] and Kenig, Ponce, and Vega [13]. First we introduce a special Sobolev-type space used by Kenig, Ponce, and Vega in [13].

For any $s, b \in R$, let $Y_{s,b}$ be the completion of the space $S(R^2)$ of tempered test functions with respect to the norm

$$\|f\|^2_{Y_{s,b}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (1 + |\tau - \xi - \xi^3|)^{2b}(1+|\xi|)^{2s}|\hat{f}(\xi,\tau)|^2 d\xi d\tau,$$

where $\hat{f}(\xi, \tau)$ denotes the Fourier transform of $f(x,t)$. As shown in [13], if $u \in Y_{s,b}$ with $s > -1$ and $b > 1/2$, one has

$$u \in C^{[1+s],\alpha}_{loc}(R; L^2_t(R))$$

for any $0 \le \alpha \le 1 + s - [1+s]$, and consequently

$$u \in L^p_{x,loc}(R; L^2_t(R)),$$

for $1 \le p \le \infty$. Here $[1+s]$ denotes the integer part of $1+s$. In particular, if $s \ge 0$, then $u \in C^1(R; L^2_t(R))$.

The following lemmas can be found in [13] or followed directly from the results there (cf. also [3], [33]).

LEMMA 2.3. *Let there be given $s > -3/4$ and $\sigma \in C_0^\infty(R)$. Then there exists a $\beta_0 \in (1/2, 1)$ such that for any $b \in (1/2, \beta_0)$, there is a constant $c > 0$ for which*

$$\|\sigma(t)\partial_x(uv)\|_{Y_{s,b-1}} \le c\|u\|_{Y_{s,b}}\|v\|_{Y_{s,b}} \tag{2.5}$$

for any u, $v \in Y_{s,b}$.

Equation (2.5) is a key estimate, which reveals a subtle smoothing property of the linear KdV equation. When $s = 0$ and $b = 1/2$, the estimate was established by Bourgain in [5]. Then it was proved by Kenig, Ponce, and Vega in [13] for $s > -5/8$, and for $s > -3/4$ in [14]. Estimate (2.5) is not true for $s < -3/4$ (see [14]).

LEMMA 2.4. *Let $b > 1/2$ and $s \in R$ be given. Then $Y_{s,b} \subset C(R; H^s(R))$ and there is a constant $c > 0$ such that for any $f \in Y_{s,b}$,*

$$\sup_{t \in R} \|f(\cdot, t)\|_{H^s(R)} \leq c \|f\|_{Y_{s,b}}.$$

LEMMA 2.5. *Suppose $\sigma \in C_0^\infty(R)$. For given $s \in R$ and $b \in (1/2, 1]$, there is a constant $c > 0$ such that*

$$\|\sigma(t) W(t) u_0\|_{Y_{s,b}} \leq c \|u_0\|_{H^s(R)} \tag{2.6}$$

and

$$\left\| \sigma(t) \int_0^t W(t-\tau) f(\tau) d\tau \right\|_{Y_{s,b}} \leq c \|f\|_{Y_{s,b-1}} \tag{2.7}$$

for any $u_0 \in H^s(R)$ and $f \in Y_{s,b-1}$.

Remark 2.1. Combining (2.5) and (2.7), one has

$$\left\| \sigma_1(t) \int_0^t W(t-\tau) \sigma^2(\tau) (\partial_x(uv))(\cdot, \tau) d\tau \right\|_{Y_{s,b}} \leq c \|u\|_{Y_{s,b}} \|v\|_{Y_{s,b}}. \tag{2.8}$$

This constitutes a global smoothing property of the linear KdV equation.

In order to prove the main exact control results in the paper we need to extend the above smoothing properties to solutions of the linearized KdV equation with variable coefficient:

$$u_t + u_x + (a(x,t)u)_x + u_{xxx} = 0, \qquad u(x,0) = \phi(x) \tag{2.9}$$

for $x, t \in R$. We assume $a \in C(R; H^\infty(R))$ for the sake of simplicity. By the standard semigroup theory, for each given real s and $\phi \in H^s(R)$, (2.9) possesses a unique solution $u \in C(R; H^s(R))$ satisfying

$$\sup_{t \in [-T,T]} \|u(\cdot, t)\|_{H^s(R)} \leq c \|\phi\|_{H^s(R)}, \tag{2.10}$$

where the constant c depends only on s, T, and a. In addition, if $e^{bx}\phi \in H^s(R)$ for some $b > 0$, then $e^{bx}u \in C(R; H^s(R))$. The following lemma is an extension of Lemma 2.1.

LEMMA 2.6. *For given $T > 0$, $s \in R$, and $b > 0$, suppose $e^{bx}\phi(x) \in H^s(R)$ and u is the solution of (2.9). Then $e^{bx}u(x,t)$ belongs to the space $C((0,T]; H^{s'}(R))$ for any $s \leq s' < s+1$. Moreover,*

$$\|e^{b\cdot} u(\cdot, t)\|_{H^{s'}(R)} \leq c t^{-(s'-s)/2} \|\phi(\cdot)\|_{H^s(R)} \tag{2.11}$$

for any $0 < t < T$, where $c > 0$ depends only on s, s', a, and T.

Proof. Since $e^{bx}u \in C([0,T]; H^s(R))$ and

$$u_t + u_x + u_{xxx} = -(a(x,t)u)_x$$

with $e^{bx}(a(x,t)u)_x \in C([0,T]; H^{s-1}(R))$, it follows from Lemma 2.2 that $e^{bx}u \in C((0,T]; H^{s'}(R))$ for any $s' < s+1$ and

$$e^{bx}u(t) = W_b(t)\phi(x) - \int_0^t W_b(t-\tau) e^{bx} (a(x,\tau)u(x,\tau))_x d\tau.$$

Applying Lemma 2.1 yields

$$\|e^{b\cdot}u(\cdot,t)\|_{H^{s'}(R)} \le \|W_b(t)\phi\|_{H^{s'}(R)} + \int_0^t \|W_b(t-\tau)e^{b\cdot}(a(\cdot,t)u(\cdot,t))_x\|_{H^{s'}(R)}d\tau$$

$$\le ct^{-(s'-s)/2}e^{b^3t}\|\phi\|_{H^s(R)} + c\int_0^t (t-\tau)^{-(s'-s+1)/2}e^{b^3(t-\tau)}\|e^{b\cdot}(a(\cdot,\tau)u(\cdot,\tau))_x\|_{H^{s-1}(R)}d\tau$$

$$\le ct^{-(s'-s)/2}e^{b^3t}\|\phi\|_{H^s(R)} + c\int_0^t (t-\tau)^{-(s'-s+1)/2}e^{b^3(t-\tau)}\|e^{b\cdot}u(\cdot,\tau)\|_{H^s(R)}d\tau$$

for any $t > 0$. Let

$$w(t) = t^{(s'-s)/2}e^{-b^3t}\|e^{b\cdot}u(\cdot,\tau)\|_{H^{s'}(R)}.$$

Then

$$w(t) \le c\|\phi\|_{H^s(R)} + c\int_0^t t^{(s'-s)/2}\tau^{-(s'-s)/2}(t-\tau)^{-(s'-s+1)/2}w(\tau)d\tau.$$

One can choose T^* small enough so that

$$c\int_0^t t^{(s'-s)/2}\tau^{-(s'-s)/2}(t-\tau)^{-(s'-s+1)/2}d\tau < 1/2$$

for $0 < t \le T^*$. Then

$$w(t) \le c\|\phi\|_{H^s(R)} + \frac{1}{2}\sup_{0\le\tau\le t} w(\tau)$$

for any $0 < t \le T^*$. This implies that

$$w(t) \le c\|\phi\|_{H^s(R)}$$

for any $0 < t \le T^*$, i.e.,

$$\|e^{b\cdot}u(\cdot,t)\|_{H^{s'}(R)} \le ct^{-(s'-s)/2}e^{b^3t}\|\phi\|_{H^s(R)}$$

for any $t \in (0, T^*]$. As the chosen T^* does not depend on ϕ, a standard argument enables us to extend T^* to $T^* = T$ so that for any $t \in (0, T)$,

$$\|e^{b\cdot}u(\cdot,t)\|_{H^{s'}(R)} \le ct^{-(s'-s)/2}e^{3t}\|\phi\|_{H^s(R)}.$$

Here $c > 0$ depends only on T, s, and s' as well as a. The proof is complete. □

The next lemma is an extension of Lemma 2.5 for the linearized KdV equation (2.9). Its proof can be found in [3].

LEMMA 2.7. *Suppose $s > -3/4$ is given, $T > 0$, b is as in Lemma 2.3 with $1/2 < b \le 1$, and $a \in Y_{s,b}$. Then for any $u_0 \in H^s(R)$ and $f \in Y_{s,b-1}$, the IVP (2.9) has a unique solution $u \in C(-T, T; H^s(R))$ which is the restriction to $(-T, T)$ of a function $\bar{u} \in Y_{s,b}$ that satisfies the estimate*

$$\|\bar{u}\|_{Y_{s,b}} \le c_1\left(\|u_0\|_{H^s(R)} + \|f\|_{Y_{s,b-1}}\right), \tag{2.12}$$

where $c_1 = c_1(T, \|a\|_{Y_{s,b}})$.

3. The linear system. In this section we consider the initial value control problem for the linearized KdV equation posed on the whole line R:

$$(3.1)\qquad \begin{cases} z_t + z_x + (a(x,t)z)_x + z_{xxx} = 0, & x,\ t \in R, \\ \\ z(x,0) = h(x). \end{cases}$$

Again we assume $a \equiv a(x,t)$ is H^∞ smooth for simplicity. The theorem below is a linear version of Theorem 1.2 but without the restriction on the amplitude of the initial and terminal states.

THEOREM 3.1. *Let $s \geq 0$ and $T > 0$ be given. There exists a bounded linear operator G: $H^s(\alpha,\beta) \times H^s(\alpha,\beta) \to H^s(R)$ such that for any ϕ and ψ in $H^s(\alpha,\beta)$ if one chooses $h = G(\phi,\psi) \in H^s(R)$, then the corresponding solution z of* (3.1) *satisfies*

$$z(x,0) = \phi(x), \qquad z(x,T) = \psi(x)$$

on the interval $\in (\alpha,\beta)$ and

$$\|h\|_{H^s(R)} = \|G(\phi,\psi)\|_{H^s(R)} \leq c(\|\phi\|_{H^s(\alpha,\beta)} + \|\psi\|_{H^s(\alpha,\beta)}),$$

where $c > 0$ is independent of ϕ and ψ.

Remark 3.1. By the standard semigroup theory, the solution

$$u \in C([0,T]; H^s(R)) \cap L^2(0,T; H^{(s+1)}_{loc}(R)).$$

However, from the proof of the theorem one can see that, in fact, $u \in C^\infty(R \times (0,T))$ even though its initial state and the terminal state belong only to $L^2(R)$.

Proof. Without loss of generality, we may assume that $\psi \equiv 0$ since the system is time reversible. We split the proof into several steps.

Step 1. Let $\epsilon > 0$ be small and let $p(x)$ be a smooth function satisfying

$$p(x) = \begin{cases} 1 & \text{for } x \in (\alpha,\beta), \\ \\ 0 & \text{for } x \in (\alpha-\epsilon, \beta+\epsilon). \end{cases}$$

We define an operator $\mathcal{P}$ from $H^s(R)$ to $H^s(R)$ by

$$\mathcal{P}(f)(x) = p(x)f(x)$$

for any $f \in H^s(R)$. Obviously $\mathcal{P}$ is a bounded linear operator and the support of $\mathcal{P}(f)$ is in the interval $(\alpha-\epsilon, \beta+\epsilon)$ for any $f \in H^s(R)$. In addition, let $\mathcal{E}$ denote the usual extension operator from $H^s(\alpha,\beta)$ to $H^s(R)$: for any $f \in H^s(\alpha,\beta)$,

- $\mathcal{E}(f) \in H^s(R)$ and $\mathcal{E}(f)(x) = f(x)$ on (α,β);
- $\mathcal{E}(f)$ has a compact support;
- $\|\mathcal{E}(f)\|_{H^s(R)} \leq c\|f\|_{H^s(\alpha,\beta)}$, where $c > 0$ does not depend on f.

Let $W_a(t)$ denote the continuous group associated to the IVP (3.1). Then for any given $h \in H^s(R)$, the unique solution of (3.1) is represented by

$$z(t) = W_a(t)h.$$

For given $T > 0$, let $K = W_a(-T)\mathcal{P}W_a(T)\mathcal{E}$. Consider the following function equation in the space $H^s(\alpha,\beta)$:

$$(3.2)\qquad (I - K)f = \phi.$$

Claim 3.1. *For any $\phi \in H^s(\alpha,\beta)$, if* (3.2) *has a solution $f \in H^s(\alpha,\beta)$, then there exists an $h \in H^s(R)$ such that the corresponding solution z of the IVP* (3.1) *satisfies*

$$z(x,0) = \phi(x), \qquad z(x,T) = 0$$

in the interval (α,β).

Indeed, if $f \in H^s(\alpha,\beta)$ is a solution of (3.2) for given ϕ, let

$$h = \mathcal{E}f - W_a(-T)\mathcal{P}W_a(T)\mathcal{E}f \tag{3.3}$$

be the initial value of (3.1); then the unique solution $z = W_a(t)h$ of (3.1) satisfies

$$z(x,0) = W_a(0)h = \phi(x), \quad \text{on } (\alpha,\beta)$$

and

$$\begin{aligned} z(x,T) &= W_a(T)h \\ &= W_a(T)\mathcal{E}f - \mathcal{P}W_a(T)\mathcal{E}f \\ &= 0 \qquad \text{on } (\alpha,\beta). \end{aligned}$$

Remark 3.2. If h is given by (3.3), then $z = W_a(t)h \in C^\infty(R \times (0,T))$. This is because $h_1 \equiv \mathcal{E}f$ has compact support, and $h_2 \equiv W_a(-T)\mathcal{P}W_a(T)\mathcal{E}f \in H^\infty(R)$ for $\mathcal{P}W_a(T)\mathcal{E}f \in H^\infty(R)$.

Step 2. We seek a necessary and sufficient condition for (3.2) to have a solution.

By the definition of the operator $\mathcal{E}$, for any $f \in H^s(\alpha,\beta)$, the function $\mathcal{E}f$ belongs to the space $H^s(R)$ and has compact support in R. Therefore $e^{bx}\mathcal{E}f \in H^s(R)$. By Lemma 2.6,

$$\|e^{b\cdot}W_a(T)\mathcal{E}f\|_{H^{s'}(R)} \le cT^{-(s'-s)/2}\|\mathcal{E}f\|_{H^s(R)} \le cT^{-(s'-s)/2}\|f\|_{H^s(\alpha,\beta)}$$

for any $s \le s' < s+1$. It then follows that

$$\|W_a(T)\mathcal{E}f\|_{H^{s'}(R)} \le c\|e^{b\cdot}W_a(T)\mathcal{E}f\|_{H^{s'}(R)} \le cT^{-(s'-s)/2}\|f\|_{H^s(\alpha,\beta)}.$$

Consequently, as an operator from $H^s(\alpha,\beta)$ to $H^s(\alpha,\beta)$, $W_a(T)\mathcal{E}$ is a compact operator and so is the operator $K = W_a(-T)\mathcal{P}W_a(T)\mathcal{E}$. Equation (3.2) is then a Fredholm equation in the space $H^s(\alpha,\beta)$. According to the Fredholm theory, the following statements are true for the operator $(I-K)$:

(a) the range $\mathcal{R}(I-K)$ of the operator $I-K$ is a closed subspace of $H^s(\alpha,\beta)$;

(b) the null space $\mathcal{N}(I-K)$ of $I-K$ is a finite dimensional subspace of $H^s(\alpha,\beta)$, and its dimension equals the dimension of the null space $\mathcal{N}(I-K^*)$ of the operator $I-K^*$. Here K^* is the adjoint of K;

(c) $\phi \in \mathcal{R}(I-K)$ if and only if ϕ is orthogonal to the space $\mathcal{N}(I-K^*)$;

(d) considered as an operator from $(\mathcal{N}(I-K))^\perp$, the orthogonal complement of $\mathcal{N}(I-K)$ in the space $H^s(\alpha,\beta)$, to $\mathcal{R}(I-K)$, the operator $I-K$ has a bounded inverse $(I-K)^{-1}$ from$\mathcal{R}(I-K)$ to $(\mathcal{N}(I-K))^\perp$, i.e., for any $\phi \in \mathcal{R}(I-K)$ there exists a unique $f \in (\mathcal{N}(I-K))^\perp$ satisfying

$$\|f\|_{H^s(\alpha,\beta)} = \|(I-K)^{-1}\phi\|_{H^s(\alpha,\beta)} \le c\|\phi\|_{H^s(\alpha,\beta)}, \tag{3.4}$$

where $c > 0$ is independent of ϕ.

As a result, we may assume that $\mathcal{N}(I-K^*)$ is a space of dimension m with $\{\psi_1, \psi_2, \dots, \psi_m\}$ as its orthonormal basis; $(\psi_i, \psi_j)_{H^s(\alpha,\beta)} = \delta_{ij}$ for i, $j = 1, 2, \dots, m$. The following claim then holds.

CLAIM 3.2. *For a given function* $\phi \in H^s(\alpha, \beta)$, (3.2) *has a solution, i.e.,* $\phi \in \mathcal{R}(I-K)$, *if and only if*

$$(\phi, \psi_j)_{H^s(\alpha,\beta)} = 0, \qquad j = 1, 2, \dots, m.$$

Step 3. We show that the set of all $\phi \in H^s(\alpha, \beta)$ such that (3.2) has a solution is a dense subset of the space $H^s(\alpha, \beta)$.

CLAIM 3.3. *For any given* $\phi \in H^s(\alpha, \beta)$ *and* $\epsilon > 0$, *one can find* $h \in H^\infty(R)$ *and* $\phi_1 \in H^\infty(\alpha, \beta)$ *with*

$$\|\phi - \phi_1\|_{H^s(\alpha,\beta)} \leq \epsilon$$

such that the solution $z = W_a(t)h$ *of* (3.1) *satisfies*

$$z(x, 0) = \phi_1(x) \quad \textit{and} \quad z(x, T) = 0 \quad \textit{on } (\alpha, \beta).$$

The following technical lemma is needed for the proof of Claim 3.3.

LEMMA 3.2. *Let* $Q = (\alpha, \beta) \times (0, T)$,

$$L \equiv \partial_t + \partial_x + \partial_x(a(x,t)u) + \partial_x^3,$$

and $\mathcal{L}$ *be the set of the collection of all smooth functions* $f(x,t)$ *in the cylinder* $R \times [-2T, 2T]$ *vanishing near the top and the bottom of* $\bar{Q}$. *Then the set*

$$L(\mathcal{L}) = \{\textit{the restriction of } Lf \textit{ on } (\alpha, \beta) \times (0, T) : \quad f \in \mathcal{L}\}$$

is dense in the space $L^2(0, T; H^s(\alpha, \beta))$.

We prove Lemma 3.2 by contradiction. Suppose that the set $L(\mathcal{L})$ is not dense in $L^2(0, T : H^s(\alpha, \beta))$. Then there exists a nonzero $g \in L^2(0, T; H^s(\alpha, \beta))$ which is orthogonal to $\overline{L(\mathcal{L})}$. Thus

$$L^* g = 0 \qquad \text{in } Q$$

and g has Cauchy data zero on the lateral boundary of Q (as a matter of fact, g can be extended to be zero outside the lateral boundary of Q). In other words, g solves

$$g_t + g_x + a(x,t) g_x + g_{xxx} = 0, \qquad x \in R, \ t \in (0, T),$$

and vanishes outside $(\alpha, \beta) \times (0, T)$. By the smoothing properties of the equation, we know that $g(\cdot, t) \in C^\infty(R)$ for any $t \in (0, T)$, which leads to $g \equiv 0$ by the unique continuation property of the equation (cf. [30]). This is a contradiction.

Now we prove Claim 3.3. First we choose $\psi \in H^\infty(R)$ such that

$$\|\psi - \phi\|_{H^s(\alpha,\beta)} \leq \epsilon.$$

Let v be the solution of the IVP

$$v_t + v_x + (a(x,t)v)_x + v_{xxx} = 0, \quad v(x, 0) = \psi(x)$$

for $x \in R$ and $t \in (0, T)$. Suppose that $\theta(t)$ is a given cutoff function: $\theta(t) \equiv 1$ near $t = 0$ and $\theta(t) \equiv 0$ near T. Let $L(\theta(t)v) = F(x, t)$. Note that F is C^∞-smooth and $F \in C^k(R; H^m(R))$ for any $k \geq 1$, $m \geq 1$. Then there exists a smooth function $u \in C^\infty(R, H^\infty(R))$ vanishing near the top and the bottom of $\overline{Q}$ such that $F_1 := Lu \in C^\infty(R, H^\infty(R))$ satisfying

$$\|F - F_1\|_{L^2(0,T;H^s(\alpha,\beta))} \leq \epsilon$$

for $L(\mathcal{L})$ is dense in the space $L^2(0,T;H^s(\alpha,\beta))$ (Lemma 3.2). Let w be the unique solution of

$$w_t + w_x + (a(x,t)w)_x + w_{xxx} = F - F_1, \quad w(x,0) = 0$$

for $x \in R$ and $t \in (0,T)$ and set

$$z(x,t) = \theta(t)v - u - w - y,$$

where $y(x,t)$ solves the linear KdV equation on R posed backward in time:

$$y_t + y_x + (a(x,t)y)_x + y_{xxx} = 0, \qquad y(x,T) = w(x,T).$$

Then $Lz = 0$ and $z(x,T) = w(x,T) - y(x,T) = 0$ on (α,β) and

$$\begin{aligned} \|z(\cdot,0) - \phi(\cdot)\|_{H^s(\alpha,\beta)} &\leq \|z(\cdot,0) - \psi(\cdot)\| + \|\psi(\cdot) - \phi(\cdot)\| \\ &\leq \epsilon + \|y(\cdot,0)\|_{H^s(\alpha,\beta)} \\ &\leq \epsilon + c\|w(\cdot,T)\|_{H^s(\alpha,\beta)} \\ &\leq \epsilon + c\|F - F_1\|_{L^2(0,T;H^s(\alpha,\beta))} \\ &\leq c\epsilon. \end{aligned}$$

Consequently, we only need to choose $h(x) = \phi_1(x) = z(x,0)$ to complete the proof of Claim 3.3.

Step 4. We construct the needed control operator G to complete the proof of Theorem 3.1.

Recall from Step 2 that the set $\{\psi_1, \ldots, \psi_m\}$ is an orthonormal basis for the space $\mathcal{N}(I - K^*)$ and

$$(\psi_i, \psi_j)_{H^s(\alpha,\beta)} = \delta_{ij}$$

for i, $j = 1, 2, \ldots, m$. According to Claim 3.3, we can find $\phi_j \in H^\infty(\alpha,\beta)$ sufficiently close to ψ_j, and $h_j \in H^\infty(R)$ for $j = 1, 2, \cdot, m$ such that

$$\det((\phi_j, \psi_i)_{H^s(\alpha,\beta)}) \neq 0 \tag{3.5}$$

and $w_j(t) = W_a(t)h_j$ satisfies

$$w_j(x,0) = \phi_j(x), \qquad w_j(x,T) = 0 \tag{3.6}$$

on the interval (α,β) for $j = 1, 2, \ldots, m$.

For any given $\phi \in H^s(\alpha,\beta)$, consider the following system of linear equations for a_1, $a_2, \ldots, a_m$:

$$\sum_{i=1}^{n} (\psi_j, \phi_i)_{H^s(\alpha,\beta)} a_i = (\phi, \psi_j), \ j = 1, 2, \ldots, m. \tag{3.7}$$

By (3.5), it has a unique solution $\vec{a} = (a_1, a_2, \ldots, a_m)$. Moreover,

$$\sum_{j=1}^{m} |a_j| \leq c\|\phi\|_{H^s(\alpha,\beta)} \tag{3.8}$$

for some constant $c >$ independent of ϕ. If we let

$$\phi_N = \sum_{j=1}^{m} a_j\phi_j, \qquad h_N = \sum_{j=1}^{m} a_j h_j, \tag{3.9}$$

then $z_N(t) = W_a(t)h_N = \sum_{j=1}^{m} W_a(t)a_j h_j$ satisfies

$$z_N(x,0) = \sum_{j=1}^{m} a_j\phi_j(x) = \phi_N(x), \qquad z_N(x,T) = 0 \tag{3.10}$$

on the interval (α, β). In addition,

$$\begin{aligned}\|h_N\|_{H^s(R)} &\leq \sum_{j=1}^{m} |a_j|\|h_j\|_{H^s(R)} \\ &\leq \max_{j=1,2,\ldots,m} \|h_j\|_{H^s(R)} \sum_{j=1}^{m} |a_j|.\end{aligned}$$

By (3.8), there exists a constant $c > 0$ independent of ϕ such that

$$\|h_N\|_{H^s(R)} \leq c\|\phi\|_{H^s(\alpha,\beta)}. \tag{3.11}$$

Furthermore, if we let

$$\phi_R = \phi - \phi_N,$$

then it follows from (3.7) that

$$(\phi_R, \psi_i)_{H^s(\alpha,\beta)} = 0, \quad i = 1, 2, \ldots, m.$$

Hence $\phi_R \in \mathcal{R}(I - K)$ by Claim 3.2. According to what we have proved in Step 2 there exists a unique $f_R \in (\mathcal{N}(I-K))^{\perp}$ which solves (3.2) and satisfies

$$\|f_R\|_{H^s(\alpha,\beta)} \leq c\|\phi_R\|_{H^s(\alpha,\beta)} \leq c\|\phi\|_{H^s(\alpha,\beta)} \tag{3.12}$$

for some constant $c > 0$ independent of ϕ. Define

$$h_R = \mathcal{E}f_R - W_a(-T)\mathcal{P}W_a(T)\mathcal{E}f_R. \tag{3.13}$$

By Claim 3.1, the solution $w_R(t) \equiv W_a(t)h_R$ of (3.1) satisfies

$$w_R(x,0) = \phi_R(x), \quad w_R(\cdot,T) = 0$$

on the interval (α, β) and

$$\|h_R\|_{H^s(R)} \leq c\|\phi_R\|_{H^s(\alpha,\beta)} \leq c\|\phi\|_{H^s(\alpha,\beta)}, \tag{3.14}$$

where c is independent of ϕ. Consequently, if we define

$$G\phi = h_N + h_R \quad \text{for any } \phi \in H^s(\alpha,\beta),$$

where h_N and h_R are as given by (3.9) and (3.14), respectively, then G is a bounded linear operator from $H^s(\alpha,\beta)$ to $H^s(R)$ and $z(t) = W_a(t)G\phi$ is a solution of (3.1) with $h = G(\phi)$, which satisfies

$$z(x,0) = \phi(x), \qquad z(x,T) = 0$$

on the interval (α,β). Moreover, $W_a(t)G(\phi) \in C^\infty(R \times (0,T))$ since both $W_a(t)h_N$ and $W_a(t)h_R$ belong to $C^\infty(R \times (0,T))$. The proof is complete. □

As a corollary of Theorem 3.1, we have the following result for the two point (in time) boundary value problem for the linearized KdV equation:

$$\text{(3.15)} \qquad \begin{cases} u_t + u_x + (a(x,t)u)_x + u_{xxx} = 0, \quad x \in R,\ t \in (0,T), \\ u(x,0) = \phi(x) \quad \text{and} \quad u(x,T) = \psi(x) \quad \text{on the interval } (\alpha,\beta). \end{cases}$$

COROLLARY 3.3. *Let $s \geq 0$ and $T > 0$ be given. Then for any $\phi, \psi \in H^s(R)$, the problem* (3.15) *has a solution*

$$u \in C([0,T]; H^s(R)) \cap L^2(0,T; H^{s+1}_{loc}(R)) \cap C^\infty(R \times (0,T)).$$

Remark 3.3. In fact, there are infinitely many such solutions.

As another corollary to Theorem 3.1, we have the following result for boundary control of the linearized KdV equation posed on a bounded domain (α,β):

$$\text{(3.16)} \qquad \begin{cases} u_t + u_x + (a(x,t)u)_x + u_{xxx} = 0, \quad x \in (\alpha,\beta),\ t \in (0,\infty), \\ u(\alpha,t) = h_1(t), \qquad u(\beta,t) = h_2(t), \qquad u_x(\beta,t) = h_3(t), \end{cases}$$

where $h_j(t)$, $j = 1,2,3$, are considered as control inputs.

COROLLARY 3.4. *Let $s \geq 0$ and $T > 0$ be given. For any $\phi \in H^s(\alpha,\beta)$, $\psi \in H^s(\alpha,\beta)$, there exist h_1, h_2, and $h_3 \in L^2(0,T)$ ($h_j \in C[0,T]$, $j = 1,2,3$, if $s > 3/2$) such that* (3.16) *has a solution*

$$u \in C([0,T]; H^s(\alpha,\beta)) \cap L^2(0,T; H^{s+1}(\alpha,\beta))$$

satisfying

$$u(x,0) = \phi(x), \qquad u(x,T) = \psi(x)$$

in the interval (α,β).

Remark 3.4. In fact we have $h_j \in C^\infty(0,T)$, $j = 1,2,3$, and the corresponding solution

$$u \in C^\infty((\alpha,\beta) \times (0,T)).$$

Proof. For given ψ, $\phi \in H^s(\alpha,\beta)$, let $h = G(\phi,\psi)$ and $w(t) = W_a(t)h$. Then $w \in C([0,T]; H^s(R)) \cap L^2(0,T; H^{s+1}_{loc}(R))$ and by Theorem 3.1,

$$w(x,0) = \phi(x), \qquad w(x,T) = \psi(x)$$

in the interval (α,β). Thus if we let $u(x,t)$ be the restriction of $w(x,t)$ to the domain $[\alpha,\beta]\times[0,T]$, then $u\in C([0,T];H^s(\alpha,\beta))\cap L^2(0,T;H^{s+1}(\alpha,\beta))$ solves (3.16) with the boundary value functions there given by

$$h_1(t)=w(\alpha,t),\qquad h_2(t)=w(\beta,t),\qquad h_2(t)=w_x(\beta,t)$$

and satisfies

$$u(x,0)=\phi(x),\qquad u(x,T)=\psi(x).$$

The proof is complete. □

4. Nonlinear system. Before presenting the proofs for our main results, Theorems 1.1 and 1.2, we consider initial value control of the following nonlinear system described by the KdV equation with variable coefficient:

$$\begin{cases} v_t+v_x+vv_x+(a(x,t)v)_x+v_{xxx}=0, & x\in R,\ t\in R,\\ v(x,0)=h(x). \end{cases}\tag{4.1}$$

Here $a(x,t)$ is assumed to be smooth for simplicity. When the initial value h is considered as a control input, the system possesses the following type of exact controllability.

PROPOSITION 4.1. *Let $s\geq 0$, $T>0$, and $b>0$ as given in Lemma 2.3. In addition, suppose $a\in Y_{s,b}$. Then there exists $\delta>0$ such that if ϕ, $\psi\in H^s(\alpha,\beta)$ with*

$$\|\phi\|_{H^s(\alpha,\beta)}\leq\delta \quad \textit{and} \quad \|\psi\|_{H^s(\alpha,\beta)}\leq\delta,$$

one can find $h\in H^s(R)$ such that the corresponding solution v of (4.1) satisfies

$$v(x,0)=\phi(x),\quad v(x,T)=\psi(x)$$

on the interval (α,β).

Proof. Using the notation of the C^0-group $W_a(t)$ we write (4.1) in its equivalent integral equation form:

$$v(t)=W_a(t)h-\int_0^t W_a(t-\tau)(vv_x)(\tau)d\tau.\tag{4.2}$$

Let

$$\omega(T,v)\equiv\int_0^T W_a(T-\tau)(vv_x)(\tau)d\tau.$$

Then, according to Theorem 3.1, for given ϕ, $\psi\in H^s(\alpha,\beta)$, if one chooses

$$h=G(\phi,\psi+\omega(T,v))$$

in (4.2), then

$$v(t)=W_a(t)G(\phi,\psi+\omega(T,v))-\int_0^t W_a(t-\tau)(vv_x)(\tau)d\tau$$

satisfies

$$v(x,0)=\phi(x),\qquad v(x,T)=\psi(x)$$

on the interval (α, β) by virtue of the definition of the operator G. This suggests that we consider the map

$$\Gamma(v) = W_a(t)G(\phi, \psi + \omega(T, v)) - \int_0^t W_a(t-\tau)(vv_x)(\tau)d\tau.$$

If we can show that the map Γ is a contraction in an appropriate Banach space, then its fixed point v is a solution of (4.1) with $h = G(\phi, \psi + \omega(T, v))$ which satisfies $v(x, 0) = \phi(x)$ and $v(x, T) = \psi(x)$ on the interval (α, β). We show that this is the case in the space $Y_{s,b}$ given earlier in section 2.

To this end, we modify the map Γ as follows:

$$\Gamma(v) = \sigma_1(t)W_a(t)\Phi(\phi, \psi + \omega(T, v)) - \sigma_1(t)\int^t W_a(t-\tau)\sigma_2(\tau)(vv_x)(\tau)d\tau$$

for any $v \in Y_{s,b}$, where $\sigma_1(t)$ is a smooth nonnegative function satisfying $\sigma_1(t) = 1$ for any $t \in (-T, T)$ but with its support inside the interval $(-T - 1/2, T + 1/2)$, and σ_2 is also a nonnegative smooth function with its support contained in $(-T - 1/2, T + 1/2)$ but is identically equal to 1 on the support of σ_1. By Lemma 2.7,

$$\begin{aligned}\|\Gamma(v)\|_{Y_{s,b}} &\le c\|G(\phi, \psi + \omega(T, v))\|_{H^s(R)} + c\|v\|^2_{Y_{s,b}} \\ &\le c(\|\phi\|_{H^s(R)} + \|\psi\|_{H^s(R)} + \|\omega(T, v)\|_{H^s(R)} + c\|v\|^2_{Y_{s,b}}).\end{aligned}$$

As

$$\begin{aligned}\|\omega(T, v)\|_{H^s(R)} &= \left\| \int_0^T W_a(T-\tau)(vv_x)(\tau)d\tau \right\|_{H^s(R)} \\ &\le \sup_{t\in R} \|\sigma_1(t)\int_0^t W_a(t-\tau)\sigma_2(\tau)(vv_x)(\tau)d\tau\|_{H^s(R)} \\ &\le c\|\sigma_1(t)\int_0^t W_a(t-\tau)\sigma_2(\tau)(vv_x)(\tau)d\tau\|_{Y_{s,b}} \\ &\le c\|v\|^2_{Y_{s,b}},\end{aligned}$$

we have

$$\|\Gamma(v)\|_{Y_{s,b}} \le c\left(\|\phi\|_{H^s(\alpha,\beta)} + \|\psi\|_{H^s(\alpha,\beta)}\right) + c\|v\|^2_{Y_{s,b}}.$$

For $M > 0$, let

$$S_M = \{v \in Y_{s,b}; \quad \|v\|_{Y_{s,b}} \le M\}.$$

Then for any $v \in S_M$,

$$\|\Gamma(v)\|_{Y_{s,b}} \le c\left(\|\phi\|_{H^s(\alpha,\beta)} + \|\psi\|_{H^s(\alpha,\beta)}\right) + cM^2$$

for some appropriate constant $c > 0$ independent of v.

Choose $\delta > 0$ and M such that

$$2c\delta + CM^2 \le M, \qquad cM < 1/2. \tag{4.3}$$

Then

$$\|\Gamma(v)\|_{Y_{s,b}} \leq M$$

for any $v \in S_M$ if $\|\phi\|_{H^s(\alpha,\beta)} \leq \delta$, $\|\psi\|_{H^s(\alpha,\beta)} \leq \delta$. In addition, for any $v_1,\ v_2 \in S_M$,

$$\Gamma(v_1) - \Gamma(v_2) = \sigma_1(t)W_a(t)(G(\phi, \psi + \omega(T, v_1)) - G(\phi, \psi + \omega(T, v_2)))$$

$$- \sigma_1(t)\int_0^t W_a(t-\tau)\sigma_2(\tau)\left(\frac{1}{2}\partial_x((v_1+v_2)(v_1-v_2))\right)(\tau)d\tau$$

$$= \quad \sigma_1(t)W_a(t)G(\phi, \omega(T, v_1) - \omega(T, v_2))$$

$$- \sigma_1(t)\int_0^t W_a(t-\tau)\sigma_2(\tau)\left(\frac{1}{2}\partial_x((v_1+v_2)(v_1-v_2))\right)(\tau)d\tau$$

and

$$\omega(T, v_1) - \omega(T, v_2) = \int_0^T W_a(T-\tau)\left(\frac{1}{2}\partial_x((v_1+v_2)(v_1-v_2))\right)(\tau)d\tau.$$

A similar argument shows that

$$\begin{aligned}\|\Gamma(v_1) - \Gamma(v_2)\|_{Y_{s,b}} &\leq \frac{1}{2}\|v_1+v_2\|_{Y_{s,b}}\|v_1-v_2\|_{Y_{s,b}} \\ &\leq cM\|v_1-v_2\|_{Y_{s,b}} \\ &\leq \frac{1}{2}\|v_1-v_2\|_{Y_{s,b}}.\end{aligned}$$

Thus the map Γ is a contraction on S_M provided that δ and M are chosen according to (4.3). As a result, its fixed point $v \in S_M$ is a solution of the integral equation

$$v(t) = \sigma_1(t)W_a(t)G(\phi, \psi + \omega(T, v)) - \sigma_1(t)\int^t \sigma_2(\tau)W_a(t-\tau)(vv_x)(\tau)d\tau$$

for any $t \in R$. In particular, for $t \in (0, T)$,

$$v(t) = W_a(t)G(\phi, \psi + \omega(T, v)) - \int_0^t W_a(t-\tau)(vv_x)(\tau)d\tau.$$

That is to say, $v \in C([0, T]; H^s(R))$ solves the IVP

$$\begin{cases} v_t + vv_x + v_x + (a(x,t)v)_x + v_{xxx} = 0, \\ v(x, 0) = G(\phi, \psi + \omega(T, v)) = h(x) \end{cases}$$

for any $t \in (0, T)$ and satisfies

$$v(x, 0) = \phi(x), \qquad v(x, T) = \psi(x), \qquad \text{on the interval } (\alpha, \beta).$$

The proof is complete. □

Now we turn to the proof of the main theorems described in the introduction of this paper.

Proof of Theorem 1.1. Let u be a solution of (1.5)–(1.6) satisfying

$$u(x,0) = \phi(x)$$

and let

$$z(x,t) = u(x,t) - w(x,t),$$

where w is as given in the assumption. Then $z(x,t)$ solves

$$\begin{cases} z_t + zz_x + z_x + (w(x,t)z)_x + z_{xxx} = 0, \\ z(x,0) = \phi(x) - w(x,0), \\ z(\alpha,t) = \tilde{h_1}(t),\ z(\beta,t) = \tilde{h_2}(t),\ z_x(\beta,t) = \tilde{h_3}(t) \end{cases} \tag{4.4}$$

with

$$\tilde{h_1}(t) = h_1(t) - w(\alpha,t), \quad \tilde{h_2}(t) = h_2(t) - w(\beta,t), \quad \tilde{h_3}(t) = h_3(t) - w_x(\beta,t).$$

Let $\xi(x,t)$ be a smooth function of x and t with compact support in $(\alpha_1,\beta_1)\times(-T,T+1)$ and $\xi(x,t) = 1$ for any $(x,t) \in (\alpha,\beta)\times[0,T]$. Setting $a(x,t) = \xi(x,t)w(x,t)$, we consider the following KdV equation posed on the whole line R:

$$v_t + v_x + vv_x + (a(x,t)v)_x + v_{xxx} = 0, \quad v(x,0) = h(x). \tag{4.5}$$

Note that $a \in Y_{s,b}$ by its definition. It follows from Proposition 4.1 that there exists a $\delta > 0$ such that if $\|(\phi(\cdot) - w(\cdot,0))\|_{H^s(\alpha,\beta)} \le \delta$ and $\|(\psi(\cdot) - w(\cdot,T))\|_{H^s(\alpha,\beta)} \le \delta$, then one can find an $h \in H^s(R)$ such that the corresponding solution v of (4.5) satisfies

$$v(x,0) = \phi(x) - w(x,0), \qquad v(x,T) = \psi(x) - w(x,T)$$

on the interval (α,β). When restricted to the interval (α,β), $a(x,t) \equiv w(x,t)$ and $v(x,t)$ solves (4.4) on the domain $(\alpha,\beta)\times(0,T)$ with

$$h_1(t) = v(\alpha,t) + w(\alpha,t), \quad h_2(t) = v(\beta,t) + w(\beta,t), \quad h_3(t) = v_x(\beta,t) + w_x(\beta,t).$$

Consequently, $u = z(x,t) + w(x,t)$ is the desired solution we are looking for. The proof is complete. □

Proof of Theorem 1.2. Let

$$\tilde{\phi}(x) = \phi(x) - w(x,0) \quad \text{and} \quad \tilde{\psi}(x) = \psi(x) - w(x,T).$$

Applying Proposition 4.1 to

$$v_t + v_x + vv_x + (a(x,t)v)_x + v_{xxx} = 0, \quad v(x,0) = \tilde{h}(x) \tag{4.6}$$

for $x,\ t \in R$, where $a(x,t) = \sigma(t)w(x,t)$ and $\sigma(t)$ is a smooth function with compact support and $\sigma(t) = 1$ for $t \in [0,T]$, yields that there exists an $\tilde{h} \in H^s(R)$ such that (4.6) has a solution v satisfying

$$v(x,0) = \tilde{\phi}(x) \quad \text{and} \quad v(x,T) = \tilde{\psi}(x)$$

on the interval (α, β). Consequently, if we let

$$h(x) = \tilde{h}(x) + w(x, 0)$$

in (1.9), then the corresponding solution u satisfies

$$u(x, 0) = \phi(x) \quad \text{and} \quad u(x, T) = \psi(x)$$

on the interval (α, β). The proof is complete. □

Acknowledgments. The author is grateful for the hospitality offered by Institut de Recherche Mathématique Avancée, Université Louis Pasteur et CNRS, and Institut Elie Cartan, Université de Nancy I, where the paper was completed. The author is also grateful to Professors W. Littman, V. Komornik, and F. Conrad for various stimulating discussions.

REFERENCES

[1] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, *The inverse scattering transform—Fourier analysis for nonlinear problems,* Stud. Appl. Math., 53 (1974), pp. 249–315.

[2] J. L. Bona and R. Smith, *The initial value problem for the Korteweg-de Vries equation,* Philos. Trans. Roy. Soc. London Ser. A, 278 (1975), pp. 555–601.

[3] J. L. Bona and B.-Y. Zhang, *The initial-value problem for the forced Korteweg-de Vries equation,* Proc. Roy. Soc. Edinburgh Ser. A, 126 (1996), pp. 571–598.

[4] J. Bourgain, *Fourier transform restriction phenomena for certain lattice subsets and applications to non-linear evolution equations, part* I: *Schrödinger equations,* Geom. Funct. Anal., 3 (1993), pp. 107–156.

[5] J. Bourgain, *Fourier transform restriction phenomena for certain lattice subsets and applications to non-linear evolution equations, part* II: *The KdV equation,* Geom. Funct. Anal., 3 (1993), pp. 209–262.

[6] A. Cohen, *Solutions of the Korteweg-de Vries equation from irregular data,* Duke Math. J., 45 (1978), pp. 149–181.

[7] P. Constantin and J.-C. Saut, *Local smoothing properties of dispersive equations,* J. Amer. Math. Soc., 1 (1988), pp. 413–446.

[8] W. Craig, T. Kappeler, and W. A. Strauss, *Gain of regularity for equations of KdV type,* Ann. Inst. H. Poincaré Anal. Non Linéaire, 9 (1992), pp. 147–186.

[9] C. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, *Method for solving the Korteweg-de Vries equation,* Phys. Rev. Lett., 19 (1967), pp. 1095–1097.

[10] T. Kato, *On the Cauchy problem for the (generalized) Korteweg-de Vries equations,* in Advances in Mathematics Supplementary Studies, Stud. Appl. Math. 8, Academic Press, New York, 1983, pp. 93–128.

[11] C. E. Kenig, G. Ponce, and L. Vega, *Well-posedness of the initial value problem for the KdV equation,* J. Amer. Math. Soc., 4 (1991), pp. 323–347.

[12] C. E. Kenig, G. Ponce, and L. Vega, *Well-posedness and scattering results for the generalized Korteweg-de Vries equations via the contraction principle,* Comm. Pure Appl. Math., 46 (1993), pp. 527–620.

[13] C. E. Kenig, G. Ponce, and L. Vega, *The Cauchy problem for the Korteweg-de Vries equation in Sobolev spaces of negative indices,* Duke Math. J., 71 (1993), pp. 1–21.

[14] C. E. Kenig, G. Ponce, and L. Vega, *A bilinear estimate with applications to the KdV equation,* J. Amer. Math. Soc., 9 (1996), pp. 573–603.

[15] V. Komornik, D. L. Russell, and B.-Y. Zhang, *Stabilisation de l'equation de Korteweg-de Vries,* C. R. Acad. Sci. Paris Sér. I Math., 312 (1991), pp. 841–843.

[16] D. J. Korteweg and G. deVries, *On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves,* Philos. Mag., 39 (1895), pp. 422–443.

[17] W. Littman, *Boundary control theory for hyperbolic and parabolic partial differential equations with constant coefficients,* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 5 (1978), pp. 567–580.

[18] W. Littman and S. Taylor, *Smoothing evolution equations and boundary control theory. Festschrift on the occasion of the* 70*th birthday of Shmuel Agmon,* J. Anal. Math., 59 (1992), pp. 117–131.

[19] R. M. Miura, *The Korteweg-de Vries equation: A survey of results,* SIAM Rev., 18 (1976), pp. 412–459.

[20] A. Pazy, *Semigroup of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.

[21] L. Rosier, *Exact boundary controllability for the Korteweg-de Vries equation on a bounded domain,* ESIAM Control Optim. Cal. Var., 2 (1997), pp. 33–55.

[22] D. L. Russell, *Boundary value control of the higher-dimensional wave equation,* SIAM J. Control Optim., 9 (1971), pp. 29–42.

[23] D. L. Russell, *Boundary value control theory of the higher-dimensional wave equation part* II, SIAM J. Control Optim., 9 (1971), pp. 401–419.

[24] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions,* SIAM Rev., 20 (1978), pp. 639–739.

[25] D. L. Russell and B.-Y. Zhang, *Controllability and stabilizability of the third-order linear dispersion equation of a periodic domain,* SIAM J. Control Optim., 31 (1993), pp. 659–676.

[26] D. L. Russell and B.-Y. Zhang, *Smoothing and decay properties of solutions of the Korteweg-de Vries equation on a periodic domain with point dissipation,* J. Math. Anal. Appl., 190 (1995), pp. 449–488.

[27] D. L. Russell and B.-Y. Zhang, *Exact controllability and stabilizability of the Korteweg-de Vries equation,* Trans. Amer. Math. Soc., 348 (1996), pp. 3643–3672.

[28] J. C. Saut and R. Temam, *Remarks on the Korteweg-de Vries equation,* Israel J. Math., 24 (1976), pp. 78–87.

[29] R. Temam, *Sur un problème non linéaire,* J. Math. Pures Appl., 48 (1969), pp. 159–172.

[30] B.-Y. Zhang, *Unique continuation for the Korteweg–De Vries equation,* SIAM J. Math. Anal., 23 (1992), pp. 55–71.

[31] B.-Y. Zhang, *Taylor series expansion for solutions of the Korteweg-de Vries equation with respect to their initial values,* J. Funct. Anal., 129 (1995), pp. 293–324.

[32] B.-Y. Zhang, *Analyticity of solutions of the generalized Korteweg–de Vries equations with respect to their initial values,* SIAM J. Math. Anal., 26 (1995), pp. 1488–1513.

[33] B.-Y. Zhang, *A remark on the Cauchy problem of the periodic Korteweg-de Vries equation,* Differential Integral Equations, 8 (1995), pp. 1191–1204.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 2, pp. 566–588

CENTRAL PATHS, GENERALIZED PROXIMAL POINT METHODS, AND CAUCHY TRAJECTORIES IN RIEMANNIAN MANIFOLDS*

ALFREDO N. IUSEM†, B. F. SVAITER†, AND JOÃO XAVIER DA CRUZ NETO‡

Abstract. We study the relationships between three concepts which arise in connection with variational inequality problems: central paths defined by arbitrary barriers, generalized proximal point methods (where a Bregman distance substitutes for the Euclidean one), and Cauchy trajectory in Riemannian manifolds. First we prove that under rather general hypotheses the central path defined by a general barrier for a monotone variational inequality problem is well defined, bounded, and continuous and converges to the analytic center of the solution set (with respect to the given barrier), thus generalizing results which deal only with complementarity problems and with the logarithmic barrier. Next we prove that a sequence generated by the proximal point method with the Bregman distance naturally induced by the barrier function converges precisely to the same point. Furthermore, for a certain class of problems (including linear programming), such a sequence is contained in the central path, making the concepts of central path and generalized proximal point sequence virtually equivalent. Finally we prove that for this class of problems the central path also coincides with the Cauchy trajectory in the Riemannian manifold defined on the positive orthant by a metric given by the Hessian of the barrier (i.e., a curve whose direction at each point is the negative gradient of the objective function at that point in the Riemannian metric).

Key words. convex programming, linear programming, variational inequalities, complementarity problems, interior point methods, central path, generalized distances, proximal point method, Riemannian manifolds

AMS subject classifications. 90C25, 90C30

PII. S0363012995290744

1. Introduction.

We study in this paper the connection among three concepts which arise in relation to convex optimization problems and, more generally, monotone variational inequality problems.

These three concepts are central paths derived from barrier functions, proximal point algorithms with generalized distances, and Cauchy trajectories in Riemannian manifolds. In recent years these notions have been the object of intense study, and many results have been obtained regarding each of them in an independent way.

We will prove that in some cases, including linear programming, these three concepts are in a certain way equivalent, opening the road to a process of cross-fertilization, exchanging the results obtained with each approach. For other problems this equivalence breaks down, but there are nevertheless striking connections, which will be brought forth in our discussion.

It happens to be the case that though many results obtained for these concepts are quite similar, the technical hypotheses made in their proofs are somewhat different. As a consequence, the comparison among results becomes difficult. To overcome this obstacle, we present in section 2 a development of the concept and convergence properties of central paths which generalizes previous expositions both with regard to the type of problems being considered and to the class of barrier functions.

*Received by the editors August 23, 1995; accepted for publication (in revised form) December 18, 1997; published electronically January 28, 1999. The research of the first author was partially supported by CNPq grant 301280/86.

http://www.siam.org/journals/sicon/37-2/29074.html

†Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina, 110, Rio de Janeiro, RJ, CEP 22460-320, Brazil (iusp@impa.br, benar@impa.br).

‡Departamento de Matemática, CNN, Universidade Federal do Piauí, Teresina, PI, CEP 64000-00, Brazil (jxavier@ufpi.br).

We start with an informal presentation of these three concepts, beginning with central paths. Given a monotone operator $T : \mathbf{R}^n \to \mathcal{P}(\mathbf{R}^n)$ and a closed and convex set $C \subset \mathbf{R}^n$, the variational inequality problem VIP(T,C) consists of finding $z \in C$ such that, for some $u \in T(z)$, it holds that $\langle u, z - x\rangle \geq 0$ for all $x \in C$ (see [Kin]). Assuming that the interior C^o of C is nonempty, we consider a *barrier* h for C. The function $h : C^o \to \mathbf{R}$ is assumed to be strictly convex and differentiable, and its gradient ∇h must diverge at the boundary ∂C of C. For $\mu \in \mathbf{R}_{++}$ let $x(\mu) \in C$ be such that

$$-\frac{1}{\mu}\nabla h(x(\mu)) \in T(x(\mu)). \tag{1}$$

The set $\{x(\mu) : \mu > 0\}$ is called the *central path* for VIP(T,C) with barrier h. We analyze in section 2 conditions on T, C, and h guaranteeing existence and uniqueness of the central path, which is contained in C^o. (For this reason h is said to be a barrier for VIP(T,C).) If $T = \partial f$ for a convex $f : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$, then (1) is equivalent to

$$x(\mu) = \operatorname*{argmin}_{x \in C}\{\mu f(x) + h(x)\}. \tag{2}$$

The problem of interest is the behavior of $x^* = \lim_{\mu\to\infty} x(\mu)$. One expects this limit to exist and to be a solution of VIP(T,C). We will prove in section 2 that this result holds under quite general hypotheses. Another issue to be considered is the characterization of such a limit in a more precise way, particularly when VIP(T,C) has multiple solutions. Here we have to consider two possibilities. Let $S(T, C)$ be the set of solutions of VIP(T,C). We will prove that if the effective domain ED(h) of h intersects $S(T, C)$, then x^* is the solution of

$$\min h(x) \tag{3}$$

$$\text{subject to } x \in S(T, C). \tag{4}$$

The solution of this problem (unique if it exists, by strict convexity of h) is called the *analytic center* of $S(T, C)$ with respect to the barrier h.

Most of the work concerned with central paths for linear programming use the logarithmic barrier $h(x) = -\sum_{j=1}^n \log x_j$, which in general diverges at all optimal solutions, since they belong to the boundary of $\mathbf{R}^n_+$. For cases such as this, we are still able to provide a characterization of x^* in some special cases. Theorem 2 of this paper says that x^* is the solution of

$$\min \tilde{h}(x) \tag{5}$$

$$\text{s.t. } x \in S(T, C) \tag{6}$$

for some appropriate $\tilde{h}$. This theorem covers the cases of linear programming, convex quadratic programming, some monotone linear complementarity problems, and also some cases of linearly constrained nonlinear convex optimization problems and nonlinear complementarity problems.

At this point, it is convenient to relate these results with the existing literature on central paths. The concept of central path has aroused a keen interest in connection with interior point methods for linear programming, because sequences which lie close enough to the central path with respect to the logarithmic barrier $h(x) = -\sum_{j=1}^{n} \log x_j$ enjoy very interesting convergence properties (see [Gon]). The concept can be traced back to [McL], where a very special case is considered. The central path with the logarithmic barrier for the linear programming problem has been considered in [Meg1], and its convergence to the analytic center of the solution set has been proved in [Meg2]. [Adl2] considers the weighted logarithmic barrier $h(x) = -\sum_{j=1}^{n} \sigma_j \log x_j$, with $\sigma_j > 0$, and proves furthermore that $\lim_{\mu \to \infty} \dot{x}(\mu)$ exists and can be well characterized, again in the case of linear programming. The central path with logarithmic barrier for linear complementarity problems is considered in [Koj] and [Mon]. The fact that the central path with logarithmic barrier coincides with the affine scaling path in the case of linear programming has been established in [Bay] and is a particular case of Theorem 4 (see section 4). The central path with the weighted logarithmic barrier for the nonlinear complementarity problem has been studied in [Gul], where it is proved that under adequate assumptions the path is continuous and bounded and that its cluster points are solutions of the problem.

Now we discuss generalized proximal point methods for variational inequality problems. We start with a strictly convex function $g : C \to \mathbf{R}$, continuously differentiable in C^o, and we define $D_g : C \times C^o \to \mathbf{R}$ as $D_g(x, y) = g(x) - g(y) - \langle \nabla g(y), x - y \rangle$. Several technical hypotheses are made on g. (One of them is akin to divergence of ∇g at ∂C; see section 3.) The generalized proximal point algorithm (GPPA) for VIP(T,C) using D_g generates a sequence $\{x^k\} \subset C^o$ in the following way. It starts with $x^0 \in C^o$, and given x^k it defines the operator T_k as

$$T_k(x) = T(x) + \lambda_k \partial_1 D_g(x, x^k), \tag{7}$$

where λ_k is an exogenously given positive number (called the regularization parameter) and ∂_1 indicates the subdifferential with respect to the first argument. Then x^{k+1} is chosen so that

$$0 \in T_k(x^{k+1}). \tag{8}$$

Existing results on this method, established in [Bur2] and [Ius3], are formally presented in section 3, where we study the connection between central paths and GPPA sequences. We prove that the GPPA sequence $\{x^k\}$ converges to the same point as the central path with barrier $h(x) = D_g(x, x^0)$. One could be tempted to conjecture that the sequence $\{x^k\}$ is contained in this central path. We prove that this is the case for linear programming and, more generally, for problems of the type min $c^t x$ s.t. $Ax = b, x \in C$, where C is any closed and convex set with nonempty interior. Moreover, any sequence $\{x(\mu_k)\}$ with increasing $\{\mu_k\}$ can be realized as a GPPA sequence $\{x^k\}$ for a specific choice of the regularization parameters λ_k, making the concepts of central path and GPPA sequence virtually identical in this case and, in particular, for linear programming problems.

On the other hand, for other problems for which convergence to the analytic center holds both for central paths and GPPA sequences (e.g., convex quadratic programming, monotone linear complementarity problems, etc.), the GPPA sequence is not contained in the central path, though both converge to the same point.

We summarize now the history of GPPA. It is an extension of the classical proximal point method (see [Lem]), with $D_g(x, y)$ substituting for $\|x - y\|^2$. It was considered first only for the case of linear programming with $g(x) = \sum_{j=1}^{n} x_j \log x_j$ in [Eri] and then for the same g and general convex optimization problems in [Egg]. The next four steps also dealt only with the convex optimization case ($T = \partial f$ with convex f). In [Cen1] a general g is allowed, but it is assumed that there exist solutions in C^o (so that the constraint set C is superfluous). This hypothesis was removed in [Che], which imposed an additional condition on g (see H4 in section 2) and then in [Ius1], which resulted in a weaker condition (see H2 in section 2). More recently, the method was further generalized in [Kiw2], the results of which are discussed in section 3.

Moving beyond the convex optimization case, the method was studied in [Eck] for the problem of finding zeros of monotone operators and later on in [Bur1], [Bur2] for the more general case of variational inequality problems.

Finally, we discuss in section 4 Cauchy trajectories in Riemannian manifolds. Given a function f defined on a Riemannian manifold M, it is possible to define its gradient grad f with respect to the metric of M, and a *Cauchy trajectory* $x(t)$ is a curve contained in M such that $\dot{x}(t) = -\text{grad } f(x(t))$ for all t. We prove that the central paths introduced in section 2 can be seen in certain cases as Cauchy trajectories in a Riemannian manifold, whose metric is given by the Hessian matrix of the barrier h. The ability of looking at central paths in this way has given, in the case of linear programming with the logarithmic barrier, new insights into the properties of the path. For instance, computational complexity of path following methods has been related to the total Riemannian curvature of the central path (see [Kar]). We hope that our results for general barriers will give rise to similar new insights.

The Cauchy trajectory, also called *gradient trajectory*, has been studied in connection with algorithms for optimization problems in Riemannian manifolds; see, e.g., [Hel].

2. Central paths with general barriers for variational inequalities.

The main goal of this section is to prove that, under quite general hypotheses, the central path $\{x(\mu)\}$ for VIP(T,C) with barrier h converges, as μ goes to ∞, to a solution of VIP(T,C) and that, under more restrictive assumptions, the limit of $\{x(\mu)\}$ is the analytic center of the solution set $S(T,C)$. The set C is assumed to be closed and convex, with nonempty interior C^o.

Next we enumerate the various assumptions on h, T, and C which will be used in this section. Some of them are required in all our results, while others are alternative hypotheses required for some specific results. The first group of assumptions does not involve the operator T. $h : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ is assumed to satisfy the following conditions:

H1. h is strictly convex, closed, and continuous on its effective domain and continuously differentiable in C^o.

H2. If $\{x^k\}$ is a sequence contained in C^o which converges to a point $\bar{x}$ in the boundary ∂C of C, and y is any point in C^o, then $\lim_{k\to\infty}\langle \nabla h(x^k), y - x^k\rangle = -\infty$.

These two assumptions are needed in practically all our results. H2, called *boundary coerciveness*, has been introduced in [Ius1]. It is closely connected to *essential smoothness*, as defined in [Roc1].

Some of our results will require the following additional assumptions:

H3. h is finite, continuous, and strictly convex on C.

We will say that a function satisfying H3 is *finite at the boundary.*

H4. For all $y \in \mathbf{R}^n$ there exists $x \in C^o$ such that $\nabla h(x) = y$.

H3, which entails finiteness of h at any solution of VIP(T,C), simplifies the proof of convergence of the central path to the analytic center. In its absence, the analytic center has to be defined in a more complicated way, with an auxiliary function $\tilde{h}$

instead of h. H4, called *zone coerciveness*, has been introduced in [Cen2], where the relations among H2, H4, and essential smoothness are treated in detail. It has been proved in [Cen2] that when h is twice continuously differentiable with nonsingular Hessian matrix at any point in C^o, H4 implies H2 and they are equivalent when C is bounded. With H4 it is possible to prove existence of the central path without any hypotheses on T beyond monotonicity.

For the case in which C is a *generalized box*, i.e., $C = [\alpha_1, \beta_2] \times \cdots \times [\alpha_n, \beta_n]$, with $\alpha_j \in [-\infty, \infty)$, $\beta_j \in (-\infty, \infty]$, $\alpha_j < \beta_j$, we also consider the following condition, called *separability*:

H5. $h(x) = \sum_{j=1}^n h_j(x_j)$ with $h_j : (\alpha_j, \beta_j) \to \mathbf{R}$.

H5 will be used to prove convergence of the central path to the analytic center, in the absence of H3.

For the case of $C = \mathbf{R}^n_+$, two examples of barrier functions are $h(x) = -\sum_{j=1}^n \log x_j$, which satisfies H1, H2, and H5 but neither H3 nor H4, and $h(x) = \sum_{j=1}^n x_j \log x_j$, which satisfies H1–H5 (with the convention that $0 \log 0 = 0$).

The second group of hypotheses refers to the operator $T : \mathbf{R}^n \to \mathcal{P}(\mathbf{R}^n)$. The first two of them, which are needed in most of our results, are the following:

H6. T is *maximal monotone*; i.e., it is monotone, meaning that $\langle u - v, x - y\rangle \geq 0$ for all $x, y \in \mathbf{R}^n$, $u \in T(x)$, $v \in T(y)$, and if $T(x) \subset T'(x)$ for some monotone operator T' and all $x \in \mathbf{R}^n$, then $T(x) = T'(x)$ for all $x \in \mathbf{R}^n$.

H7. T is *paramonotone*, meaning that it is maximal monotone and additionally $\langle u - v, x - y\rangle = 0$ with $u \in T(x)$, $v \in T(y)$ implies $u \in T(y)$, $v \in T(x)$.

Paramonotonicity has been introduced in [Cen2]. The main properties of paramonotone operators are summarized in the following proposition. For a matrix $B \in \mathbf{R}^{n \times n}$, B^s will denote its symmetric part, i.e., $B^s = 1/2(B + B^t)$ and $rk(B)$ its rank. The Jacobian matrix of a point-to-point and differentiable operator U at a point x will be denoted by $J_U(x)$.

Proposition 1.

i) *If $T = \partial f$ for some convex $f : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$, then T is paramonotone.*

ii) *If T is paramonotone, $z \in S(T, C)$ and $\langle u, z - x\rangle \geq 0$ for some $x \in C$ and some $u \in T(x)$, then $x \in S(T, C)$.*

iii) *If T_1 and T_2 are paramonotone, then $T_1 + T_2$ is paramonotone.*

iv) *If T is point-to-point and differentiable, $J_T(x)$ is positive semidefinite (not necessarily symmetric) and* $\operatorname{Ker}(J_T(x)) = \operatorname{Ker}(J_T(x)^s)$ *for all $x \in \mathbf{R}^n$, then T is paramonotone.*

v) *If T is point-to-point and differentiable, $J_T(x)$ is positive semidefinite and $rk(J_T(x)) = rk(J_T(x)^s)$ for all $x \in \mathbf{R}^n$, then T is paramonotone.*

Proof. See [Ius2, Propositions 1, 2, and 6]. □

The following assumption is needed to prove existence of the central path in the absence of H4. Before presenting it, we recall certain known facts on normal cones. For a closed, convex, and nonempty set $V \subset \mathbf{R}^n$, let δ_V be the indicator function of V, i.e.,

$$\delta_V(x) = \begin{cases} 0 & \text{if } x \in V, \\ +\infty & \text{otherwise.} \end{cases}$$

We define the normality operator N_V of V as $N_V(x) = \partial\delta_V(x)$ (i.e., the set of subgradients of δ_V at x). It is easy to check that $N_V(x) = \emptyset$ if $x \notin V$; $N_V(x) = \{0\}$ if $x \in V^o$; $N_V(x)$ is a positive cone for all x; and, given $x \in V$, $v \in N_V(x)$ if and only if

$$\langle v, x' - x\rangle \leq 0. \tag{9}$$

The assumption is the following.

H8. $T = \hat{T} + N_V$ for some point-to-point, continuous, and paramonotone $\hat{T} : \mathbf{R}^n \to \mathbf{R}^n$ and some nonempty, closed, and convex $V \subset \mathbf{R}^n$.

The rationale behind H8 is the following. It is easy to check (e.g., [Roc2]) that the set of solutions of VIP(T,C) is the set of zeros of $T + N_C$ and more generally that

$$S(T, V \cap C) = S(T + N_V, C). \tag{10}$$

Now, in many cases the problem of interest is a variational inequality problem with a well-behaved operator T in a constraint set with empty interior, which makes the introduction of barrier functions difficult. The trick to avoid this obstacle is to write the constraint set as $C \cap V$ with $C^o \neq \emptyset$ and then to transfer the additional constraints, given by V, to the operator, as in (10). For instance, for linear programming we take $C = \mathbf{R}^n_+$, $V = \{x \in \mathbf{R}^n : Ax = b\}$. It is easy to check that for this V we have $N_V(x) = \mathrm{Im}(A^t)$ for all $x \in \mathbf{R}^n$. Our proofs would become much simpler assuming that T is point-to-point and continuous, or at least bounded over bounded sets, but then we lose the option of including affine constraints in the operator.

It follows from Proposition 1(i) and (iii) and the definition of N_V that H8 implies paramonotonicity of T (see H7). It can be easily checked that H8 also implies maximal monotonicity of T (see H6).

The following assumption is needed in order to prove convergence to the analytic center without finiteness of h at the boundary (see H3).

H9. Let $\hat{T}$ and V be as given in H8. V is an affine manifold, $\hat{T}$ is continuously differentiable, and there exists a subspace W such that $\mathrm{Ker}(J_{\widehat{T}}(x)) = W$ for all $x \in C \cap V$.

As examples for which H9 holds we have linear programming, in which case $\hat{T}(x) = c \in \mathbf{R}^n$ so that $J_{\widehat{T}}(x) = 0$ and $W = \mathbf{R}^n$; monotone linear complementarity problems (including convex quadratic programming problems), in which case $\hat{T}(x) = Qx$ with $Q \in \mathbf{R}^{n \times n}$ and $W = \mathrm{Ker}(Q)$; and also many cases of nonlinear operators. Among them we can mention the gradients of self-concordant functions, introduced in [Nes], which satisfy H9 by [Nes, Corollary 2.1.1]. Another case of interest is $\hat{T}(x) = A^t \nabla f(Ax)$, with f twice continuously differentiable and $\nabla^2 f(x)$ positive definite for all $x \in C \cap V$. Then $J_{\widehat{T}}(x) = A^t \nabla^2 f(Ax) A$ and it is easy to check that H9 holds with $W = \mathrm{Ker}(A)$. This case occurs in maximum likelihood estimation problems. A proof of convergence of the central path to the analytic center of the solution set with respect to a special smoothing barrier for this type of problem can be found in [Ius4], from which we took some of the main ideas for this paper.

Finally the third group of assumptions involve both T and C (i.e., VIP(T,C)) and in some cases also the barrier h. First we have two basic and elementary assumptions on VIP(T,C). Let $\mathrm{dom}(T) = \{x \in \mathbf{R}^n : T(x) \neq \emptyset\}$.

H10. $\mathrm{dom}(T) \cap C^o \neq \emptyset$.

A variational inqulity problem satisfying H10 will be said to be *regular*.

H11. $S(T, C) \neq \emptyset$.

We remark that in the presence of H8, H10 becomes equivalent to $V \cap C^o \neq \emptyset$, because $\mathrm{dom}(T) = \mathrm{dom}(\hat{T}) \cap \mathrm{dom}(N_V)$, $\mathrm{dom}(\hat{T}) = \mathbf{R}^n$, and $\mathrm{dom}(N_V) = V$.

Next we need two assumptions in order to guarantee existence of the central path in the absence of zone coerciveness (see H4).

H12. h attains its minimum on $\mathrm{dom}(T) \cap C^o$ at some point $\tilde{x}$.

H12 holds automatically when $h(x) = D_g(x, \tilde{x})$ for some $\tilde{x} \in C^o$ and D_g is a Bregman function (see section 3). The next assumption requires introduction of the gap function $q_{T,C} : \mathbf{R}^n \to \mathbf{R}$ for VIP(T,C), defined as

$$q_{T,C}(x) := \sup_{(v,y)\in G_C(T)} \langle v, x-y\rangle, \tag{11}$$

where $G_C(T) := \{(v,y) : v \in T(y), y \in C\}$. Note that, taking $y = x$ in (11), we get that $q_{T,C}(x) \geq 0$ for all $x \in \mathrm{dom}(T) \cap C$. The assumption on the gap function is the following:

H13. $q_{T,C}(x) < \infty$ for all $x \in \mathrm{dom}(T) \cap C$.

Similar gap functions for variational inequality problems have been considered, e.g., in [Mar] and [Ngu]. We take our $q_{T,C}$ from [Bur1], where it has been proved that H13 holds, e.g., when T is coercive, when T is the subdifferential of a convex function bounded below, when T is strongly monotone, or when $\mathrm{dom}(T) \;\cap\; C$ is bounded (see [Bur1, Proposition 3.1]). It follows easily from (17) that q is convex and that $q_{T,C}(x) \geq 0$ for all $x \in C$. Also, it is easy to check, and is proved in [Bur1] that in the presence of H10 and H11, $q_{T,C}(x) = 0$ if and only if $x \in S(T,C)$. Finiteness of $q_{T,C}$ is a regularity condition which ensures that $T + \partial h$ has zeros for any strictly convex h which attains its global minimizer, even when ∂h is not onto and T does not have zeros (see Proposition 2(ii)).

The last assumption is needed to ensure boundedness of the central path without H4.

H14. Either

i) $S(T,C)$ is bounded and $T = \partial f$ for some convex $f : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$,

ii) there exists $y \in C^o \cap \mathrm{dom}(T)$ such that $\lim_{k\to\infty}\langle u^k, x^k - y\rangle = \infty$ for all $\{x^k\}$ such that $\lim_{k\to\infty}\left\|x^k\right\| = \infty$ and all $\{u^k\}$ such that $u^k \in T(x^k)$, or

iii) $\mathrm{dom}(T) \cap C$ is bounded.

H14(ii) holds, e.g., when T is coercive, though it is much weaker than coerciveness. We conjecture that H14 can be replaced, in the proof of boundedness of the central path, by the much simpler and weaker hypothesis of boundedness of $S(T,C)$, but we have not been able to prove the result under this weaker assumption.

We end this preliminary discussion with a few technical lemmas on monotone operators.

LEMMA 1. *If h satisfies* H1 *and is either zone or boundary coercive* (*see* H4 *and* H2, *respectively*), *then dom*$(\partial h) = C^o$.

Proof. See [Bur2, Lemma 1]. □

LEMMA 2. i) *If T_1 and T_2 are maximal monotone operators such that dom*$(T_1) \cap$ *dom*$(T_2)^o \neq \emptyset$, *then $T_1 + T_2$ is maximal monotone.*

ii) *If furthermore T_2 is the subdifferential of a closed and proper convex function, and T_2 is onto, then $T_1 + T_2$ is onto.*

Proof. See [Bur2, Corollary 1]. □

LEMMA 3. *If T is maximal monotone and dom*(T) *is bounded, then T is onto.*

Proof. See [Brez, Corollary 2.2]. □

Finally, after all these preliminaries, we start the study of the central path. For an operator $T : \mathbf{R}^n \to \mathcal{P}(\mathbf{R}^n)$, a convex function h, and a scalar $\mu > 0$, let $T_\mu = \mu T + \partial h$.

PROPOSITION 2. *Assume that h satisfies* H1, *T is maximal monotone* (*see* H6), *VIP(T,C) is regular* (*see* H10), *and either*

i) *h is zone coercive* (*see* H4), *or*

ii) *h is boundary coercive* (*see* H2) *and attains its minimum on dom*$(T) \cap C^o$ (*see* H12), *the gap function is finite everywhere* (*see* H13), *and T satisfies* H8.
Then the operator T_μ has a unique zero $x(\mu)$ which belongs to dom$(T) \cap C^o$.

Proof. i) Let $T_1 = \mu T$, $T_2 = \partial h$. By Lemma 1, $\mathrm{dom}(T_2) = C^o$. By regularity, $\mathrm{dom}(T_1) \cap \mathrm{dom}(T_2)^o \neq \emptyset$. By Lemma 2(i), using H1 and maximal monotonicity of T, $T_\mu = T_1 + T_2$ is maximal monotone. By zone coerciveness of h, T_2 is onto. By Lemma 2(ii) T_μ is onto; i.e., there exists $x \in \mathrm{dom}(T_\mu)$ such that $0 \in T_\mu(x)$. Since h is strictly convex, ∂h is strictly monotone, and therefore, by monotonicity of T, T_μ is strictly monotone, so that the zero is unique. Note that $x \in \mathrm{dom}(T_\mu) = \mathrm{dom}(T_1) \cap \mathrm{dom}(T_2) = \mathrm{dom}(T_1) \cap C^o$. The result holds.

ii) Now, in the absence of zone coerciveness, ∇h is not onto and the argument is more involved. Let $\tilde{x}$ be the minimizer of h in $\mathrm{dom}(T) \cap C^o$ (which exists by H12). By H13, $q_{T,C}(\tilde{x}) < \infty$. We consider two cases:

a) $q_{T,C}(\tilde{x}) = 0$. In this case, as mentioned above, $\tilde{x} \in S(T, C)$. Since $\tilde{x} \in C^o$, it follows that $0 \in T(\tilde{x})$; i.e., using H8,

$$0 = \hat{T}(\tilde{x}) + v \tag{12}$$

with $v \in N_V(\tilde{x})$. On the other hand, since $\tilde{x}$ minimizes h in $\mathrm{dom}(T)$ and $\tilde{x}$ belongs to C^o, $\tilde{x}$ solves VIP(∂h,dom(T)); i.e., $0 \in (\partial h + N_{\mathrm{dom}(T)})(\tilde{x})$ (note that $C^o \subset \mathrm{dom}(T) \subset C$ by Lemma 1). By H8, $\mathrm{dom}(T) = \mathrm{dom}(\hat{T}) \cap \mathrm{dom}(N_V) = \mathbf{R}^n \cap V = V$; i.e.,

$$0 = \nabla h(\tilde{x}) + v', \tag{13}$$

with $v' \in N_V(\tilde{x})$. Let $\bar{v} = \mu v + v'$. Since $N_V(\tilde{x})$ is a cone, $\bar{v} \in N_V(\tilde{x})$. From (12) and (13), $0 = \mu\hat{T}(\tilde{x}) + \bar{v} + \nabla h(\tilde{x})$; i.e., $0 \in (\mu T + \partial h)(\tilde{x}) = T_\mu(\tilde{x})$. Since $\tilde{x} \in \mathrm{dom}(T) \cap C^o$, the result follows.

b) $q_{T,C}(\tilde{x}) > 0$. Let $W = \{x \in C : h(x) \leq \mu q_{T,C}(\tilde{x}) + h(\tilde{x})\}$. W is closed and convex by H1. Observe that $W^o = \{x \in C^o : h(x) < \mu q_{T,C}(\tilde{x}) + h(\tilde{x})\}$. We claim that $V \cap W$ is bounded. By H8, $\mathrm{dom}(T) = V$, so that $V \cap W$ is the intersection of a level set of h with $\mathrm{dom}(T) \cap C$, i.e., a level set of $\hat{h}$ defined as

$$\hat{h}(x) = \begin{cases} h(x) & \text{if } x \in \mathrm{dom}(T) \cap C, \\ \infty & \text{otherwise.} \end{cases}$$

$\hat{h}$ attains its minimum by H12, and the minimizer $\tilde{x}$ is unique by strict convexity of h (see H1). So $\{x \in C : \hat{h}(x) \leq \hat{h}(\tilde{x})\} = \{\tilde{x}\}$. Since $\hat{h}$ has one bounded level set, all its level sets are bounded, in particular $V \cap W$, and the claim is established, noting that $\mu q_{T,C}(\tilde{x})$ is finite by H13.

Let $U = T_\mu + N_W$. As proved in item (i), T_μ is maximal monotone. By assumption, $\tilde{x} \in \mathrm{dom}(T_\mu)$. Since $q_{T,C}(\tilde{x}) > 0$, $\tilde{x} \in W^o = \mathrm{dom}(N_W)^o$. By Lemma 2(i), U is maximal monotone. Note that $\mathrm{dom}(U) = \mathrm{dom}(T_\mu) \cap \mathrm{dom}(N_W) = \mathrm{dom}(\hat{T}) \cap V \cap \mathrm{dom}(\partial h) \cap W = V \cap C^o \cap W \subset W$, using Lemma 1. Since W is bounded, $\mathrm{dom}(U)$ is bounded, and then U is onto by Lemma 3. Therefore there exists $x \in \mathrm{dom}(U) = V \cap C^o \cap W$ such that $0 \in U(x)$. By H8,

$$0 = \mu\hat{T}(x) + \mu v + w + \nabla h(x) \tag{14}$$

with $v \in N_V(x)$, $w \in N_W(x)$. We claim that $x \in W^o$. This is immediate if $x = \tilde{x}$, because $q_{T,C}(\tilde{x}) > 0$. Otherwise, multiplying (14) by $x - \tilde{x}$ and using strict convexity of h,

$$h(x) - h(\tilde{x}) < \langle \nabla h(x), x - \tilde{x}\rangle = \mu\langle \hat{T}(x), \tilde{x} - x\rangle + \mu\langle v, \tilde{x} - x\rangle + \langle w, \tilde{x} - x\rangle. \tag{15}$$

Note that both $\tilde{x}$ and x belong to $V \cap W$. It follows from (9) that the last two inner products in the rightmost expression of (15) are nonpositive, so that

$$h(x) < h(\tilde{x}) + \mu\langle \hat{T}(x), \tilde{x} - x\rangle \le h(\tilde{x}) + \mu q_{T,C}(\tilde{x}) \tag{16}$$

using the definition of $q_{T,C}$ in the last inequality of (16). It follows from (16) and the definition of W that $x \in W^o$, and the claim is established. Then $N_W(x) = \{0\}$, i.e., $w = 0$. Thus, we get from (14) that $0 = \mu\hat{T}(x) + v + \nabla h(x)$, with $v \in N_V(x)$. By H8, $0 \in (\mu T + \partial h)(x) = T_\mu(x)$. Uniqueness follows as in item (i). □

Proposition 2 states that the central path $\{x(\mu) : \mu > 0\}$ given by (1) is well defined and contained in C^o. As mentioned in section 1, if $T(x) = \partial f(x)$, then

$$x(\mu) = \operatorname*{argmin}_{x \in C}\{\mu f(x) + h(x)\}. \tag{17}$$

corresponding to the standard definition of the central path. We prove now that the central path has cluster points. A cluster point of $\{x(\mu)\}$ is a point $\bar{x}$ such that $\bar{x} = \lim_{k\to\infty} x(\mu_k)$ for some sequence $\{\mu_k\}$ such that $\lim_{k\to\infty}\mu_k = \infty$. We need a preparatory result, of some interest on its own.

PROPOSITION 3. *Under the assumptions of Proposition* 2*,*

i) $h(x(\mu))$ *is nondecreasing in* μ*,*

ii) $x(\mu)$ *is continuous at any* $\mu > 0$.

Proof. Take $\mu_1, \mu_2 > 0$ and let $x^i = x(\mu_i)$ $(i = 1, 2)$. By definition of $x(\mu)$, $0 \in T_\mu(x(\mu))$, which is equivalent to $-\mu^{-1}\nabla h(x(\mu)) \in T(x(\mu))$. Let $u^i = -\mu_i^{-1}\nabla h(x^i)$ $(i = 1, 2)$. Then, it holds that

$$u^i \in T(x^i) \qquad (i = 1, 2), \tag{18}$$

$$\frac{1}{\mu_1}(h(x^1) - h(x^2)) \le \frac{1}{\mu_1}\langle \nabla h(x^1), x^1 - x^2\rangle = \langle u^1, x^2 - x^1\rangle, \tag{19}$$

$$\frac{1}{\mu_2}(h(x^2) - h(x^1)) \le \frac{1}{\mu_2}\langle \nabla h(x^2), x^2 - x^1\rangle = \langle u^2, x^1 - x^2\rangle, \tag{20}$$

using convexity of h in the inequalities. Adding (19) and (20)

$$\left(\frac{1}{\mu_1} - \frac{1}{\mu_2}\right)(h(x^1) - h(x^2)) \le \frac{1}{\mu_1}\langle \nabla h(x^1), x^1 - x^2\rangle + \frac{1}{\mu_2}\langle \nabla h(x^2), x^2 - x^1\rangle$$

$$= \langle u^1 - u^2, x^2 - x^1\rangle \le 0, \tag{21}$$

using (18) and monotonicity of T in the second inequality. Assume now that $\mu_1 > \mu_2$. Looking at the extreme expressions in (21), we get

$$h(x^1) \ge h(x^2), \tag{22}$$

establishing (i). By (19), (21), and (22)

$$0 \le \frac{1}{\mu_1}[h(x^1) - h(x^2)] \le \frac{1}{\mu_1}\langle \nabla h(x^1), x^1 - x^2\rangle \le \frac{1}{\mu_2}\langle \nabla h(x^2), x^1 - x^2\rangle. \tag{23}$$

Fix now $\bar{\mu} > 0$, and take $\hat{\mu}$, $\tilde{\mu}$ such that $\tilde{\mu} > \bar{\mu} > \hat{\mu}$. For any $\mu \in (\hat{\mu}, \tilde{\mu})$ we get

$$h(x(\mu)) \leq h(x(\tilde{\mu})) \tag{24}$$

using (22) with $\mu_1 = \tilde{\mu}$, $\mu_2 = \mu$, and

$$0 \leq \langle \nabla h(x(\hat{\mu})), x(\mu) - x(\hat{\mu}) \rangle \tag{25}$$

using (23) with $\mu_1 = \mu$, $\mu_2 = \hat{\mu}$. Let $L_1 = \{x \in C : h(x) \leq h(x(\tilde{\mu}))\}$, $L_2 = \{x \in C : \langle \nabla h(x(\hat{\mu})), x - x(\hat{\mu}) \rangle \geq 0\}$, and $L = L_1 \cap L_2$. By Proposition 2, (24), and (25), $\{x(\mu) : \hat{\mu} \leq \mu \leq \tilde{\mu}\} \subset L$. We claim that L is bounded. Define $\bar{h}$ as

$$\bar{h}(x) = \begin{cases} h(x) & \text{if } x \in L_2, \\ \infty & \text{otherwise.} \end{cases}$$

It is easy to see that $\bar{h}$ is strictly convex and that $x(\hat{\mu})$ is its unique minimizer, so that $\bar{h}$ has one bounded level set, and therefore all its level sets are bounded, among them L, which can be written as $\{x \in \mathbf{R}^n : \bar{h}(x) \leq \bar{h}(x(\tilde{\mu}))\}$. The claim is established, and therefore $\{x(\mu) : \hat{\mu} \leq \mu \leq \tilde{\mu}\}$ is bounded. In order to prove that $x(\mu)$ is continuous at $\bar{\mu}$, it suffices to prove that $\lim_{k\to\infty} x(\mu_k) = x(\bar{\mu})$ for any sequence $\{\mu_k\}$ such that $\lim_{k\to\infty} \mu_k = \bar{\mu}$. Given such a sequence $\{\mu_k\}$, $\mu_k \in (\hat{\mu}, \tilde{\mu})$ for large enough k, so that $\{x(\mu_k)\}$ is bounded. Let y be any cluster point of $\{x(\mu_k)\}$, $\bar{x} = x(\bar{\mu})$, and $x^k = x(\mu_k)$. Without loss of generality (i.e., refining the sequence if necessary), we may assume that $y = \lim_{k\to\infty} x^k$.

Let $K_1 = \{k : \mu_k \leq \bar{\mu}\}$ and take $k \in K$. Using (23) with $\mu_1 = \bar{\mu}$, $\mu_2 = \mu_k$, we get, by convexity of h,

$$\frac{1}{\bar{\mu}}[h(\bar{x}) - h(x^k)] \leq \frac{1}{\bar{\mu}}\langle \nabla h(\bar{x}), \bar{x} - x^k \rangle \leq \frac{1}{\mu_k}\langle \nabla h(x^k), \bar{x} - x^k \rangle \leq \frac{1}{\mu_k}[h(\bar{x}) - h(x^k)]. \tag{26}$$

Note that $\{x^k\} \subset L_1 \subset \mathrm{ED}(h)$. By H1, h is continuous at $y = \lim_{k\to\infty} x^k$, and hence $y \in ED(h)$. Then, if K_1 is infinite, we take limits in (26) as k goes to ∞, with $k \in K_1$, obtaining

$$\frac{1}{\bar{\mu}}[h(\bar{x}) - h(y)] \leq \frac{1}{\bar{\mu}}\langle \nabla h(\bar{x}), \bar{x} - y \rangle \leq \frac{1}{\bar{\mu}}[h(\bar{x}) - h(y)]. \tag{27}$$

From (27), we get $h(\bar{x}) - h(y) = \langle \nabla h(\bar{x}), \bar{x} - y \rangle$. By strict convexity of h, $y = \bar{x}$. If K_1 is finite (or empty), then we consider the complement K_2 of K_1, which certainly is infinite, and apply (23) with $\mu_1 = \mu_k$, $\mu_2 = \bar{\mu}$, obtaining a chain of inequalities similar to (26). After passing to the limit as k goes to ∞, with $k \in K_2$, we get a chain of inequalities like (27) with the roles of $\bar{x}$ and y reversed and obtain again that $y = \bar{x}$. We have proved that in any case all cluster points of x^k are equal to $\bar{x}$. We conclude that $\lim_{k\to\infty} x^k = \bar{x}$, i.e., that $x(\mu)$ is continuous at $\bar{\mu}$. □

With the help of Proposition 3(i), we prove now, under several alternative hypotheses, that the set $\{x(\mu) : \mu \geq \bar{\mu}\}$ is bounded for any $\bar{\mu} > 0$.

PROPOSITION 4. *Assume that T satisfies* H8, *that VIP(T,C) has solutions* (*see* H11), *that h attains its minimum in dom(T) $\cap$ C^o* (*see* H12) *and that all the hypotheses of Proposition* 2 *hold. If either*

i) *h is finite at the boundary* (*see* H3), *or*

ii) H14 *holds,*

then $\{x(\mu)\}$ has cluster points.

Proof. We start with a basic chain of inequalities which will be used in this and other proofs. By definition of $x(\mu)$, $-\mu^{-1}\nabla h(x(\mu)) \in T(x(\mu))$. Then, for all $y \in V$,

$$\frac{1}{\mu}\langle \nabla h(x(\mu)), x(\mu) - y\rangle = \langle \hat{T}(x(\mu)), y - x(\mu)\rangle + \langle v, y - x(\mu)\rangle$$

$$\leq \langle \hat{T}(x(\mu)), y - x(\mu)\rangle \leq \langle \hat{T}(y), y - x(\mu)\rangle, \tag{28}$$

with $v \in N_V(x(\mu))$, using H8 in the equality, (9) in the first inequality, and monotonicity of T in the second one.

Consider the first case (i). By H11, $S(T,C) \neq \emptyset$. Take $z \in S(T,C)$. Since h is finite at the boundary (see H3), $z \in \mathrm{ED}(h)$. Take $y = z$ in (28), and get

$$\frac{1}{\mu}[h(x(\mu)) - h(z)] \leq \frac{1}{\mu}\langle \nabla h(x(\mu)), x(\mu) - z\rangle \leq \langle \hat{T}(x(\mu)), z - x(\mu)\rangle, \tag{29}$$

using H1 in the first inequality. Since $N_V(x(\mu))$ is a positive cone, $0 \in N_V(x(\mu))$, i.e., $\hat{T}(x(\mu)) \in T(x(\mu))$, and, since $z \in S(T,C)$, $\langle \hat{T}(x(\mu)), z - x(\mu)\rangle \leq 0$. It follows from (29) that $h(x(\mu)) \leq h(z)$ for all $\mu > 0$. As in the proof of Proposition 2, H12 implies that the intersections of V with the level sets of h are bounded, and $\{x(\mu)\}$ is contained in one of them. The result follows.

Now we consider hypothesis (ii). If h fails to be finite at the boundary, we may have $h(z) = \infty$ for all $z \in S(T,C)$ and the previous argument breaks down. Now, under H14(i), $T = \partial f$, so that $-\mu^{-1}\nabla h(x(\mu)) \in T(x(\mu)) = \partial f(x(\mu))$. Fix $\bar{\mu} > 0$ and take any $\mu \geq \bar{\mu}$. Therefore

$$\frac{1}{\mu}[h(x(\mu)) - h(x(\bar{\mu}))] \leq \frac{1}{\mu}\langle \nabla h(x(\mu)), x(\mu) - x(\bar{\mu})\rangle \leq f(x(\bar{\mu})) - f(x(\mu)) \tag{30}$$

using convexity of h in the first inequality and convexity of f in the second one. Use Proposition 3 and (30) to conclude that $f(x(\mu)) \leq f(x(\bar{\mu}))$. So, $\{x(\mu) : \mu \geq \bar{\mu}\}$ is contained in a level set of f. By C5, VIP(T,C) is equivalent to min $f(x)$ s.t. $x \in C$ and the set of solutions of this problem, which is a level set of the restriction of f to C, is bounded. Then all such level sets are bounded, and it follows that $\{x(\mu) : \mu \geq \bar{\mu}\}$ is bounded. Thus, $\{x(\mu)\}$ has cluster points.

Consider next assumption H14(ii). Fix $\bar{\mu} > 0$. Let $u = -\mu^{-1}\nabla h(x(\mu)) \in T(x(\mu))$. By H1, with $\mu \geq \bar{\mu}$ and y as in H14(ii),

$$\langle u, x(\mu) - y\rangle \leq \frac{1}{\mu}[h(y) - h(x(\mu))] \leq \frac{1}{\mu}[h(y) - h(x(\bar{\mu}))]$$

$$\leq \frac{1}{\mu}\left|h(y) - h(x(\bar{\mu}))\right| \leq \frac{1}{\bar{\mu}}\left|h(y) - h(x(\bar{\mu}))\right|, \tag{31}$$

using Proposition 3(i) in the second inequality. Let θ be the rightmost expression in (31). If $\{x(\mu) : \mu \geq \bar{\mu}\}$ is unbounded, there exists a sequence $\{\mu_k\}$ such that $\lim_{k\to\infty}\|x(\mu_k)\| = \infty$ and $\langle u^k, x(\mu_k) - y\rangle \leq \theta$ for all k, with $u^k = -\mu_k^{-1}\nabla h(x(\mu_k)) \in T(x(\mu_k))$, in contradiction with H14(ii). So $\{x(\mu) : \mu \geq \bar{\mu}\}$ is bounded and the result holds.

The result is immediate under assumption H14(iii), because $\{x(\mu)\} \subset \mathrm{dom}(T) \cap C$. □

Now we prove that the cluster points of $\{x(\mu)\}$ solve VIP(T,C). This is the point where the paramonotonicity assumption implicit in H8 is used for the first time.

PROPOSITION 5. *Under the assumptions of Proposition* 4*, all cluster points of* $\{x(\mu)\}$ *are solutions of VIP(T,C).*

Proof. Take $z \in S(T,C)$ and $y \in V \cap C^o$ (both sets are nonempty by H10, H11). Let $\bar{x}$ be a cluster point of $\{x(\mu)\}$ (which exists by Proposition 4) and $\{\mu_k\}$ a sequence such that $\lim_{k\to\infty} \mu_k = \infty$ and $\lim_{k\to\infty} x(\mu_k) = \bar{x}$. Let $x^k = x(\mu_k)$ and $y(\varepsilon) = (1-\varepsilon)z + \varepsilon y$ with $\varepsilon \in (0,1)$. Then $y(\varepsilon) \in V \cap C^o$ and using (28)

$$\frac{1}{\mu_k}\langle \nabla h(y(\varepsilon)), x^k - y(\varepsilon)\rangle \le \langle \hat{T}(x^k), y(\varepsilon) - x^k\rangle. \tag{32}$$

Since $\lim_{k\to\infty}\langle \nabla h(y(\varepsilon)), x^k - y(\varepsilon)\rangle$ has the finite value $\langle \nabla h(y(\varepsilon)), \bar{x} - y(\varepsilon)\rangle$, the leftmost expression in (32) converges to 0 as k goes to ∞. Thus, taking limits in (32) we get

$$0 \le \langle \hat{T}(\bar{x}), y(\varepsilon) - \bar{x}\rangle \tag{33}$$

using H8. Taking limits in (33) as ε goes to 0, we get $0 \le \langle \hat{T}(\bar{x}), z - \bar{x}\rangle$. Since $\bar{x} \in V$ because V is closed, $0 \in N_V(\bar{x})$, and therefore $\hat{T}(\bar{x}) \in T(\bar{x})$. By Proposition 1(i) and (iii) and paramonotonicity of $\hat{T}$, T is paramonotone. By Proposition 1(ii), $\bar{x}$ solves VIP(T,C). □

Now we establish convergence of $\{x(\mu)\}$ to the analytic center of $S(T,C)$ when h is finite at the boundary (see H3). In this case H9 is not needed.

PROPOSITION 6. *Under the assumptions of Proposition* 4 *if h is finite at the boundary* (*see* H3), *then* $\lim_{\mu\to\infty} x(\mu)$ *exists and is the solution of*

$$\min h(x) \tag{34}$$

$$s.t.\ x \in S(T,C). \tag{35}$$

Proof. Let $\bar{x}$ be a cluster point of $\{x(\mu)\}$. $\bar{x} \in S(T,C)$ by Proposition 5. Take a sequence $\{x^k\}$ as in the proof of Proposition 5, and $z \in S(T,C)$. Then, using (28),

$$\frac{1}{\mu_k}\langle \nabla h(x^k), x^k - z\rangle \le \langle \hat{T}(x^k), z - x^k\rangle. \tag{36}$$

Since $x^k \in V$, $0 \in N_V(x^k)$, and so $\hat{T}(x^k) \in T(x^k)$. Since $z \in S(T,C)$ and $x^k \in C$, it follows that the right-hand side of (36) is nonpositive, so that, using convexity of h in (36), we get

$$h(x^k) - h(z) \le \langle \nabla h(x^k), x^k - z\rangle \le 0. \tag{37}$$

Since h is finite at the boundary, taking limits in (37) as k goes to ∞, we get $h(\bar{x}) \le h(z)$. Since z is an arbitrary element in $S(T,C)$, it follows that $\bar{x}$ minimizes h in $S(T,C)$. Since $S(T,C)$ is convex for monotone T and h is strictly convex in C by H3, the minimizer is unique. Therefore all cluster points coincide and $\{x(\mu)\}$ converges, as μ goes to ∞, to the solution of (34)–(35). □

Before proving a result similar to Proposition 6 without finiteness of h at the boundary, we summarize the results of Propositions 2–6 in the following theorem.

Theorem 1. *Assume that h satisfies* H1, *is boundary coercive* (*see* H2), *and attains its minimum on dom(T) $\cap C^o$* (*see* H12); *that T satisfies* H8; *that VIP(T,C) is regular and has solutions* (*see* H10 *and* H11); *and additionally that either*

i) *h is zone coercive and finite at the boundary* (*see* H4 *and* H3), *or*

ii) *the gap function is finite* (*see* H13) *and any of the alternatives in* H14 *holds. Then for any $\bar{\mu} > 0$ the curve $\{x(\mu) : \mu \geq \bar{\mu}\}$ is well defined, continuous, bounded, and contained in C^o, and all its limit points are solutions of VIP(T,C). In the case in which h is finite at the boundary, the path converges to the analytic center of the solution set with respect to the barrier h, i.e., to the solution of* (34)–(35).

Proof. We get that $\{x(\mu)\}$ is well defined and contained in C^o from Proposition 2. We remark that the hypotheses of Proposition 2 hold in all the cases we are considering here, because H8 implies maximal monotonicity of T. The remaining results follow from Propositions 3(ii), 4, 5, and 6, all of whose hypotheses are included in the assumptions of this theorem. □

We mention here specifically the case of the optimization problem, i.e., min $f(x)$ s.t. $x \in C$, where f is the restriction of a convex and differentiable function $\hat{f} : \mathbf{R}^n \to \mathbf{R}$ to a closed and convex set V. In this case H8 holds with $\hat{T} = \nabla \hat{f}$, and we get boundedness of the central path and optimality of the cluster points under H1, H2, H10, H11, H12, and additionally either H3 and H4, or boundedness of the solution set and convergence of the whole path to the analytic center of the solution set under H3. The case of a nondifferentiable function is not precisely covered here, because, taking $\hat{T} = \partial f$, $\hat{T}$ is not point-to-point in general. This is part of the bit of generality lost when we decided to use H8 instead of pseudomonotonicity (see section 3). The result can be obtained for this case also, at the cost of quite a few technical complications in the proofs.

Finally we proceed to study the limit of $x(\mu)$ without finiteness of h at the boundary (see H3). We impose more restrictive assumptions. First we assume that C is a box and that h is separable (see H5). For the sake of simplicity we will take $C = \mathbf{R}^n_+$, but the result holds without any changes in an arbitrary box, either bounded or unbounded. We will assume also that T satisfies both H8 and H9. We need two preparatory results. The first one establishes that under these hypotheses $S(T,C)$ is a polyhedron.

Proposition 7. *Assume that T satisfies* H8–H9, *and that VIP(T,C) is regular and has solutions* (*see* H10 *and* H11). *Fix any $z \in S(T,C)$ and any $\tilde{x} \in C \cap V$. Then*

$$S(T,C) = \{x \in V \cap C : \hat{T}(\tilde{x})^t x = \hat{T}(\tilde{x})^t z, J_{\hat{T}}(\tilde{x})x = J_{\hat{T}}(\tilde{x})z\}. \tag{38}$$

Proof. See [Ius3, Proposition 2.4]. □

We remark that the proof of Proposition 7 does not require V to be an affine manifold, as in H9, but just a closed and convex set, as in H8. But it requires that the kernel of the Jacobian matrix of $\hat{T}$ be constant. When $C \cap V$ is a polyhedron, it follows that $S(T,C)$ is also a polyhedron, namely the intersection of $V \cap C$ with the affine manifold defined by the linear equations in (38).

We need also an intermediate optimality result, for which polyhedrality of $C \cap V$ is essential.

Proposition 8. *Let $C = \mathbf{R}^n_+$. Under the assumptions of Propositions* 2 *and* 7, *the point $x(\mu)$ belonging to the central path is the solution of*

$$\min h(x) \tag{39}$$

$$s.t.\ x \in S(\mu), \tag{40}$$

where

$$S(\mu)=\{x\in\mathbf{R}^n:\hat{T}(\tilde{x})^tx=\hat{T}(\tilde{x})^tx(\mu),J_{\widehat{T}}(\tilde{x})x=J_{\widehat{T}}(\tilde{x})x(\mu),Ax=b,x\geq 0\},\tag{41}$$

$V=\{x\in\mathbf{R}^n:Ax=b\}$, *and* $\tilde{x}$ *is any point in* $C\cap V$.

Proof. $x(\mu)$ is well defined and contained in C^o by Proposition 2. Since $x(\mu)\in C^o\cap V$, it follows that $x(\mu)\in S(\mu)$. By convexity of h, it suffices to check the Karush–Kuhn–Tucker conditions for (39)–(40), which are $x(\mu)\in S(\mu)$ (already checked), and existence of $u\in\mathbf{R}^m$, $w\in\mathbf{R}^n$, and $\xi\in\mathbf{R}$ such that

$$\nabla h(x(\mu))+A^tu+J_{\widehat{T}}(\tilde{x})^tw+\xi\hat{T}(\tilde{x})\geq 0,\tag{42}$$

$$x(\mu)^t[\nabla h(x(\mu))+A^tu+J_{\widehat{T}}(\tilde{x})^tw+\xi\hat{T}(\tilde{x})]=0.\tag{43}$$

We will exhibit u, w, and ξ such that (42) holds with equality and consequently (43) also holds. By definition of $x(\mu)$, $-1/\mu\nabla h(x(\mu))\in T(x(\mu))$; i.e., there exists $v\in N_V(x(\mu))$ such that

$$-\frac{1}{\mu}\nabla h(x(\mu))=\hat{T}(x(\mu))+v.\tag{44}$$

It is easy to check that $N_V(x)=\text{Im}(A^t)$ for all $x\in V$. Thus $\mu v\in\text{Im}(A^t)$; i.e., $\mu v=A^tu$ for some $u\in\mathbf{R}^m$ and (44) can be rewritten as

$$\nabla h(x(\mu))+\mu\hat{T}(x(\mu))+A^tu=0.\tag{45}$$

Now, we write $\hat{T}(x(\mu))$ as

$$\hat{T}(x(\mu))=\hat{T}(\tilde{x})+\int_0^1 J_{\widehat{T}}(y(\tau))(x(\mu)-\tilde{x})d\tau,\tag{46}$$

with $y(\tau)=\tilde{x}+\tau(x(\mu)-\tilde{x})$. It is well known and easy to prove (e.g., see [Ius3, Proposition 2.2]) that for a monotone and differentiable operator U it holds that $\text{Ker}(J_U(x))=\text{Ker}(J_U(x)^t)$. Then, using H9,

$$J_{\widehat{T}}(y(\tau))(x(\mu)-\tilde{x})\in\text{Im}[J_{\widehat{T}}(y(\tau))]=\text{Ker}[J_{\widehat{T}}(y(\tau))^t]^\perp=\text{Ker}[J_{\widehat{T}}(y(\tau))]^\perp=W^\perp.$$

It follows that $\int_0^1 J_{\widehat{T}}(y(\tau))(x(\mu)-\tilde{x})d\tau\in W^\perp$, and therefore, by (46) and H9, $\mu\hat{T}(x(\mu))=\mu\hat{T}(\tilde{x})+p$, with $p\in W^\perp=\text{Ker}[J_{\widehat{T}}(\tilde{x})]^\perp=\text{Im}[J_{\widehat{T}}(\tilde{x})^t]$, so that $p=J_{\widehat{T}}(\tilde{x})^tw$ for some $w\in\mathbf{R}^n$. We conclude that

$$\mu\hat{T}(x(\mu))=\mu\hat{T}(\tilde{x})+J_{\widehat{T}}(\tilde{x})^tw.\tag{47}$$

Replacing (47) in (45),

$$\nabla h(x(\mu))+\mu\hat{T}(\tilde{x})+J_{\widehat{T}}(\tilde{x})^tw+A^tu=0,\tag{48}$$

which gives equality in (42) with $\xi=\mu$. □

The idea now is to push the optimality property of $x(\mu)$ given in Proposition 8 to the limit as $\mu\to\infty$, so that $S(\mu)$ becomes $S(T,C)$ by Proposition 7. The problem is that $\lim_{\mu\to\infty}x(\mu)$ may belong to ∂C where h is possibly ∞. We must work now with the optimal face of the orthant and use the separability of h (H5), under which

$h(x) = \sum_{j=1}^{n} h_j(x_j)$. Let $J = \{j \in \{1, \dots, n\} : \exists z \in S(T,C) \text{ such that } z_j > 0\}$. By convexity of $S(T,C)$ the set $\{z \in S(T,C) : z_j > 0 \text{ for all } j \in J\}$ is nonempty, and it is in fact the relative interior of $S(T,C)$. We define $\tilde{h} : \mathbf{R}^n_{++} \to \mathbf{R}$ as

$$\tilde{h}(x) = \sum_{j \in J} h_j(x_j). \tag{49}$$

We prove now the optimality property of $\lim_{\mu\to\infty} x(\mu)$ without finiteness of h at the boundary (see H3).

THEOREM 2. *Let $C = \mathbf{R}^n_+$. Assume that h satisfies* H1, *is boundary coercive* (*see* H2) *and separable* (*see* H5) *and attains its minimum in dom* $(T) \cap C^o$ (*see* H12); *that T satisfies* H8–H9; *that VIP(T,C) is regular* (*see* H10) *and has solutions* (*see* H11), *and its gap function is finite* (*see* H13); *and that some of the alternatives in* H14 *hold. Then the central path $\{x(\mu)\}$ converges as $\mu \to \infty$ to the solution x^* of*

$$\min \tilde{h}(x) \tag{50}$$

$$s.t.\ x \in S(T,C), \tag{51}$$

with $\tilde{h}$ as in (49).

Proof. By Theorem 1, $\{x(\mu)\}$ has cluster points. Let $\bar{x}$ be a cluster point of $\{x(\mu)\}$ and $\{\mu_k\}$ a sequence such that $\lim_{k\to\infty} \mu_k = \infty$, $\lim_{k\to\infty} x(\mu_k) = \bar{x}$. Call $x^k = x(\mu_k)$. Fix some $\tilde{x} \in V \cap C$ as in Propositions 7 and 8. Take any $z \in S(T,C)$. We will prove that

$$\tilde{h}(\bar{x}) \le \tilde{h}(z). \tag{52}$$

We consider first the case in which z belongs to the relative interior of $S(T,C)$; i.e., $z_j > 0$ for all $j \in J$. Define $y^k = z - \bar{x} + x^k$. We claim that $y^k \in S(\mu_k)$ for large enough k, with $S(\mu_k)$ as in (41). Let $I = \{1, \dots, n\} \setminus J$. It follows from the definition of J that $x_j = 0$ for all $j \in I$ and all $x \in S(T,C)$. Since $\bar{x} \in S(T,C)$ by Theorem 1 and $z \in S(T,C)$ by assumption, we get

$$y^k_j = x^k_j \qquad (j \in I). \tag{53}$$

Since $x^k \in C^o$ by Proposition 2, we have $y^k_j > 0$ for $j \in I$. For $j \in J$, we have $y^k_j = z_j - \bar{x}_j + x^k_j$. Since $\lim_{k\to\infty} x^k_j = \bar{x}_j$ and $z_j > 0$ for $j \in J$, we have $y^k_j > 0$ for $j \in J$ and k large enough. We conclude that $y^k > 0$ for large enough k.

Let the set V of H8, H9 be equal to $\{x \in \mathbf{R}^n : Ax = b\}$. Since $\bar{x}$, $z \in S(T,C) \subset \text{dom}(T) = \mathbf{R}^n \cap V = V$, we have $Az = A\bar{x} = b$, and therefore $Ay^k = Ax^k = b$, because $x(\mu) \in \text{dom}(T) \subset V$. So, it suffices to check that $\hat{T}(\tilde{x})^t(x^k - y^k) = 0$ and that $J_{\widehat{T}}(\tilde{x})(y^k - x^k) = 0$. Since $x^k - y^k = \bar{x} - z$, this follows from Proposition 7, because $\bar{x} \in S(T,C)$ by Theorem 1. We have proved that $y^k \in S(\mu_k)$. By Proposition 8, $h(x^k) \le h(y^k)$; i.e.,

$$\tilde{h}(x^k) + \sum_{j \in I} h_j(x^k_j) \le \tilde{h}(y^k) + \sum_{j \in I} h_j(y^k_j). \tag{54}$$

From (53) and (54),

$$\tilde{h}(x^k) \le \tilde{h}(y^k). \tag{55}$$

Now, since we do not have finiteness of h at the boundary (see H3), we must be careful with the behavior of $\tilde{h}$ at the boundary of $\mathbf{R}^n_+$. It follows from H1 and the separability of h (see H5) that $h_j : \mathbf{R}_{++} \to \mathbf{R}$ is continuous and closed. Thus, $\lim_{t\to 0} h_j(t)$ is well defined (possibly infinity), and we can take limits in (55) as k goes to ∞. Since $\lim_{k\to\infty} y^k = z$, $\lim_{k\to\infty} x^k = \bar{x}$, and $z_j > 0$ for all $j \in J$, we conclude that

$$\tilde{h}(\bar{x}) \leq \tilde{h}(z) < \infty \tag{56}$$

for any z in the relative interior of $S(T, C)$. By the same argument as above, (56) holds in fact for all $z \in S(T, C)$ (note that $\tilde{h}(z)$ might be infinite for some z in the relative boundary of $S(T, C)$, i.e., such that $z_j = 0$ for some $j \in J$). So $\bar{x}$ solves (50)–(51). Observe that $\tilde{h}$ is not strictly convex in $\mathbf{R}^n_{++}$, but it is strictly convex in $S(T, C)$ because h_j is strictly convex by H1 and $x_j = 0$ for all $j \notin J$ and all $x \in S(T, C)$, by definition of J. Therefore problem (50)–(51) has a unique solution, and since all cluster points of $\{x(\mu)\}$ are equal to this solution, we conclude that $\lim_{\mu\to\infty} x(\mu)$ exists and that the result holds. □

3. Generalized proximal point methods for variational inequalities. Let C be a closed and convex subset of $\mathbf{R}^n$, with nonempty interior. Consider a function $g : C \to \mathbf{R}$ satisfying H1 and H3 of section 2, and define $D_g : C \times C^o \to \mathbf{R}$ as

$$D_g(x, y) = g(x) - g(y) - \langle \nabla g(y), x - y \rangle.$$

g is said to be a Bregman function if it satisfies H1 and H3 of section 2 and, additionally, the following hold.

B1. For all $\nu \in \mathbf{R}$ the partial level sets $\Gamma(x, \nu) = \{y \in C^o : D_g(x, y) \leq \nu\}$ are bounded for all $x \in C$.

B2. If $\{y^k\} \subset C^o$ converges to y^*, then $D_g(y^*, y^k)$ converges to 0.

B3. If $\{x^k\} \subset C$ and $\{y^k\} \subset C^o$ are sequences such that $\{x^k\}$ is bounded, $\lim_{k\to\infty} y^k = y^*$ and $\lim_{k\to\infty} D_g(x^k, y^k) = 0$, then $\lim_{k\to\infty} x^k = y^*$.

C^o is called the *zone* of g. This definition originates in the results of [Breg]. It is easy to check that $D_g(x, y) \geq 0$ for all $x \in C$, $y \in C^o$, and $D_g(x, y) = 0$ if and only if $x = y$. Also, $D_g(\cdot, y)$ is strictly convex and continuous in C for all $y \in C^o$.

As examples of Bregman functions for the case of $C = \mathbf{R}^n_+$ we have the following.

Example 1. $g(x) = \sum_{j=1}^n x_j \log x_j$, extended with continuity to $\partial\mathbf{R}^n_+$ with the convention that $0 \log 0 = 0$. In this case $D_g(x, y) = \sum_{j=1}^n (x_j \log \frac{x_j}{y_j} + y_j - x_j)$, which is the Kullback–Leibler divergence, widely used in statistics (see [Lie]).

Example 2. $g(x) = \sum_{j=1}^n (x_j^\alpha - x_j^\beta)$ with $\alpha \geq 1$, $0 < \beta < 1$. For $\alpha = 2$, $\beta = \frac{1}{2}$ we get $D_g(x, y) = \|x - y\|^2 + \sum_{j=1}^n \frac{1}{2\sqrt{y_j}}(\sqrt{x_j} - \sqrt{y_j})^2$, and for $\alpha = 1$, $\beta = \frac{1}{2}$, we get $D_g(x, y) = \sum_{j=1}^n \frac{1}{2\sqrt{y_j}}(\sqrt{x_j} - \sqrt{y_j})^2$.

The functions g of Examples 1 and 2 satisfy B1–B3 and H1–H5, except for the case of g as in Example 2 with $\alpha = 1$, which fails to satisfy only H4. Examples of Bregman functions for other convex sets, like balls, boxes, and polyhedra with nonempty interiors, can be found in [Cen2].

If we discard H3, we can consider also $g(x) = -\sum_{j=1}^n \log x_j$, which satisfies H1, H2, H5, and B1–B3. In this case $D_g : C^o \times C^o \to \mathbf{R}$ is given by

$$D_g(x, y) = \sum_{j=1}^n \left[\frac{x_j}{y_j} - \log\left(\frac{x_j}{y_j}\right) - 1 \right] \tag{57}$$

and is called the Itakura–Saitu distance (see [Csi]).

The GPPA algorithm for VIP(T,C) is defined as follows. Take a Bregman function g with zone C^o and a sequence $\{\lambda_k\}$ of positive numbers bounded above by some $\bar{\lambda} > 0$. Let $\{x^k\}$ be defined by the following.

INITIALIZATION:

$$x^0 \in C^o.$$

ITERATIVE STEP: Given x^k, consider the operator T_k defined by

$$T_k(\cdot) = T(\cdot) + \lambda_k[\nabla g(\cdot) - \nabla g(x^k)]. \tag{58}$$

Then, take x^{k+1} such that

$$0 \in T_k(x^{k+1}). \tag{59}$$

We summarize the convergence properties of this algorithm in the following proposition.

PROPOSITION 9. *Assume that g satisfies* H1, H3, *and* B1–B3, *T satisfies* H8–H9, *and VIP(T,C) satisfies* H10–H11. *If either* H2 *and* H13 *or* H4 *holds, then the sequence $\{x^k\}$ generated by GPPA* (*i.e., by* (58)–(59)) *converges to the solution of*

$$\min D_g(x, x^0) \tag{60}$$

$$s.t.\ x \in S(T, C). \tag{61}$$

Proof. See [Ius3, Theorem 5.1]. □

The proof of Proposition 9 addresses mainly the issue of characterizing the limit of the sequence as the solution of (60)–(61), using H9. The fact that the sequence does converge was previously proved in [Bur2, Corollary 3], and does not require H9. This proof uses H6, H7, and a hypothesis which is weaker than H8, namely *pseudomonotonicity*, which is rather technical.

We study next the connection between the central path and the GPPA. Our first result is just a corollary of Theorem 1 and Proposition 9 and shows that in many relevant cases the GPPA sequence and the central path (with the Bregman function as a barrier) converge precisely to the same point.

COROLLARY 1. *Assume that T satisfies* H8 *and* H9; *that g satisfies both* H1 *and the Bregman assumptions* B1–B3 *and is zone coercive* (H4) *and finite at the boundary* (H3); *and that VIP(T,C) is regular and has solutions* (H10–H11). *Let $\{x^k\}$ be the sequence generated by GPPA with Bregman function g starting at some $x^0 \in$ dom$(T) \cap C^o$ and $\{x(\mu)\}$ the central path with barrier $h(x) = D_g(x, x^0)$. Then both* $\lim_{k\to\infty} x^k$ *and* $\lim_{\mu\to\infty} x(\mu)$ *are equal to the solution of* (60)–(61), *i.e., the analytic center of $S(T, C)$ with respect to the barrier h.*

Proof. The result for $\{x^k\}$ follows from Proposition 9, and the result for $\{x(\mu)\}$ follows from Theorem 1(i), which holds because $g(\cdot)$ and $D_g(\cdot, x^0)$ differ by the affine term $\langle \nabla h(x^0), x - x^o \rangle$, and H1, H3, and H4 are invariant through additions of affine functions, so that h also satisfies H1, H3, and H4. Also, in this case H12 holds because the function $h(x) = D_g(x, x^0)$ attains its minimum at $x^0 \in \text{dom}(T) \cap C^o$. □

We remark that in the central path approach, convergence to the analytic center under finiteness of h at the boundary (see H3) does not require $S(T, C)$ to be a polyhedron, as is the case when H9 holds. In the case of GPPA assumption, H9 seems to be necessary. The validity of Proposition 9 without demanding H9 remains as an open problem.

Another interesting issue is the convergence of GPPA without finiteness of g at the boundary (see H3). An extension of the notion of Bregman function has been recently presented in [Kiw1], where H3 is omitted and H1 is weakened (g need not be differentiable in C^o). GPPA for the kind of g's considered by Kiwiel is studied in [Kiw2], where only the optimization case ($T = \partial f$) is considered. The GPPA of [Kiw2] works for g's which do not satisfy H3, like the Itakura–Saitu distance (57), which, when looked at as a barrier function, produces the logarithmic barrier considered in the references mentioned in section 1. However, the convergence results in [Kiw2] request $S(T, C) \cap \mathrm{ED}(g) \neq \emptyset$. If the effective domain of g is C^o, as in the case of (57), this implies existence of solutions in the interior of C, which leaves out most interesting problems (e.g., linear programming).

Corollary 1 raises the following question: Is it the case that the sequence $\{x^k\}$ generated by GPPA is contained in the central path with barrier $h(x) = D_g(x, x^o)$? The answer is affirmative when the operator $\hat{T}$ is constant and V is an affine manifold (e.g., in the linear programming case) and negative otherwise, as we show next.

THEOREM 3. *Consider VIP(T,C) with $T(x) = c + N_V$ and $V = \{x \in \mathbf{R}^n : Ax = b\}$. Assume that VIP(T,C) is regular and has solutions* (*see* H10–H11) *and that g satisfies* H1 *and is boundary coercive* (*see* H2). *Let $\{x^k\}$ be the sequence generated by GPPA with $x^0 \in V \cap C^o$ and $\{x(\mu)\}$ be the central path with barrier $h(x) = D_g(x, x^0)$. Then $\{x^k\} \subset \{x(\mu) : \mu > 0\}$, and for each increasing sequence $\{\mu_k\} \subset \mathbf{R}_{++}$ there exists a sequence $\{\lambda_k\} \subset \mathbf{R}_{++}$ such that $x(\mu_k) = x^k$, where $\{x^k\}$ is the sequence generated by GPPA with Bregman function g and regularization parameters λ_k.*

Proof. First we check that under these hypotheses both x^k and $x(\mu)$ are well defined. We start with $x(\mu)$. We are within hypothesis (ii) of Proposition 2 because it has been proved in [Bur1, Proposition 3.1] that H13 follows from H11 in the optimization case. For x^k, the result follows from [Bur2, Theorem 2].

Next, we observe that in this case (59) can be written as

$$\frac{1}{\lambda_\ell}(c + A^t w^\ell) + \nabla g(x^{\ell+1}) - \nabla g(x^\ell) = 0 \tag{62}$$

for some $w^\ell \in \mathbf{R}^m$, because $N_V(x) = \mathrm{Im}(A^t)$ for all $x \in V$. Summing (62) from $\ell = 0$ to $k-1$ and taking $\mu_k = \sum_{\ell=1}^{k-1} \lambda_\ell^{-1}$, $\bar{w}^k = \mu_k^{-1} \sum_{\ell=0}^{k-1} \lambda_\ell^{-1} w^\ell$, we get

$$0 = \mu_k(c + A^t \bar{w}^k) + \nabla g(x^k) - \nabla g(x^0) = \mu_k(c + A^t \bar{w}^k) + \nabla h(x^k). \tag{63}$$

By (63), $x^k{=}x(\mu_k)$. If an increasing sequence $\{\mu_k\}$ is given, then we get by the same argument $x^k = x(\mu_k)$, where $\{x^k\}$ is generated with $\lambda_k = (\mu_k - \mu_{k-1})^{-1} > 0$. □

Theorem 3 says that in the case of linear programming (i.e., when $C = \mathbf{R}^n_{++}$) the notions of central path and generalized proximal point sequence coincide. The result depends in an essential way on the fact that $\hat{T}$ is constant. It does not hold in more general cases, not even for quadratic programming, as we show next.

Let $n = 2$, and consider the problem $\min \frac{1}{2}\|x\|^2$ s.t. $x \geq 0$. Take $g(x) = -\sum_{j=1}^{2} \log x_j$ and $x^0 = (1/8, 1/2)$. Then $x(\mu) = \mathrm{argmin}\{\frac{\mu}{2}\|x\|^2 + D_g(x, x^0)\}$ with D_g as in (57), and a direct calculation gives $x(\mu) = \mu^{-1}(\sqrt{16+\mu}-4, \sqrt{1+\mu}-1)$. Now we look at the GPPA sequence, given by $x^{k+1} = \mathrm{argmin}\{\frac{1}{2}\|x\|^2 + \lambda_k D_g(x, x^k)\}$. Take $\lambda_0 = 1/48$, $\lambda_1 = 1/9$. Then $x^1 = (1/12, 1/8)$, which coincides with $x(\lambda_0^{-1}) = x(48)$, but $x^2 = ((\sqrt{5}-2)/6, 1/18)$, and it follows from the formula above for $x(\mu)$ that

$x^2 \neq x(\mu)$ for all $\mu > 0$. Therefore, the GPPA sequence is not contained in the central path.

It is interesting to discuss some consequences of Theorem 3. Consider, for instance, the case of $h(x) = -\sum_{j=1}^{n} \log x_j$. Since h is not continuous on the boundary of $\mathbf{R}^n_+$ (i.e., it does not satisfy H3) convergence of the GPPA sequence with the Itakura–Saitu distance given by (57) is not dealt with by most papers on GPPA (e.g., [Cen1], [Che], [Eck], [Ius1], [Bur2], or [Kiw2] for the case of solutions in ∂C). On the other hand, it follows from our Theorem 2 that, for a linear programming problem with a bounded set of solutions, the central path converges to the analytic center of the solution set; and then from our Theorem 3, that such central path contains the GPPA sequence $\{x^k\}$; and furthermore that $\lim_{k\to\infty} x^k = \lim_{k\to\infty} x(\mu_k)$ with $\mu_k = \sum_{\ell=0}^{k-1} \lambda_\ell^{-1}$. Since $\{\lambda_k\}$ is bounded above, $\lim_{k\to\infty} \mu_k = \infty$, and so $\{x^k\}$ converges to the analytic center of the solution set.

Of course, the central path trajectory for this h converges also in the more general cases included in the hypotheses of Theorem 2 (quadratic programming, paramonotone linear complementarity problems, etc.), but in these cases the GPPA sequence is not contained in the central path, and therefore its convergence for the case of the solution set contained in the boundary of C remains as an open problem.

Another interesting consequence of Theorem 3 is related to the limiting direction of the central path, i.e., $\lim_{\mu\to\infty} \dot{x}(\mu)$. For the case of linear programming with the logarithmic barrier, existence of this limit has been established in [Adl2], where a precise characterization of this limiting direction is given. Similar results with the same barrier for monotone linear complementarity problems have been proved in [Mon]. In view of Theorem 3, such results apply to the GPPA for linear programming with the Itakura–Saitu distance given by (57), since the limiting direction $\lim_{\mu\to\infty} \dot{x}(\mu)$ can be seen as the limiting direction of the GPPA sequence $\{x^k\}$. This limiting direction has not been considered in the GPPA literature. Extensions of the results in [Adl2] to more general barriers (e.g., $h(x) = \sum_{j=1}^{n} x_j \log x_j$) will, by virtue of Theorem 3, hold also for the GPPA sequence.

4. The relation between central paths and Cauchy trajectories in Riemannian manifolds. Let M be a Riemannian manifold of dimension s, with metric $\langle \cdot, \cdot \rangle$. Given a local coordinate system in a neighborhood U of a point $p \in M$, the metric in M is given by the symmetric and positive definite matrix $H(q)$, with $H(q)_{ij} = \langle \frac{\partial}{\partial x_i}|_q, \frac{\partial}{\partial x_j}|_q \rangle$ for all $q \in U$, where $\{\frac{\partial}{\partial x_i}\}$ is a basis of the tangent space to M. If $f : M \to \mathbf{R}$ is differentiable, the gradient of f is the vector field grad f defined by

$$\langle \text{grad } f, X \rangle = df(X) = \frac{\partial f}{\partial X},$$

where X is any vector field and $\frac{\partial f}{\partial X}$ is the derivative of f in the direction of X. It is well known that grad $f(q) = H(q)^{-1} \nabla f(q)$, where $\nabla f(q) = (\frac{\partial f}{\partial x_1}(q), \dots, \frac{\partial f}{\partial x_s}(q))$.

Let $N \subset M$ be a submanifold of M. For $x \in N$, let T_xM and T_xN be the tangent spaces to M and N at x, respectively, and let $\Pi_x : T_xM \to T_xN$ be the orthogonal projection onto T_xN with respect to the metric of the manifold. If $f_{/N}$ is the restriction of f to N, then the gradient of $f_{/N}$ with respect to the induced metric in N turns out to be given by grad $f_{/N}(x) = \Pi_x(\text{grad } f(x))$. A *Cauchy trajectory* for f in N is a curve $x : [0, \beta) \to N$ given by

$$x(0) = x^0, \tag{64}$$

$$\dot{x}(t) = -\text{grad } f_{/N}(x(t)) \tag{65}$$

for a given $x^0 \in N$ and some $\beta > 0$. It is well known that for each $x^0 \in N$ there exists $\beta > 0$ such that (64)–(65) has a unique solution.

We remark that if M is an open subset of $\mathbf{R}^s$, then the representation of the metric given by $H(x)$ is global, rather than local, $H : M \to \mathbf{R}^{s\times s}$ is differentiable, and ∇f coincides with the ordinary gradient in $\mathbf{R}^s$.

We prove now that, under the hypotheses of Theorem 3, the Cauchy trajectory and the central path for a certain barrier associated with the metric coincide.

THEOREM 4. *Let M be an open subset of $\mathbf{R}^n$, $N = M \cap \{x \in \mathbf{R}^n : Ax = b\}$, and $f(x) = c^t x$. Let H represent the metric in M and assume that there exists $h : M \to \mathbf{R}$ such that $\nabla^2 h(x) = H(x)$ for all $x \in M$ and $\nabla h(x^0) = 0$. If the Cauchy trajectory $\{x(t)\}$ and the central path $\{x(\mu)\}$ exist, then they coincide.*

Proof. The central path in this case, as in (63) with an arbitrary μ, satisfies

$$0 = \mu(c + A^t w(\mu)) + \nabla h(x(\mu)). \tag{66}$$

By (66), $\nabla h(x(0)) = 0$. Since $\nabla h(x^0) = 0$, by strict convexity of h we get $x(0) = x^0$; i.e., (64) holds. We must check that (65) holds. Let P_A be the orthogonal projection onto $\operatorname{Ker}(A)$. By (66), $\mu c + \nabla h(x(\mu)) \in \operatorname{Im}(A^t)$, so that, for all $\mu > 0$,

$$0 = P_A[\mu c + \nabla h(x(\mu))]. \tag{67}$$

Since P_A is linear, differentiating with respect to μ in (67) we get

$$0 = P_A[c + \nabla^2 h(x(\mu))\dot{x}(\mu)] = P_A[c + H(x(\mu))\dot{x}(\mu)] \tag{68}$$

or, equivalently,

$$H(x(\mu))\dot{x}(\mu) - c \in \operatorname{Im}(A^t). \tag{69}$$

Now we look at grad $f_{/N}(x(\mu))$. As observed above,

(70)
$$\operatorname{grad} f_{/N}(x(\mu)) = \Pi_{x(\mu)} \operatorname{grad} f(x(\mu)) = \Pi_{x(\mu)} H(x(\mu))^{-1} \nabla f(x(\mu)) = \Pi_{x(\mu)} H(x(\mu))^{-1} c,$$

where $\Pi_{x(\mu)}$ is the orthogonal projection of $T_{x(\mu)}M$ onto $T_{x(\mu)}N$ with respect to the inner product induced by the metric; i.e., $\langle u, v\rangle = u^t H(x(\mu))v$. Since M is open and N is an affine manifold, we have $T_{x(\mu)}M = \mathbf{R}^n$ and $T_{x(\mu)}N = \operatorname{Ker}(A)$, so that, for any $y \in \mathbf{R}^n$, $\Pi_{x(\mu)}(y)$ is the unique solution z of

$$\min (z - y)^t H(x(\mu))(z - y) \tag{71}$$

$$\text{s.t.} \qquad Az = 0, \tag{72}$$

whose sufficient Karush–Kuhn–Tucker conditions are (72), plus

$$H(x(\mu))(z - y) = A^t w \tag{73}$$

for some $w \in \mathbf{R}^m$. Note that $H(x(\mu))$ is the Hessian of h at $x(\mu)$, and so it is symmetric and positive semidefinite. For $y = H(x(\mu))^{-1}c$, (73) reduces to

$$-H(x(\mu))z + c \in \operatorname{Im}(A^t). \tag{74}$$

By (69), $z = -\dot{x}(\mu)$ satisfies (74). Since $x(\mu) \in \operatorname{dom}(T) \subset V = N$, we have $Ax(\mu) = b$ for all $\mu > 0$, which implies $0 = A\dot{x}(\mu) = A(-\dot{x}(\mu))$, so that $z = -\dot{x}(\mu)$ satisfies (72) and (74) and is therefore equal to $\Pi_{x(\mu)} y = \Pi_{x(\mu)} H(x(\mu))^{-1} c$. By (70), $\dot{x}(\mu) = -\text{grad } f_{/N}(x(\mu))$; i.e., (65) holds. We have proved that the path $\{x(\mu)\}$ coincides with the Cauchy trajectory. □

As mentioned above, the result of Theorem 4 covers the case of linear programming (with $M = \mathbf{R}^n_{++}$). It is worthwhile to remark that only one of the two conditions imposed on h in the hypotheses of Theorem 4 is indeed relevant, namely $\nabla^2 h(x) = H(x)$. If this relation is satisfied by some $\bar{h}$, then we define $h(x) = \bar{h}(x) - \nabla\bar{h}(x^0)^t x$, and we get $\nabla^2 h(x) = H(x)$ and $\nabla h(x^0) = 0$.

In view of (70) and (74), it is easy to prove that, when A has full rank, (65) can be rewritten as

$$\dot{x}(t) = -H(x(t))^{-1}[I - A^t(AH(x(t))^{-1}A^t)^{-1}AH(x(t))^{-1}]\nabla f(x(t)). \tag{75}$$

Theorem 4 says that, when $H(x) = \nabla^2 h(x)$ and $\nabla h(x^0) = 0$, the curve given by (64), (75) coincides with the curve

$$x(\mu) = \operatorname{argmin}\{\mu c + h(x)\} \tag{76}$$

$$\text{s.t. } Ax = b. \tag{77}$$

It is worthwhile to consider the form of (75) under different choices of H. For h as in Example 1, i.e., $h(x) = \sum_{j=1}^n x_j \log x_j$, we get

$$H(x)^{-1} = \operatorname{diag}(x_1, \dots, x_n). \tag{78}$$

The curve given by (64) and (75), (78) has been considered in [Fay], while, by Theorems 3 and 4, this curve contains any sequence generated by GPPA for the linear programming problem with the Kullback–Leibler divergence, a particular case of the methods for linear programming considered in [Ius5].

For h as in Example 2, with $\alpha = 1$, $\beta = 1/2$, we get

$$H(x)^{-1} = 4\operatorname{diag}(x_1^{3/2}, \dots, x_n^{3/2}). \tag{79}$$

This "scaling" has not been particularly studied, either from the perspective of interior point methods for linear programming or from the point of view of generalized proximal point methods. We feel that it deserves more attention.

Finally, for $h(x) = -\sum_{j=1}^n \log x_j$, we get

$$H(x)^{-1} = \operatorname{diag}(x_1^2, \dots, x_n^2). \tag{80}$$

In this case, (64), (75), with $H(x)^{-1}$ as in (80) gives the affine scaling trajectory, widely studied from the point of view of interior point methods for linear programming (see, e.g., [Adl1]).

REFERENCES

[Adl1] I. Adler, M. Resende, G. Veiga, and N. Karmarkar, *An implementation of Karmarkar's method for linear programming*, Math. Programming, 44 (1989), pp. 297–335.

[Adl2] I. Adler and R.D.C. Monteiro, *Limiting behavior of the affine scaling continuous trajectories for linear programming*, Math. Programming, 5 (1991), pp. 29–51.

[Bay] D.A. Bayer and J.C. Lagarias, *The nonlinear geometry of linear programming*, Trans. Amer. Math. Soc., 314 (1989), pp. 499–581.

[Breg] L.M. Bregman, *The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming*, USSR Comput. Math. Math. Phys., 7 (1967), pp. 200–217.

[Brez] H. Brézis, *Opérateurs Monotones Maximaux et Semigroups de Contractions dans les Espaces de Hilbert*, Université de Paris-CNRS, Paris, 1971.

[Bur1] R.S. Burachik, *Generalized Proximal Point Methods for the Variational Inequality Problem*, Ph.D. thesis, Instituto de Matemática Pura e Aplicada, Rio de Janeiro, Brazil, 1995.

[Bur2] R.S. Burachik and A.N. Iusem, *A generalized proximal point algorithm for the variational inequality problem in a Hilbert space*, SIAM J. Optim., 8 (1998), pp. 197–216.

[Cen1] Y. Censor and S. Zenios, *The proximal minimization algorithm with D-functions*, J. Optim. Theory Appl., 73 (1992), pp. 451–464.

[Cen2] Y. Censor, A.N. Iusem, and S.A. Zenios, *An interior point method with Bregman functions for the variational inequality problem with paramonotone operators*, Math. Programming, 81 (1998), pp. 373–400.

[Che] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like optimization algorithm using Bregman functions*, SIAM J. Optim., 3 (1993), pp. 538–543.

[Csi] I. Csiszár, *Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems*, Ann. Statist., 19 (1991), pp. 2032–2066.

[Eck] J. Eckstein, *Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming*, Math. Oper. Res., 18 (1993), pp. 202–226.

[Egg] P.P.B. Eggermont, *Multiplicative iterative algorithms for convex programming*, Linear Algebra Appl., 130 (1990), pp. 25–42.

[Eri] J. Eriksson, *An Iterative Primal-Dual Algorithm for Linear Programming*, Technical report 85–10, Department of Mathematics, Linköping University, Linköping, Sweden, 1985.

[Fay] L. Faybusovich, *Hamiltonian structure of dynamical systems which solve linear programming problems*, Phys. D, 53 (1991), pp. 217–232.

[Gon] C.G. Gonzaga, *Path following methods for linear programming*, SIAM Rev., 34 (1992), pp. 167–224.

[Gul] O. Güler, *Existence of interior points and interior paths in nonlinear monotone complementarity problems*, Math. Oper. Res., 18 (1993), pp. 128–147.

[Hel] U. Helmke and J.B. Moore, *Optimization and Dynamical Systems*, Springer-Verlag, London, 1994.

[Ius1] A.N. Iusem, *On some properties of generalized proximal point methods for quadratic and linear programming*, J. Optim. Theory Appl., 85 (1995), pp. 593–612.

[Ius2] A.N. Iusem, *On some properties of paramonotone operators*, J. Convex Anal., 5 (1998), pp. 269–278.

[Ius3] A.N. Iusem, *On some properties of generalized proximal point methods for variational inequalities*, J. Optim. Theory Appl., 96 (1998), pp. 337–362.

[Ius4] A.N. Iusem and B.V. Svaiter, *A new smoothing-regularization approach for a maximum likelihood estimation problem*, Appl. Math. Optim., 29 (1994), pp. 225–241.

[Ius5] A.N. Iusem and M. Teboulle, *Convergence rate analysis of nonquadratic proximal and augmented Lagrangian methods for convex and linear programming*, Math. Oper. Res., 20 (1995), pp. 655–677.

[Kar] N. Karmarkar, *Riemannian geometry underlying interior-point methods for linear programming*, in Mathematical Developments Arising from Linear Programming, Contemporary Mathematics 114, J.C. Lagarias and M.J. Todd, eds., American Mathematical Society, Providence, RI, 1990, pp. 51–75.

[Kin] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[Kiw1] K. Kiwiel, *Free steering relaxation methods for problems with strictly convex costs and linear constraints*, Math. Oper. Res., 22 (1997), pp. 326–349.

[Kiw2] K.C. Kiwiel, *Proximal minimization methods with generalized Bregman functions*, SIAM J. Control Optim., 35 (1997), pp. 1142–1168.

[Koj] M. Kojima, S. Mizuno, and T. Noma, *Limiting behavior of trajectories by a continuation method for monotone complementarity problems*, Math. Oper. Res., 15 (1990), pp. 662–675.

[Lem] B. Lemaire, *The proximal algorithm*, in New Methods in Optimization and Their Industrial Uses, Intern. Ser. Numer. Math. 87, J.P. Penot, ed., Birkhauser-Verlag, Basel, Switzerland, 1989, pp. 73–87.

[Lie] F. Liese and I. Vajda, *Convex Statistical Distances*, Teubner, Leipzig, 1987.

[Mar] P. Marcotte, *A new algorithm for solving variational inequalities with applications to the traffic assignment problem*, Math. Programming, 33 (1985), pp. 339–351.

[McL] L. McLinden, *An analogue of Moreau's proximation theorem, with applications to nonlinear complementarity problem*, Pacific J. Math., 88 (1990), pp. 101–161.

[Meg1] N. Megiddo, *Pathways to the optimal set in linear programming*, in Progress in Mathematical Programming: Interior Point and Related Methods, N. Megiddo, ed., Springer-Verlag, New York, 1989, pp. 131–158.

[Meg2] N. Megiddo and M. Shub, *Boundary behavior of interior point algorithms in linear programming*, Math. Oper. Res., 14 (1989), pp. 97–146.

[Mon] R.D.C. Monteiro and T. Tsuchiya, *Limiting behavior of the derivatives of certain trajectories associated with a monotone horizontal linear complementarity problem*, Math. Oper. Res., 21 (1996), pp. 793–814.

[Nes] Y. Nesterov and A. Nemirovskii, *Interior Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, PA, 1994.

[Ngu] S. Nguyen and C. Dupuis, *An efficient method for computing traffic equilibria in networks with asymmetric transportation costs*, Transportation Sci., 18 (1984), pp. 185–202.

[Roc1] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[Roc2] R.T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 2, pp. 589–616

ON HOMOTOPY-SMOOTHING METHODS FOR BOX-CONSTRAINED VARIATIONAL INEQUALITIES*

XIAOJUN CHEN† AND YINYU YE‡

Abstract. A variational inequality problem with a mapping $g : \Re^n \to \Re^n$ and lower and upper bounds on variables can be reformulated as a system of nonsmooth equations $F(x) = 0$ in $\Re^n$. Recently, several homotopy methods, such as interior point and smoothing methods, have been employed to solve the problem. All of these methods use parametric functions and construct perturbed equations to approximate the problem. The solution to the perturbed system constitutes a smooth trajectory leading to the solution of the original variational inequality problem. The methods generate iterates to follow the trajectory. Among these methods Chen–Mangasarian and Gabriel–Moré proposed a class of smooth functions to approximate F. In this paper, we study several properties of the trajectory defined by solutions of these smooth systems. We propose a homotopy-smoothing method for solving the variational inequality problem, and show that the method converges globally and superlinearly under mild conditions. Furthermore, if the involved function g is an affine function, the method finds a solution of the problem in finite steps. Preliminary numerical results indicate that the method is promising.

Key words. smoothing approximation, variational inequalities, P_0 function, finite convergence, homotopy

AMS subject classifications. 65H10, 90C30, 90C33

PII. S0363012997315907

1. Introduction.

We consider the following variational inequality problem (VIP). Given a continuously differentiable mapping $g : \Re^n \to \Re^n$, $l \in \{\Re \cup -\infty\}^n$, and $u \in \{\Re \cup \infty\}^n$, where $l < u$, find $x \in [l, u]$ such that

$$(y - x)^T g(x) \geq 0 \qquad \text{for any } y \in [l, u].$$

Such problem is called a box-constrained VIP or mixed complementarity problem in [6, 9, 11, 15, 17, 18, 19, 24]. We denote this problem by VI(l, u, g).

Several algorithms have been developed for solving VIP [17, 20]. In the last few years, smoothing methods have been studied extensively. Smoothing methods for solving VI(l, u, g) are based on the reformulation of nonsmooth equations

$$F(x) := x - \text{mid}(l, u, x - g(x)) = 0, \tag{1.1}$$

where mid$(\cdot)$ denotes the componentwise median operator, i.e.,

$$\text{mid}(l_i, u_i, x_i - g_i(x)) = \begin{cases} l_i, & x_i - g_i(x) \leq l_i, \\ x_i - g_i(x), & l_i \leq x_i - g_i(x) \leq u_i, \\ u_i, & x_i - g_i(x) \geq u_i. \end{cases}$$

These methods use parametric smooth functions and construct perturbed equations to approximate nonsmooth equations (1.1). The solution to the perturbed system constitutes a smooth trajectory leading to the solution of the original VIP.

*Received by the editors January 31, 1997; accepted for publication (in revised form) June 2, 1998; published electronically January 28, 1999.
http://www.siam.org/journals/sicon/37-2/31590.html

†Department of Mathematics, Shimane University, Matsue 690, Japan (chen@math.shimane-u.ac.jp). The research of this author was supported in part by the Australian Research Council.

‡Department of Management Sciences, The University of Iowa, Iowa City, IA 52242 (yyye@dollar.biz.uiowa.edu). The research of this author was supported in part by NSF grants DDM-9207347 and DMI-9522507.

In contrast with Newton methods using generalized Jacobian [3, 15, 16, 35, 38], smoothing methods have a notable advantage in that they can be extended to solve nonsmooth problems in function spaces [7, 10, 26]. In fact, smoothing methods are closely related to splitting methods for solving nonsmooth equations in function spaces for which the nonlinear function can be split into smooth and nonsmooth parts. Such problems arise from differential equations with nondifferentiable terms, nonsmooth compact fixed point problems, etc. In particular, a class of problems arises in optimal control problems for parabolic partial differential equations with bound constraints on the control. In a survey paper [26], Kelley considered the splitting methods for solving the nonsmooth equations

$$x(t) - P(K(x))(t) = 0, \tag{1.2}$$

where K is a completely continuous map from $L^\infty(\Omega)$ to $C(\Omega)$ for some bounded $\Omega \subseteq \Re^m$, and P is the map on $C(\Omega)$ given by

$$P(K(x))(t) = \begin{cases} l(t), & K(x)(t) \le l(t), \\ K(x)(t), & l(t) \le K(x)(t) \le u(t), \\ u(t), & K(x)(t) \ge u(t), \end{cases}$$

for given l and u in $C(\Omega)$. A paradigm for problems of the form (1.2) is the Urysohn equation [27]. Nonsmooth equation (1.1) may be considered as a special case of the form (1.2) in $\Re^n$.

In this paper, we will use the Jacobian of smooth approximation functions and directional derivative consistency property [7, 19] in the smoothing algorithm and its convergence analysis. Thus, these results can be extended to function spaces without difficulties.

The box-constrained VIP includes the nonlinear complementarity problem (NCP) [29, 30, 34], where $l_i = 0$ and $u_i = \infty$ for $i = 1, \dots, n$. In the NCP case, (1.1) is reduced to

$$F(x) = x - \max(0, x - g(x)) = 0. \tag{1.3}$$

Chen and Mangasarian [6] introduced a class of parametric smooth approximation functions for the nonsmooth function F in (1.3). A family of homotopy continuation methods, including an interior point path-following method, can be constructed to solve the NCP based on Chen–Mangasarian smooth approximation equations. Using the proving techniques developed in Kojima, Megiddo, and Noma [28], Chen and Harker [5] established the existence, uniqueness, and limiting properties of the trajectory defined by these approximation equations under certain conditions. Recently, Xu and Burke [40] gave a polynomial complexity bound for an interior point smoothing method for the monotone linear complementarity problem (LCP) based on the Chen–Harker–Kanzow–Smale function [4, 23] and Tseng's convergence analysis for a class of infeasible interior point methods [39]. More recently, Burke and Xu [2] and Hotta and Yoshise [22] studied the global convergence of non–interior point algorithms using the Chen–Harker–Kanzow–Smale function.

A number of important problems in engineering and economics have lower and/or upper bounds on the variables, rather than just nonnegative constraints as in the NCP. The VI(l, u, g) model includes more general problems and attracts researchers' attention [6, 9, 11, 15, 17, 18, 19, 24]. Gabriel and Moré [19] defined a parametric function $f\colon \Re^n \times R_{++} \to \Re^n$ to approximate function F in (1.1), where $f(x, \varepsilon)$ is continuously differentiable with respect to variable vector x and $\|F(x)-f(x,\varepsilon)\| \leq \mu\varepsilon$. Here μ is a positive constant.

In section 2, we reformulate $\mathrm{VI}(l,u,g)$ as a generalized complementarity problem (GCP) in $\Re^n$:

$$q(x) \geq 0, \quad p(x) \geq 0, \quad p(x)^T q(x) = 0,$$

where the set $\{x : \; p(x) \geq 0, \; q(x) \geq 0\}$ is usually called the feasible set of the GCP. We show that for any $\varepsilon > 0$, if x is a solution of $f(x,\varepsilon) = 0$, then x is approximately feasible for the GCP—$q(x) \geq 0$ and $p(x) \geq -O(\varepsilon)e$, and for each i, $|p_i(x)q_i(x)| \leq O(\varepsilon^2)$ (which is stronger than the current result that $|p_i(x)q_i(x)|$ is bounded). Furthermore, if the VIP or GCP is merely an NCP, then any solution of $f(x,\varepsilon) = 0$ is an exactly feasible point of the NCP, i.e., $x \geq 0$ and $g(x) \geq 0$, and for each i, $x_i g_i(x) \leq O(\varepsilon^2)$.

Under the assumptions that g is a P_0 function and the level sets $\{x \; : \; \|F(x)\| \leq \Gamma\}$ with $\Gamma > 0$ are bounded, we investigate the trajectory formed by solutions of $f(x,\varepsilon) = r$ as $\varepsilon > 0$ and $r \in \Re^n$ vary. We show that the monotone complementarity problem with a feasible interior point satisfies the assumptions on certain level sets. Our trajectory analysis is related to the one established in Chen and Harker [5]. However, there are several fundamental distinctions. For example, the Gabriel–Moré approximation function is not convex. Consequently, assumptions A1, A2, and A4 in [5], which play an important role in their proof, do not hold for the Gabriel–Moré approximation function.

In section 3, we modify a smoothing Newton method for solving $\mathrm{VI}(l,u,g)$, which was originally proposed in [9]. The new method simplifies the ε update step in the Chen–Qi–Sun method without loss of its feature that it solves a system of linear equations at each step and generates iterates converging to a solution of (1.1) globally and superlinearly when g is a uniform P function. Moreover, motivated by the finite convergence result of other algorithms [16, 42], we use a Newton step in the new method. The resulting method finds a solution of $\mathrm{VI}(l,u,g)$ in finite steps if the involved function g is an affine function.

It is notable that the existence of smooth path and global convergence of the smoothing method requires the same condition that g be a P_0 function and the level sets $\{x \; : \; \|F(x)\| \leq \Gamma\}$ with $\Gamma > 0$ be bounded. Like the central path in interior point methods, the smooth path plays a key role in smoothing methods although it does not appear in algorithms.

We have tested the hybrid Newton-smoothing method with three approximation functions on seven types of samples which contain 27 linear VIPs and 4 nonlinear VIPs. Our preliminary numerical results reported in section 4 show that the new method is promising.

Some words about our notation are in order. Let $N = \{i \; : \; i = 1,2,\ldots,n\}$. The $n \times n$ identity matrix is denoted by I. The ith row of an $n \times n$ matrix M is denoted by M_i. We use $\|\cdot\|$ to denote the Euclidean norm. Let e denote the vector of ones, i.e.,

$$e = (1,1,\ldots,1)^T.$$

We denote

$$\Re_+ = \{\varepsilon \; : \; \varepsilon \in \Re, \varepsilon \geq 0\} \quad \text{and} \quad \Re_{++} = \{\varepsilon \; : \; \varepsilon \in \Re, \varepsilon > 0\}.$$

2. Smoothing approximations. Let $\rho : \Re \to \Re_+$ be a density function with a bounded absolute mean, that is,

$$\kappa := \int_{-\infty}^{\infty} |s|\rho(s)ds < \infty. \tag{2.1}$$

In addition, let ρ satisfy

$$\zeta^2 := \max\left\{\sup_{t\geq 0}\int_t^{\infty} ts\rho(s)ds, \quad \sup_{t\leq 0}\int_{-\infty}^{t} ts\rho(s)ds\right\} < \infty. \tag{2.2}$$

It is easy to verify that

$$\zeta^2 \leq \int_{-\infty}^{\infty} s^2\rho(s)ds.$$

The three density functions used in our computational experiment satisfy these two conditions (see section 4). Note that if the second moment of ρ is bounded, then ρ satisfies (2.1) and (2.2). The condition (2.2) is new. We introduce it to establish a strong complementarity error bound in Theorem 2.1.

The class of Gabriel–Moré approximation functions to function F in (1.1) is defined by

$$f_i(x,\varepsilon) = x_i - \int_{-\infty}^{\infty} \text{mid}(l_i, u_i, x_i - g_i(x) - \varepsilon s)\rho(s)ds, \quad i = 1,\ldots,n. \tag{2.3}$$

By using (2.3), a smooth approximation function can be generated by an appropriate density function. The following proposition is known.

PROPOSITION 2.1 (see [7, 9, 19]). *The function $f : \Re^n \times \Re_{++} \to \Re^n$ defined by (2.3) satisfies the following four conditions:*

(i) *For any $(x,\varepsilon) \in \Re^n \times \Re_{++}$,*

$$|F_i(x) - f_i(x,\varepsilon)| \leq \kappa\varepsilon, \quad i = 1,2,\ldots,n. \tag{2.4}$$

(ii) *f is continuously differentiable with respect to the variable x, and for any $(x,\varepsilon) \in \Re^n \times \Re_{++}$,*

$$f_x(x,\varepsilon) = I - D(x)(I - g'(x)), \tag{2.5}$$

where $D(x)$ is a diagonal matrix whose diagonal elements are

$$D_{ii}(x) = \int_{(x-g(x)-u)_i/\varepsilon}^{(x-g(x)-l)_i/\varepsilon} \rho(s)ds, \quad i = 1,2,\ldots,n.$$

(iii) *For any $x \in \Re^n$,*

$$\lim_{\varepsilon\downarrow 0} f_x(x,\varepsilon) = f^0(x), \tag{2.6}$$

where for $i \in N$,

$$f_i^0(x) = \begin{cases} g_i'(x) & \textit{if } x_i - g_i(x) \in (l_i, u_i), \\ I_i & \textit{if } x_i - g_i(x)_i \notin [l_i, u_i], \\ I_i - \left(\int_{-\infty}^{0}\rho(s)ds\right)(I_i - g_i'(x)) & \textit{if } x_i - g_i(x) = l_i, \\ I_i - \left(\int_{0}^{\infty}\rho(s)ds\right)(I_i - g_i'(x)) & \textit{if } x_i - g_i(x) = u_i. \end{cases}$$

(iv) *For any* $x \in \Re^n$, f *satisfies the* directional derivative consistency *property at* x [7], *i.e.,*

$$\lim_{h\to 0} \frac{F(x+h) - F(x) - f^0(x+h)h}{\|h\|} = 0. \tag{2.7}$$

It is shown in [9] that $f^0(x) \in \partial_C F(x)$ for any $x \in \Re^n$, where $\partial_C F(x)$ is the generalized Jacobian at x. This property is called Jacobian consistency property in [9]. However, the definition of the generalized Jacobian and the semismoothness for superlinear convergence [9, 35] are restricted in $\Re^n$. In order to make smoothing methods applicable in function spaces, we use the definition of $f^0(x)$ and the directional derivative consistency property in this paper.

We now develop the following proposition which will be used in our analysis.

PROPOSITION 2.2. *For a given* $\varepsilon \in \Re_{++}$ *and a pair* $a, b \in \Re \cup \{-\infty, \infty\}$ *such that* $a < b$, *let* $\rho(s) = \rho(-s)$ *and*

$$h(t) = \int_{-\infty}^{\infty} \text{mid}(a, b, t - \varepsilon s)\rho(s)ds.$$

Then

$$0 \le h'(t) \le 1 \tag{2.8}$$

and

$$h(t) \ge \text{mid}(a, b, t) \quad \Leftrightarrow \quad t \le \frac{a+b}{2}. \tag{2.9}$$

Furthermore, if supp$\{\rho\} = \Re$, *then*

$$0 < h'(t) < 1 \tag{2.10}$$

and

$$h(t) > \text{mid}(a, b, t) \quad \Leftrightarrow \quad t < \frac{a+b}{2}. \tag{2.11}$$

Proof. The result on $h'(t)$ was proved in Lemma 2.3 of Gabriel and Moré [19]. We focus on proving the rest of the proposition.

Using $\rho(s) = \rho(-s)$, we have

$$\int_{-\infty}^{\infty} s\rho(s)ds = 0,$$

which implies

$$\int_{-\infty}^{\infty} (t - \varepsilon s)\rho(s)ds = t.$$

By the definition of the median operator, we obtain

$$h(t) = \int_{-\infty}^{(t-b)/\varepsilon} b\rho(s)ds + \int_{(t-a)/\varepsilon}^{\infty} a\rho(s)ds + \int_{(t-b)/\varepsilon}^{(t-a)/\varepsilon} (t - \varepsilon s)\rho(s)ds$$

$$= \int_{-\infty}^{(t-b)/\varepsilon} b\rho(s)ds + \int_{(t-a)/\varepsilon}^{\infty} a\rho(s)ds + \int_{-\infty}^{\infty} (t-\varepsilon s)\rho(s)ds$$

$$- \int_{-\infty}^{(t-b)/\varepsilon} (t-\varepsilon s)\rho(s)ds - \int_{(t-a)/\varepsilon}^{\infty} (t-\varepsilon s)\rho(s)ds$$

$$= \int_{-\infty}^{(t-b)/\varepsilon} (b-t+\varepsilon s)\rho(s)ds + \int_{(t-a)/\varepsilon}^{\infty} (a-t+\varepsilon s)\rho(s)ds + t. \tag{2.12}$$

Suppose that $t \leq (a+b)/2$. Then there are only two cases: $\mathrm{mid}(a,b,t) = t$ or $\mathrm{mid}(a,b,t) = a$.

If $\mathrm{mid}(a,b,t) = t$, then from (2.12), $b-t \geq t-a \geq 0$, and $\rho(s) = \rho(-s)$ we have

$$h(t) - \mathrm{mid}(a,b,t)$$
$$= \int_{-\infty}^{(t-b)/\varepsilon} (a+b-2t)\rho(s)ds + \int_{(t-a)/\varepsilon}^{(b-t)/\varepsilon} (a-t+\varepsilon s)\rho(s)ds \geq 0. \tag{2.13}$$

If $\mathrm{mid}(a,b,t) = a$, then from (2.12) and $0 \geq t-a \geq t-b$ we have

$$h(t) - \mathrm{mid}(a,b,t)$$
$$= \int_{-\infty}^{(t-b)/\varepsilon} (b-a+\varepsilon s)\rho(s)ds + \int_{(t-b)/\varepsilon}^{\infty} (t-a)\rho(s)ds + \int_{(t-a)/\varepsilon}^{\infty} (a-t+\varepsilon s)\rho(s)ds$$
$$= \int_{-\infty}^{(t-b)/\varepsilon} (b-a)\rho(s)ds + \int_{(t-b)/\varepsilon}^{(t-a)/\varepsilon} (t-a-\varepsilon s)\rho(s)ds \geq 0. \tag{2.14}$$

Hence, we obtain the $\Leftarrow$ part of (2.9) from (2.13) and (2.14).

Assume that $t \geq (a+b)/2$. Then following the same argument above, we can show that $h(t) \leq \mathrm{mid}(a,b,t)$. This implies the $\Rightarrow$ part of (2.9). So the proof of (2.9) is completed.

Notice that $\mathrm{supp}\{\rho\} = \Re$ implies that for any α, β satisfying $\alpha < \beta$,

$$\int_{\beta}^{\alpha} \rho(s)ds > 0.$$

Using this fact and following the above argument, we can prove (2.11). □

To study the smooth path defined by the smooth equations $f(x,\varepsilon) = 0$ for $\varepsilon > 0$, we introduce a generalized complementarity problem. Let $z \in \Re^n$ and E be a diagonal matrix whose diagonal elements are

$$E_{ii}(z) = \begin{cases} 1, & z_i \geq 0, \\ -1, & z_i < 0, \quad i = 1, 2, \ldots, n. \end{cases}$$

It is easy to verify that a vector $x \in \Re^n$ solves (1.1) if and only if x solves the following complementarity problem:

$$q(x) := \min(x-l, u-x) \geq 0, \tag{2.15}$$

$$p(x) := E\left(\frac{l+u}{2} - x\right) g(x) \geq 0, \tag{2.16}$$

$$q(x)^T p(x) = 0. \tag{2.17}$$

In the complementarity formulation (2.15)–(2.17), for $i \in N$ we regard $l_i \cdot 0 = 0$ and $u_i \cdot 0 = 0$ even for $l_i = -\infty$ or $u_i = \infty$.

Based on the complementarity problem (2.15)–(2.17), we define a restricted feasible set of the variational inequality problem VI(l, u, g) by

$$X_0 = \{x \ : \ q(x) \geq 0, \ p(x) \geq 0\}.$$

$q(x) \geq 0$ implies that $x \in [l, u]$, which is desired, but $p(x) \geq 0$ implies that $g(x) \geq 0$ if $x \leq \frac{l+u}{2}$ and $g(x) \leq 0$ if $x > \frac{l+u}{2}$, which is strictly contained in the feasible set of the original VIP

$$\{x \ : \ q(x) \geq 0, \ g_i(x) \geq 0 \text{ if } x_i = l_i; \qquad g_i(x) \leq 0 \text{ if } x_i = u_i, \ i = 1, \ldots, n\}.$$

Thus, we define another set X associated with X_0 as

$$X = \left\{x \ : \ q(x) \geq 0, \ E\left(\frac{l+u}{2} + g(x) - x\right) g(x) \geq 0\right\}$$

and its interior by

$$\text{int}X = \left\{x \ : \ q(x) > 0, \ E\left(\frac{l+u}{2} + g(x) - x\right) g(x) > 0\right\}.$$

Obviously, we have

$$X_0 \subseteq X,$$

and both X_0 and X contain the solution set of the VIP if it exists. Furthermore, if $l_i = 0$ and $u_i = \infty$ for each $i \in N$, then VI(l, u, g) reduces to the NCP and

$$X_0 = X = \{x \ : \ x \geq 0, \ g(x) \geq 0\},$$

which is known as the feasible set for the NCP.

The following theorem shows that the homotopy equation $f(x, \varepsilon) = 0$ possesses several desired features and is closely related to the interior point homotopy equation for the GCP (2.15)–(2.17).

THEOREM 2.1. *Suppose that $\rho(s) = \rho(-s)$. Let x be a solution of*

$$f(x, \varepsilon) = 0.$$

Then

(i) *For any $\varepsilon \in \Re_{++}$,*

$$x \in X.$$

Furthermore, if $\text{supp}(\rho) = \{s : \rho(s) > 0\} = \Re$, *then*

$$x \in \text{int}X \cup \left\{x \ : \ x_i = \frac{l_i + u_i}{2}, \ \exists \ i \in N\right\}.$$

(ii) *For $\varepsilon \in \Re_{++}$ satisfying $\varepsilon < \frac{1}{2\kappa} \min_{1 \leq i \leq n}(u_i - l_i)$,*

$$x \in X_\varepsilon = \{x \ : \ q(x) \geq 0, \quad p(x) \geq -\kappa\varepsilon e\}.$$

(iii) *For any* $\varepsilon \in \Re_{++}$,

$$|q_i(x)p_i(x)| \le \varepsilon^2(\zeta^2 + \kappa^2/2), \qquad i = 1, 2, \ldots, n.$$

Proof. To simplify the proof, we denote $y = g(x)$.

(i) Since $l_i \le$ mid $(l_i, u_i, x_i - y_i - \varepsilon s) \le u_i$ and $\rho(s) \ge 0$, we have

$$l_i \le x_i - f_i(x, \varepsilon) = x_i \le u_i, \quad i = 1, 2, \ldots, n.$$

Now we show

$$E_{ii}\left(\frac{l+u}{2} + y - x\right) y_i \ge 0, \quad i = 1, 2, \ldots, n. \tag{2.18}$$

By Proposition 2.2, $f(x, \varepsilon) = 0$ implies that for any $i \in N$

$$x_i \ge \text{mid}(l_i, u_i, x_i - y_i) \quad \Leftrightarrow \quad x_i - y_i \le \frac{l_i + u_i}{2}. \tag{2.19}$$

Suppose that $x_i - y_i \le \frac{l_i + u_i}{2}$. Then from (2.19), we have either

$$x_i \ge x_i - y_i \ge l_i \quad \text{or} \quad x_i \ge l_i \ge x_i - y_i.$$

In either case, $y_i \ge 0$.

Suppose that $x_i - y_i \ge \frac{l_i + u_i}{2}$. Then from (2.19), we have either

$$x_i \le x_i - y_i \le u_i \quad \text{or} \quad x_i \le u_i \le x_i - y_i.$$

In either case, $y_i \le 0$.

Therefore (2.18) holds, and hence $x \in X$.

The assumption supp$(\rho) = \Re$ implies that ρ vanishes in any nontrivial interval. In such a case, from

$$\begin{aligned} x_i &= x_i - f_i(x, \varepsilon) \\ &= \int_{-\infty}^{\frac{x_i - y_i - u_i}{\varepsilon}} u_i \rho(s) ds + \int_{\frac{x_i - y_i - l_i}{\varepsilon}}^{\infty} l_i \rho(s) ds + \int_{\frac{x_i - y_i - u_i}{\varepsilon}}^{\frac{x_i - y_i - l_i}{\varepsilon}} (x_i - y_i - \varepsilon s)\rho(s) ds \end{aligned}$$

we have $l_i < x_i < u_i$. Furthermore, using (2.11) and following the same argument above, we can show

$$E_{ii}\left(\frac{l+u}{2} + y - x\right) y_i > 0 \quad \text{for} \quad x_i - y_i \ne \frac{l_i + u_i}{2} \quad i = 1, 2, \ldots, n.$$

If $x_i - y_i = (l_i + u_i)/2$, then by (2.11) $y_i = 0$. Hence, we obtain $x \in$ int $X \cup \{x : x_i = (l_i + u_i)/2, \exists\, i \in N\}$.

(ii) By (2.4) and $f(x, \varepsilon) = 0$, we have

$$|F_i(x)| = |x_i - \text{mid}(l_i, u_i, x_i - y_i)| \le \kappa\varepsilon.$$

This implies

$$x_i - \kappa\varepsilon \le \text{mid}(l_i, u_i, x_i - y_i) \le x_i + \kappa\varepsilon. \tag{2.20}$$

If $x_i \le (l_i + u_i)/2$, then from the second inequality in (2.20) and $\varepsilon < \frac{u_i - l_i}{2\kappa}$, we have either

$$\text{mid}(l_i, u_i, x_i - y_i) = l_i \tag{2.21}$$

or

$$\text{mid}(l_i, u_i, x_i - y_i) = x_i - y_i. \tag{2.22}$$

Moreover, (2.21) implies that $y_i \ge x_i - l_i \ge 0$, and (2.22) with (2.20) implies $y_i \ge -\kappa\varepsilon$.

If $x_i \ge (l_i + u_i)/2$, then from the first inequality in (2.20) and $\varepsilon < \frac{u_i - l_i}{2\kappa}$ we have either

$$\text{mid}(l_i, u_i, x_i - y_i) = u_i \tag{2.23}$$

or

$$\text{mid}(l_i, u_i, x_i - y_i) = x_i - y_i. \tag{2.24}$$

Moreover, (2.23) implies that $y_i \le x_i - u_i \le 0$, and (2.24) with (2.20) implies $y_i \le \kappa\varepsilon$.

Hence, $p_i(x) = E_{ii}(\frac{l+u}{2} - x)y_i \ge -\kappa\varepsilon$.

(iii) For a fixed i, we consider three cases.

Case 1. $\text{mid}(l_i, u_i, x_i - y_i) = l_i$. We have

$$x_i - y_i \le l_i < u_i, \quad x_i - l_i \ge 0, \quad y_i \ge 0,$$

and

$$x_i - l_i \le \kappa\varepsilon.$$

Moreover,

$$\begin{aligned}
&x_i - l_i \\
&= \int_{-\infty}^{\frac{x_i - y_i - u_i}{\varepsilon}} (u_i - x_i + y_i + \varepsilon s)\rho(s)ds - \int_{-\infty}^{\frac{x_i - y_i - l_i}{\varepsilon}} (l_i - x_i + y_i + \varepsilon s)\rho(s)ds \\
&= \varepsilon \left(\int_{-\infty}^{\frac{x_i - y_i - u_i}{\varepsilon}} \left(\frac{u_i - x_i + y_i}{\varepsilon} + s \right) \rho(s)ds - \int_{-\infty}^{\frac{x_i - y_i - l_i}{\varepsilon}} \left(\frac{l_i - x_i + y_i}{\varepsilon} + s \right) \rho(s)ds \right) \\
&\le -\varepsilon \int_{-\infty}^{\frac{x_i - y_i - l_i}{\varepsilon}} \left(\frac{l_i - x_i + y_i}{\varepsilon} + s \right) \rho(s)ds \\
&\le -\varepsilon \int_{-\infty}^{\frac{x_i - y_i - l_i}{\varepsilon}} s\rho(s)ds,
\end{aligned}$$

since the first integral in the parenthesis has a negative value and $l_i - x_i + y_i \ge 0$. Thus,

$$\begin{aligned}
y_i(x_i - l_i) &\le (l_i - x_i + y_i + x_i - l_i) \left(-\varepsilon \int_{-\infty}^{\frac{x_i - y_i - l_i}{\varepsilon}} s\rho(s)ds \right) \\
&\le (l_i - x_i + y_i + \kappa\varepsilon) \left(-\varepsilon \int_{-\infty}^{\frac{x_i - y_i - l_i}{\varepsilon}} s\rho(s)ds \right) \\
&= \varepsilon^2 \frac{x_i - y_i - l_i}{\varepsilon} \int_{-\infty}^{\frac{x_i - y_i - l_i}{\varepsilon}} s\rho(s)ds - \varepsilon^2\kappa \int_{-\infty}^{\frac{x_i - y_i - l_i}{\varepsilon}} s\rho(s)ds \\
&\le \varepsilon^2 \left(\frac{\zeta^2 + \kappa^2}{2} \right).
\end{aligned}$$

Case 2. $\mathrm{mid}(l_i, u_i, x_i - y_i) = u_i$. The proof of Case 2 is similar to Case 1, where we have

$$l_i < u_i \le x_i - y_i, \quad u_i - x_i \ge 0, \quad y_i \le 0,$$

and

$$u_i - x_i \le \kappa\varepsilon.$$

Case 3. $\mathrm{mid}(l_i, u_i, x_i - y_i) = x_i - y_i$ and $l_i < x_i - y_i < u_i$. We have

$$|y_i| \le \kappa\varepsilon.$$

Moreover,

$$y_i = \int_{-\infty}^{\frac{x_i - y_i - u_i}{\varepsilon}} (u_i - x_i + y_i + \varepsilon s)\rho(s)ds + \int_{\frac{x_i - y_i - l_i}{\varepsilon}}^{\infty} (l_i - x_i + y_i + \varepsilon s)\rho(s)ds.$$

Since the first integral has negative value and the second has positive value, we have

$$\int_{-\infty}^{\frac{x_i - y_i - u_i}{\varepsilon}} (u_i - x_i + y_i + \varepsilon s)\rho(s)ds \le y_i \le \int_{\frac{x_i - y_i - l_i}{\varepsilon}}^{\infty} (l_i - x_i + y_i + \varepsilon s)\rho(s)ds.$$

Furthermore, since $u_i - x_i + y_i \ge 0$ and $l_i - x_i + y_i \le 0$, we have

$$\int_{-\infty}^{\frac{x_i - y_i - u_i}{\varepsilon}} \varepsilon s\rho(s)ds \le y_i \le \int_{\frac{x_i - y_i - l_i}{\varepsilon}}^{\infty} \varepsilon s\rho(s)ds.$$

Thus,

$$\begin{aligned}
\min(x_i - l_i, u_i - x_i)y_i &\le (x_i - l_i)\int_{\frac{x_i - y_i - l_i}{\varepsilon}}^{\infty} \varepsilon s\rho(s)ds \\
&= (x_i - l_i - y_i + y_i)\int_{\frac{x_i - y_i - l_i}{\varepsilon}}^{\infty} \varepsilon s\rho(s)ds \\
&\le (x_i - l_i - y_i + \kappa\varepsilon)\int_{\frac{x_i - y_i - l_i}{\varepsilon}}^{\infty} \varepsilon s\rho(s)ds \\
&= \varepsilon^2 \frac{x_i - l_i - y_i}{\varepsilon}\int_{\frac{x_i - y_i - l_i}{\varepsilon}}^{\infty} s\rho(s)ds + \varepsilon^2\kappa\int_{\frac{x_i - y_i - l_i}{\varepsilon}}^{\infty} s\rho(s)ds \\
&\le \varepsilon^2\left(\frac{\zeta^2 + \kappa^2}{2}\right).
\end{aligned}$$

Similarly, we can prove

$$\min(x_i - l_i, u_i - x_i)y_i \ge -\varepsilon^2(\zeta^2 + \kappa^2/2).$$

Therefore,

$$|\min(x_i - l_i, u_i - x_i)y_i| \le \varepsilon^2(\zeta^2 + \kappa^2/2). \qquad \square$$

Theorem 2.1 presents some properties of the smooth path

$$\{x(\varepsilon)\ :\ f(x, \varepsilon) = 0, \varepsilon > 0\}.$$

Result 1 shows that the smooth path is in the box $[l, u]$, i.e., $l \leq x(\varepsilon) \leq u$ and $l < x(\varepsilon) < u$ if $\operatorname{supp}(\rho) = \{s : \rho(s) > 0\} = \Re$. Result 2, together with result 1, shows that $x(\varepsilon)$ is approximately feasible, and the distance from $x(\varepsilon)$ to the feasible set goes to zero as ε goes to zero. Result 3 shows that the complementarity gap goes to zero quadratically in ε. Hence, the solution trajectory $x(\varepsilon)$, if it exists, converges to the solution set of the VIP as $\varepsilon \to 0$.

The following corollary further illustrates the behavior of the trajectory for NCP.

COROLLARY 2.1. *Suppose that $\rho(s) = \rho(-s)$ and consider the NCP problem, i.e., $l_i = 0$ and $u_i = \infty$ for each $i \in N$. Let x be a solution of*

$$f(x, \varepsilon) = 0.$$

Then we have the following:

(i) *For any $\varepsilon \in \Re_{++}$,*

$$x \in X.$$

Furthermore, if $\operatorname{supp}(\rho) = \{s : \rho(s) > 0\} = \Re$, *then*

$$x \in \operatorname{int} X.$$

(ii) *If*

$$\xi^2 := \inf_{t \leq 0} \left(\int_{-\infty}^{t} (t - s)\rho(s)ds \right) \left(\int_{-\infty}^{t} (t - s)\rho(s)ds - t \right) > 0, \tag{2.25}$$

then for any $\varepsilon \in \Re_{++}$ and each $i \in N$,

$$x_i g_i(x) \geq \epsilon^2 \xi^2.$$

Proof. Result (i) follows from Theorem 2.1. We now prove (ii), which is similar to the proof of (iii) of Theorem 2.1. (Note that Case 2 no longer exists here.)

Case 1. $y_i := g_i(x) \geq x_i$. We have

$$x_i - y_i \leq 0, \quad x_i \geq 0, \quad y_i \geq 0.$$

Let $t = \frac{x_i - y_i}{\varepsilon}$. Then $t \leq 0$, $y_i = x_i - \varepsilon t$, and

$$\begin{aligned} x_i &= -\int_{-\infty}^{\frac{x_i - y_i}{\varepsilon}} (-x_i + y_i + \varepsilon s)\rho(s)ds \\ &= \varepsilon \int_{-\infty}^{\frac{x_i - y_i}{\varepsilon}} \left(\frac{x_i - y_i}{\varepsilon} - s \right) \rho(s)ds \\ &= \varepsilon \int_{-\infty}^{t} (t - s)\rho(s)ds. \end{aligned}$$

Thus,

$$\begin{aligned} y_i x_i &= (x_i - \varepsilon t)x_i \\ &= \varepsilon^2 \left(\int_{-\infty}^{t} (t - s)\rho(s)ds \right) \left(\int_{-\infty}^{t} (t - s)\rho(s)ds - t \right) \\ &\geq \varepsilon^2 \xi^2. \end{aligned}$$

Similarly, we can prove Case 3 where $y_i \leq x_i$. □

The density function generating the Chen–Harker–Kanzow–Smale smooth approximation function satisfies condition (2.25). Moreover, it provides

$$\left(\int_{-\infty}^{t}(t-s)\rho(s)ds\right)\left(\int_{-\infty}^{t}(t-s)\rho(s)ds-t\right)\equiv 1.$$

By the proof of Corollary 2.1, we have

$$x_i g_i(x) = \varepsilon^2, \quad i = 1, 2, \dots, n,$$

for this density function. However, the other two density functions used in section 4 do not satisfy condition (2.25).

If both (iii) of Theorem 2.1 and (ii) of Corollary 2.1 hold, then for each $i \in N$,

$$\frac{x_i g_i(x)}{x^T g(x)/n} \geq \frac{\xi^2}{\zeta^2 + \kappa^2/2} > 0.$$

This inequality implies that $x(\varepsilon)$ converges to a maximal complementary solution (the number of positive components in x and $g(x)$ is maximal) if $g(x)$ is a monotone function (see [41] and references therein). This also implies that $x(\varepsilon)$ converges to a strictly complementary solution if $g(x)$ is monotone and the NCP has a strictly complementary solution, which property is possessed by most interior point algorithms.

We now turn our attention to the existence of the trajectory defined by

$$\{x(\varepsilon, r) \ : \ f(x, \varepsilon) = r, \quad \varepsilon > 0, r \in \Re^n\}.$$

The following two assumptions will be used to establish a sufficient condition for the existence of a solution to $f(x, \varepsilon) = r$, where r is a given n-dimensional vector.

A1. The level sets

$$D(\Gamma) = \{x \in \Re^n \ : \ \|F(x)\| \leq \Gamma\}$$

are bounded for all positive numbers Γ.

A2. For any $\varepsilon > 0$ and $x \in \Re^n$, $f_x(x, \varepsilon)$ is nonsingular.

Assumptions related to A1 and A2 have been used in several papers on smoothing methods. We state some results here.

PROPOSITION 2.3. *Any of the following conditions implies assumption* A1*:*

(i) *g is a uniform P function* [24]*;*

(ii) $-\infty < l_i < u_i < \infty$ *for* $i = 1, 2, \dots, n$ [9]*;*

(iii) $l_i = 0$ *for* $i = 1, 2, \dots, n$*, and g is an R_0 function* [5].

Any of the following conditions implies assumption A2*:*

(i) *g is a uniform P function* [19]*;*

(ii) *g is a P_0 function and* $\text{supp}\{\rho\} = \Re$ [19].

Many interior point algorithms (e.g., [28, 41]) for solving the NCP use the following assumption:

AIP. g is monotone and the set of all the strictly positive feasible solutions

$$S_{++}(g) = \{x \in \Re^n : g(x) > 0,\ x > 0\}$$

is nonempty.

The existence of a feasible interior point is the standard assumption for any interior point algorithm, and the monotonicity of g is necessary to have a convex objective for the LCP [41]

$$\begin{array}{rl} \min & x^T g(x), \\ \text{subject to} & g(x) = Mx + q, (x, g(x)) \geq 0. \end{array}$$

The existence and uniqueness of trajectories defined by

$$\{x(\varepsilon, 0) \ : \ f(x, \varepsilon) = 0, \varepsilon > 0\}$$

were established in [5, 28] under assumption AIP. It is interesting to see the relation between interior point methods and smoothing methods from the difference of assumptions A1, A2, and AIP.

Since the monotonicity is stronger than the P_0 property, the assumption AIP implies A2 for smooth functions with $\text{supp}\{\rho\} = \Re$. However, neither of the two assumptions, A1 and AIP, implies the other. A1 does not imply AIP, because the uniform P property satisfies A1 but fails to satisfy the monotonicity. Kanzow, Yamashita, and Fukushima [25] gave a simple example to show that AIP does not imply A1. In their example, $g(x) \equiv 1$, for $x \in \Re$. Then g is monotone and $S_{++}(g)$ is nonempty. However, for $\Gamma \geq 1$, $[1, \infty) \subset D(\Gamma)$.

Nevertheless, the following proposition shows that AIP implies the boundedness of certain level sets for the NCP.

Proposition 2.4. *Under assumption* AIP*, the level sets* $D(\Gamma)$ *are bounded for all positive numbers* $\Gamma < \Gamma_0$, *where*

$$F(x) := \min(x, g(x))$$

and

$$\Gamma_0 := \sup\left\{\min_{i \in N} F_i(x), \ x \in S_{++}(g)\right\}.$$

Proof. Let $\Gamma < \Gamma_0$ and $\hat{x} \in S_{++}(g)$ satisfy

$$\Gamma < \min_{i \in N} F_i(\hat{x}).$$

This means that for each $i \in N$, $g_i(\hat{x}) > \Gamma$ and $\hat{x}_i > \Gamma$.

Suppose on the contrary that we have an unbounded sequence $\{x^k\} \subset D(\Gamma)$. Since $x^k \in D(\Gamma)$ implies that $|\min(x_i^k, g_i(x^k))| \leq \Gamma$ for each $i \in N$, there is no index j such that $x_j^k \to -\infty$ or $g_j(x^k) \to -\infty$. Define the index sets $J = \{i \ : \ x_i^k \to \infty, \ i \in N\}$ and $L = \{\ell \ : \ g_\ell(x^k) \to \infty, \ \ell \in N\}$. Since $\{x^k\}$ is unbounded by assumption, the set J is nonempty, while the set L can be either empty or nonempty.

Then there is a $k' \geq 0$ such that for all $k \geq k'$,

$$|\min(x_i^k, g_i(x^k))| = |g_i(x^k)| \leq \Gamma \quad \text{for} \quad i \in J$$

and

$$|\min(x_\ell^k, g_\ell(x^k))| = |x_\ell^k| \leq \Gamma \quad \text{for} \quad \ell \in L \text{ (if } L \neq \emptyset).$$

Hence we have

$$x_i^k - \hat{x}_i \to \infty, \quad |g_i(x^k)| \leq \Gamma < g_i(\hat{x}) \quad \text{for} \quad k \geq k', \ i \in J$$

and

$$g_\ell(x^k) - g_\ell(\hat{x}) \to \infty, \quad |x^k_\ell| \le \Gamma < \hat{x}_\ell \quad \text{for} \quad k \ge k',\ \ell \in L \text{ (if } L \ne \emptyset).$$

Note this results in a contradiction to the monotonicity of g, i.e,

$$(g(x^k) - g(\hat{x}))^T(x^k - \hat{x}) \ge 0.$$

Thus $D(\Gamma)$ must be bounded. □

Subsequent to the writing of this paper, several interesting results on the boundedness of the level set $D(\Gamma)$ and the existence of an interior feasible point for the P_0-function NCP were established [3, 14, 37]. We summarize these results by the following theorem.

Theorem 2.2. *Suppose that g is a P_0 function and the solution set of the NCP is nonempty and bounded. Then we have the following:*

(i) $D(\Gamma)$ *is bounded for all sufficiently small* $\Gamma > 0$ [14, 37]*;*

(ii) $S_{++}(g)$ *is nonempty* [3, 37].

The assumption of Theorem 2.2 is weaker than assumption AIP, because every monotone function is a P_0 function and that $S_{++}(g)$ is nonempty implies that the solution set of the monotone NCP is nonempty and bounded. On the other hand, under assumption AIP, by Proposition 2.4, we can have a bounded level set for smoothing methods after we find an interior point (see Corollary 3.1). Precisely, if $\tilde{x} \in S_{++}(g)$, then

$$\min_{i\in N} F_i(\tilde{x}) \le \Gamma_0 = \sup\left\{\min_{i\in N} F_i(x), x \in S_{++}(g)\right\}.$$

Hence for every point x^0 satisfying

$$\|F(x^0)\| < \min_{i\in N} F_i(\tilde{x}),$$

the level set $D(\|F(x^0)\|)$ is bounded. However, we cannot define a bounded level set by using an interior point based on Theorem 2.2.

Theorem 2.3. *Suppose that assumptions* A1 *and* A2 *hold; then for any $\varepsilon > 0$ and $r \in \Re^n$, there is a solution $x(\varepsilon, r)$ of $f(x,\varepsilon) = r$. Furthermore, if g is a P_0 function and* $\text{supp}\{\rho\} = \Re$, *then, $x(\varepsilon, r)$ is unique.*

Proof. Let $\mu = \kappa\sqrt{n}$. We choose $x^0 \in \Re^n$ and a positive number Γ satisfying $\Gamma > \mu\varepsilon + \|r\|$ and $\Gamma - \mu\varepsilon - \|r\| > \|F(x^0)\|$. We define

$$C_{\varepsilon,r}(\Gamma) = \{x \ :\ \|f(x,\varepsilon) - r\| \le \Gamma\}.$$

By (2.4), we have

$$\|f(x,\varepsilon) - r\| \le \|F(x)\| + \mu\varepsilon + \|r\| \tag{2.26}$$

and

$$\|F(x)\| \le \|f(x,\varepsilon) - r\| + \mu\varepsilon + \|r\|. \tag{2.27}$$

Since $x^0 \in D(\Gamma - \mu\varepsilon - \|r\|)$, (2.26) implies that $x^0 \in C_{\varepsilon,r}(\Gamma)$, and hence $C_{\varepsilon,r}(\Gamma)$ is nonempty. By (2.27), we have

$$C_{\varepsilon,r}(\Gamma) \subseteq D(\Gamma + \mu\varepsilon + \|r\|).$$

Hence, from assumption A1, the level set $C_{\varepsilon,r}(\Gamma)$ is bounded.

Let $\theta(x) := \frac{1}{2}\|f(x,\varepsilon)-r\|^2$. The boundedness of $C_{\varepsilon,r}(\Gamma)$ and the continuous differentiability of $f(\cdot,\varepsilon)$ imply that θ has a global minimum point x^* in $C_{\varepsilon,r}(\Gamma)$ [33, Thm. 4.2.2]. Since $\theta(x^*) \le \theta(x^0)$ and $\|f(x^0,\varepsilon)-r\| < \Gamma$, $x^* \in \operatorname{int} C_{\varepsilon,r}(\Gamma)$. Moreover, by assumption A2, $f'_{\mathbf{1}}(x^*,\varepsilon)$ is nonsingular. Therefore, by [33, Thm. 4.1.3], $\theta'(x^*) = f'_{\mathbf{1}}(x^*,\varepsilon)^T(f(x^*,\varepsilon)-r) = 0$, and so $f(x^*,\varepsilon)=r$.

Now we prove the uniqueness of $x(\varepsilon,r)$.

Let $y=g(x)$, $t=x-y$, and

$$h_i(t) = \int_{-\infty}^{\infty} \operatorname{mid}(l_i,u_i,t_i-\varepsilon s)\rho(s)ds, \quad i=1,\dots n.$$

By Proposition 2.2, $h_i(t)$ is an increasing function—therefore, an increasing function in x_i and a decreasing function in y_i. Furthermore, by Lemma 2.3 in [19], we can claim $h'_i(t) \in (0,1)$.

Suppose on the contrary that we have both x^1 and x^2 as solutions. Let $y^1=g(x^1)$ and $y^2=g(x^2)$. Since $g(x)$ is a P_0 function, we have an index i such that

$$x_i^1 \ne x_i^2 \quad \text{and} \quad (y_i^1-y_i^2)(x_i^1-x_i^2) \ge 0.$$

Assume that, without loss of generality, $x_i^1 > x_i^2$. Then we have $y_i^1 \ge y_i^2$. Moreover,

$$\begin{aligned} & x_i^1 - x_i^2 \\ & = h_i(x_i^1-y_i^1) - h_i(x_i^2-y_i^2) \\ & \le h_i(x_i^1-y_i^2) - h_i(x_i^2-y_i^2) \\ & = h'_i(\theta)(x_i^1-x_i^2) \\ & < x_i^1-x_i^2, \end{aligned}$$

where θ is between $x_i^1-y_i^2$ and $x_i^2-y_i^2$ according to the Taylor theorem. Note that this results in a contradiction. Thus, we must have $x^1=x^2$. □

3. Algorithm and convergence. In this section, we propose a hybrid Newton-smoothing algorithm for solving VI(l,u,g). We show that the algorithm converges globally and superlinearly under assumptions A1 and A2. Furthermore, the method is finitely convergent if

$$g(x) = Mx+q,$$

where M is an $n\times n$ matrix and q is an n-dimensional vector. In this case, we denote the linear variational inequality problem by VI(l,u,M,q).

We denote

$$\Theta(x) = \frac{1}{2}\|F(x)\|^2$$

and

$$\theta_k(x) = \frac{1}{2}\|f(x,\varepsilon_k)\|^2.$$

ALGORITHM 3.1. *Given $\rho,\alpha,\eta \in (0,1)$ and a starting point $x^0 \in \Re^n$, choose a scalar $\sigma \in (0,\frac{1}{2}(1-\alpha))$. Let $\nu = \frac{\alpha}{2\sqrt{n}\kappa}$. Let $\beta_0 = \|F(x^0)\|$ and $\varepsilon_0 = \nu\beta_0$.*

For $k\ge 0$:

1. *Find a solution* $\hat{d}^k$ *of the system of linear equations*

$$F(x^k) + f^0(x^k)d = 0. \tag{3.1}$$

If $\|F(x^k + \hat{d}^k)\| \le \eta\beta_k$, *let* $x^{k+1} = x^k + \hat{d}^k$. *Otherwise perform step* 2.

2. *Find a solution* d^k *of the system of linear equations*

$$F(x^k) + f_x(x^k, \varepsilon_k)d = 0. \tag{3.2}$$

Let m_k *be the smallest nonnegative integer* m *such that*

$$\theta_k(x^k + \rho^m d^k) - \theta_k(x^k) \le -2\sigma\rho^m\Theta(x^k). \tag{3.3}$$

Set $t_k = \rho^{m_k}$ *and* $x^{k+1} = x^k + t_k d^k$.

3. 3.1. *If* $\|F(x^{k+1})\| = 0$, *terminate.*

 3.2. *If*

$$0 < \|F(x^{k+1})\| \le \max\{\eta\beta_k, \alpha^{-1}\|F(x^{k+1}) - f(x^{k+1}, \varepsilon_k)\|\}, \tag{3.4}$$

let

$$\beta_{k+1} = \|F(x^{k+1})\| \quad \textit{and} \quad \varepsilon_{k+1} = \min\left\{\nu\beta_{k+1}, \frac{\varepsilon_k}{2}\right\}.$$

 3.3. *Otherwise, let* $\beta_{k+1} = \beta_k$ *and* $\varepsilon_{k+1} = \varepsilon_k$.

Algorithm 3.1 is a modification of the smoothing method proposed in [9]. This algorithm simplifies the ε update step in [9] and has finite convergence property for linear box constrained VIP.

THEOREM 3.1. *Suppose that assumptions* A1 *and* A2 *hold. Then for any starting point* $x^0 \in \Re^n$, *Algorithm* 3.1 *is well defined and the generated sequence* $\{x^k\}$ *remains in* $D_0 := D((1+\alpha)\|F(x^0)\|)$ *and satisfies*

$$\lim_{k\to\infty} \|F(x^k)\| = 0. \tag{3.5}$$

Proof. By Lemma 3.1 in [9], there exists a finite nonnegative integer m_k such that (3.3) holds. Hence, Algorithm 3.1 is well defined.

Let

$$K = \{k \ : \ \|F(x^k + \hat{d}^k)\| \le \eta\beta_k, \quad k \ge 0\}.$$

By the construction of Algorithm 3.1, for $k \in K$, $x^{k+1} = x^k + \hat{d}^k$ and $\beta_{k+1} = \|F(x^{k+1})\| \le \eta\beta_k \le (1+\alpha)\|F(x^0)\|$. Hence, $x^{k+1} \in D_0$ for $k \in K$. Moreover, following the proof of Theorem 3.1 in [9], we can show $x^{k+1} \in D_0$ for $k \notin K$. Hence, $\{x^k\}$ remains in D_0.

Now we prove (3.5).

If K is finite, then there exists $k_0 \ge 0$ such that step 2 is performed for all $k \ge k_0$. By Theorem 3.1 in [9],

$$\lim_{\substack{k\to\infty \\ k\ge k_0}} \|F(x^k)\| = 0.$$

Hence the whole sequence $\{x^k\}$ satisfies (3.5).

If K is infinite, then by the construction of Algorithm 3.1,

$$\lim_{\substack{k\to\infty \\ k\in K}} \|F(x^k)\| = 0.$$

Let

$$\bar{K} = \{k \ : \ \|F(x^{k+1})\| \le \max\{\eta\beta_k, \alpha^{-1}\|F(x^{k+1}) - f(x^{k+1}, \varepsilon_k)\|\}\}.$$

Then $K \subseteq \bar{K}$. Assume that $\bar{K}$ consists of $k_0 = 0 < k_1 < k_2 < \cdots$.

Let k be an arbitrary nonnegative integer. Let k_j be the largest number in $\bar{K}$ such that $k_j \le k$. Then $\varepsilon_k = \varepsilon_{k_j}$ and $\beta_k = \beta_{k_j}$. By the line search rule

$$\|f(x^k, \varepsilon_{k_j})\| \le \|f(x^{k_j}, \varepsilon_{k_j})\|.$$

Hence by (2.4), for $j \ge 0$,

$$\begin{aligned} \|F(x^k)\| &\le \|f(x^k, \varepsilon_k)\| + \|F(x^k) - f(x^k, \varepsilon_k)\| \\ &= \|f(x^k, \varepsilon_{k_j})\| + \|F(x^k) - f(x^k, \varepsilon_{k_j})\| \\ &\le \|f(x^{k_j}, \varepsilon_{k_j})\| + \mu\varepsilon_{k_j} \\ &\le \|F(x^{k_j})\| + 2\mu\varepsilon_{k_j} \\ &= \beta_{k_j} + 2\mu\varepsilon_{k_j}, \end{aligned}$$

where $\mu = \kappa\sqrt{n}$.

Since K is infinite, $\bar{K}$ is infinite and hence $\beta_{k_j} \to 0$ and $\varepsilon_{k_j} \to 0$ as $j \to \infty$. Therefore, the whole sequence $\{x^k\}$ satisfies (3.5). □

COROLLARY 3.1. *Suppose that assumption AIP holds and* $l_i = 0, u_i = \infty$ *for each* $i \in N$. *Then for any smooth function* f *with* $\operatorname{supp}\{\rho\} = \Re$ *and any starting point* $x^0 \in \Re^n$*, Algorithm* 3.1 *is well defined and the generated sequence* $\{x^k\}$ *remains in* $D_0 := D((1+\alpha)\|F(x^0)\|)$. *Furthermore, if* x^0 *satisfies*

$$(3.6) \qquad (1+\alpha)\|F(x^0)\| < \sup\left\{\min_{i\in N} F_i(x),\ x \in S_{++}(g)\right\},$$

then $\lim_{k\to\infty} \|F(x^k)\| = 0$.

Proof. By Theorem 4.2 in [19], the monotonicity of g implies assumption A2. By Lemma 3.1 in [9], there is a finite nonnegative integer m_k such that (3.3) holds. Hence Algorithm 3.1 is well defined. Following the proof of Theorem 3.1 in [9] and Theorem 3.1 here, we can show $\{x^k\} \subset D_0$ for any starting point $x^0 \in \Re^n$. By Proposition 2.4, if x^0 satisfies (3.6), then D_0 is bounded. Therefore, we can show that $\lim_{k\to\infty} \|F(x^k)\| = 0$ by following the proof of Theorem 3.1. □

The following lemma shows that if an iterate is sufficiently close to a solution of VI(l, u, M, q), Algorithm 3.1 finds the solution in one step. We define

$$\begin{aligned} \gamma(x) = \min_{1\le i,j\le n} \{&|x - Mx - q - l|_i, |x - Mx - q - u|_j : \\ &(x - Mx - q - l)_i \ne 0, (x - Mx - q - u)_j \ne 0\}. \end{aligned}$$

Since $l_i < u_i$ for all $i \in N$, $\gamma(x) > 0$ for any $x \in \Re^n$.

LEMMA 3.1. *Let* $x^* \in \Re^n$ *be a solution of VI*(l, u, M, q). *Let*

$$(3.7) \qquad B := \begin{cases} \{x \in \Re^n \ : \ \|x - x^*\|_\infty \le \gamma(x^*)/\|I - M\|_\infty\} & \textit{if } I \ne M, \\ \Re^n & \textit{otherwise.} \end{cases}$$

Then for any $x \in B$,

$$F(x) + f^0(x)(x^* - x) = 0. \tag{3.8}$$

Proof. If $M = I$, then $f^0(x) = I$ for any $x \in \Re^n$. Moreover, x^* =mid$(l, u, -q)$ is a solution of $F(x) = 0$. Hence, in this case for any $x \in \Re^n$,

$$F(x) + f^0(x)(x^* - x) = x - \mathrm{mid}(l, u, -q) + (x^* - x) = 0.$$

Suppose that $I \neq M$ and $x \in B$. Then

$$\begin{aligned} |x - Mx - q - x^* + Mx^* + q|_i &\leq \|I - M\|_\infty \|x - x^*\|_\infty \\ &\leq \gamma(x^*). \end{aligned} \tag{3.9}$$

For a fixed i, we consider two cases.

Case 1. $(x^* - Mx^* - q)_i \neq l_i$ and $(x^* - Mx^* - q)_i \neq u_i$.

By the definition $\gamma(x)$ and (3.9), we have

$$(x^* - Mx^* - q)_i < l_i \quad \Rightarrow (x - Mx - q)_i < l_i, \tag{3.10}$$

$$l_i < (x^* - Mx^* - q)_i < u_i \quad \Rightarrow l_i < (x - Mx - q)_i < u_i, \tag{3.11}$$

and

$$(x^* - Mx^* - q)_i > u_i \quad \Rightarrow (x - Mx - q)_i > u_i. \tag{3.12}$$

If $(x^* - Mx^* - q)_i \in (l_i, u_i)$, then $(Mx^* + q)_i = 0$. By (3.11) and (2.6), $f_i^0(x) = M_i$ and

$$F_i(x) + f_i^0(x)(x^* - x) = x_i - (x - Mx - q)_i + M_i(x^* - x) = (Mx^* + q)_i = 0. \tag{3.13}$$

If $(x^* - Mx^* - q)_i < l_i$, then $x_i^* = l_i$. By (3.10) and (2.6), $f_i^0(x) = I_i$ and

$$F_i(x) + f_i^0(x)(x^* - x) = x_i - l_i + I_i(x^* - x) = x_i^* - l_i = 0. \tag{3.14}$$

Similarly, we can show

$$F_i(x) + f_i^0(x)(x^* - x) = x_i^* - u_i = 0,$$

for $(x^* - Mx^* - q)_i > u_i$.

Case 2. $(x^* - Mx^* - q)_i = l_i$ or $(x^* - Mx^* - q)_i = u_i$.

If $(x^* - Mx^* - q)_i = l_i$, then $x_i^* = l_i$ and $(Mx^* + q)_i = 0$.

If $(x - Mx - q)_i > l_i$ or $(x - Mx - q)_i < l_i$, we can find $F_i(x) + f_i^0(x)(x^* - x) = 0$ by following the same argument in (3.13) and (3.14). If $(x - Mx - q)_i = l_i$, then $f_i^0(x) = I_i - \lambda(I_i - M_i)$, where $\lambda \in (0, 1)$. Furthermore, we have

$$\begin{aligned} &F_i(x) + f_i^0(x)(x^* - x) \\ &= x_i - l_i + (I_i - \lambda(I_i - M_i))(x^* - x) \\ &= \lambda(M_i - I_i)(x^* - x) \\ &= \lambda(x_i - l_i + (Mx^* + q - Mx - q)_i) \\ &= \lambda((x - l - Mx - q)_i + (Mx^* + q)_i) = 0. \end{aligned}$$

The proof of the case $(x^* - Mx^* - q)_i = u_i$ is similar. □

THEOREM 3.2. *Suppose that* A1 *and* A2 *hold. Assume that for an accumulation point x^* of $\{x^k\}$, there is an open ball $\hat{B} := \hat{B}(x^*, \bar{r}) = \{x \ : \ \|x - x^*\| < \bar{r}\}$ and a positive number Υ such that for any $x \in \hat{B}$, $f^0(x)$ is nonsingular and $\|f^0(x)^{-1}\| \le \Upsilon$. Then x^* is a solution of* (1.1) *and $\{x^k\}$ converges to x^* superlinearly. Moreover, if g has a locally Lipschitz continuous derivative around x^*, the convergence rate is quadratic. In addition, if g is an affine function, the convergence is finite.*

Proof. By Theorem 3.1, x^* is a solution of (1.1). By Lemma 2.4 in [7], x^* is the unique solution of (1.1) in $\hat{B}$. Let K_* be a subsequence of $\{0, 1, \ldots\}$ such that

$$\lim_{\substack{k \to \infty \\ k \in K_*}} x^k = x^*.$$

By the directional derivative consistency property of f at x^*, for $k \in K_*$,

$$\begin{aligned}
&\|x^k + \hat{d}^k - x^*\| \\
&= \|x^k - f^0(x^k)^{-1}F(x^k) - x^*\| \\
&= \|f^0(x^k)^{-1}(f^0(x^k)(x^k - x^*) - F(x^k) + F(x^*))\| \\
&\le \Upsilon \|F(x^k) - F(x^*) - f^0(x^k)(x^k - x^*)\| \\
&= o(\|x^k - x^*\|).
\end{aligned} \tag{3.15}$$

Following the proof of Theorem 3.1 in [35],

$$\|F(x^k + \hat{d}^k)\| = o(\|F(x^k)\|).$$

This implies that there is $k_* \in K_*$ such that for all $k \ge k_*$,

$$\|F(x^k + \hat{d}^k)\| \le \eta \|F(x^k)\|$$

and $x^{k+1} = x^k + \hat{d}^k$. Hence, $\{x^k\}$ converges to x^* superlinearly.

Moreover, if g has a locally Lipschitz continuous derivative, F is directionally differentiable of degree 2 at x^*. Thus

$$\|F(x^k) - F(x^*) - f^0(x^k)(x^k - x^*)\| = O(\|x^k - x^*\|^2).$$

Following the argument in (3.15), we obtain the quadratic convergence rate.

The finite convergence follows from Lemma 3.1 and the fact that $k \in K_*$ for all $k \ge k_*$. □

Remark 3.1. If g is a uniform P function, by Theorem 4.3 in [19], for any $x \in \Re^n$, $f^0(x)$ is nonsingular. Hence, by Proposition 2.3, Algorithm 3.1 converges globally and superlinearly for VI(l, u, g), and converges globally and finitely for VI(l, u, M, q), assuming only that g is a uniform P function. It is notable that global convergence of Algorithm 3.1 requires only that g be a P_0 function and the level sets $\{x \ : \ \|F(x)\| \le \Gamma\}$ with $\Gamma > 0$ be bounded. Furthermore, Algorithm 3.1 converges to the solution set of the NCP if g is monotone, $S_{++}(g)$ is nonempty, and the starting point is in a certain level set.

4. Numerical experiments.

We numerically tested Algorithm 3.1 with seven examples containing 27 linear box-constrained VIPs and 4 nonlinear VIPs. We use three smooth approximation functions which are generated by (2.3) and density functions ρ satisfying (2.1) and (2.2) (cf. [5, 6, 7, 8, 9, 15, 19, 23, 36]).

Let

$$z = x - Mx - q.$$

Neural networks smooth approximation function (S1):

Let the density function be

$$\rho(x) = \frac{e^{-s}}{(1+e^{-s})^2}.$$

The smooth approximation function is

$$f_i(x,\varepsilon) = x_i - u_i - \varepsilon \log(1 + e^{(l_i - z_i)/\varepsilon}) + \varepsilon \log(1 + e^{(u_i - z_i)/\varepsilon})$$

and $f_x(x,\varepsilon) = I - D(x)(I - M)$, where

$$d_{ii}(x) = \frac{-e^{(l_i - z_i)/\varepsilon}}{1 + e^{(l_i - z_i)/\varepsilon}} + \frac{e^{(u_i - z_i)/\varepsilon}}{1 + e^{(u_i - z_i)/\varepsilon}}.$$

Chen–Harker–Kanzow–Smale smooth approximation function (S2):

Let the density function be

$$\rho(s) = \frac{2}{(s^2+4)^{\frac{3}{2}}}.$$

The smooth approximation function is

$$f_i(x,\varepsilon) = x_i - \frac{1}{2}\left(\sqrt{(l_i - z_i)^2 + 4\varepsilon^2} - \sqrt{(u_i - z_i)_i^2 + 4\varepsilon^2} + u_i + l_i\right)$$

and $f_x(x,\varepsilon) = I - D(x)(I - M)$, where

$$d_{ii}(x) = \frac{1}{2}\left(\frac{z_i - l_i}{\sqrt{(z_i - l_i)^2 + 4\varepsilon^2}} - \frac{z_i - u_i}{\sqrt{(z_i - u_i)^2 + 4\varepsilon^2}}\right).$$

Uniform smooth approximation function (S3):

Let the density function be

$$\rho(s) = \begin{cases} 1, & |s| \le 0.5, \\ 0 & \text{otherwise.} \end{cases}$$

For $0 < \varepsilon \le \min_{1 \le i \le n}\{u_i - l_i\}$,

$$f_i(x,\varepsilon) = \begin{cases} 1/2(Mx + q + x)_i + 1/2\varepsilon(u_i - z_i)^2 + \varepsilon/8 - u_i/2 & \text{if } |u_i - z_i| \le \varepsilon/2, \\ 1/2(Mx + q + x)_i - 1/2\varepsilon(l_i - z_i)^2 - \varepsilon/8 - l_i/2 & \text{if } |l_i - z_i| \le \varepsilon/2, \\ F_i(x) & \text{otherwise,} \end{cases}$$

and

$$f_x(x,\varepsilon) = \begin{cases} 1/2(I_i + M_i) + 1/\varepsilon(u_i - z_i)(M_i - I_i) & \text{if } |u_i - z_i| \le \varepsilon/2, \\ 1/2(I_i + M_i) - 1/\varepsilon(l_i - z_i)(M_i - I_i) & \text{if } |l_i - z_i| \le \varepsilon/2, \\ F_i'(x) & \text{otherwise.} \end{cases}$$

Table 1

Example 4.1, *iterations, total number of line search steps, CPU (sec.).*

$l = 0$, $u = 10^7 e$				
n	100	200	400	500
S1	2, 0, 0.3	2, 0, 2.1	2, 0, 20.5	2, 0, 45.0
S2	2, 0, 0.2	2, 0, 1.6	2, 0, 18.3	2, 0, 45.5
S3	2, 0, 0.2	2, 0, 1.5	2, 0, 10.0	2, 0, 30.9
$l = 0.5e$, $u = e$				
n	50	100	200	400
S1	2, 0, 0.1	2, 0, 0.3	2, 0, 1.8	2, 0, 12.5
S2	2, 0, 0.0	2, 0, 0.2	2, 0, 2.0	4, 0, 24.3
S3	2, 0, 0.0	2, 0, 0.2	2, 0, 1.9	2, 0, 7.3
$l = -e$, $u = e$				
n	50	100	200	300
S1	12, 32, 0.8	16, 57, 3.4	22, 103, 25.3	28, 143, 114.1
S2	12, 32, 0.3	18, 57, 2.2	22, 97, 20.8	28, 142, 106.0
S3	13, 35, 0.5	16, 58, 2.4	24, 107, 21.9	27, 141, 66.9
$l = -e$, $u = 0$				
n	50	100	200	400
S1	3, 0, 0.1	3, 0, 0.4	3, 0, 3.0	3, 0, 29.6
S2	3, 0, 0.1	3, 0, 0.3	3, 0, 2.9	3, 0, 29.6
S3	3, 0, 0.1	3, 0, 0.3	3, 0, 2.6	3, 0, 18.9
$l = 0$, $u_i = 1$ for i is even, $u_i = 10^7$ for i is odd.				
n	50	100	150	180
S1	2, 0, 0.1	2, 0, 0.2	2, 0, 0.8	2, 0, 1.5
S2	2, 0, 0.0	2, 0, 0.2	2, 0, 0.7	2, 0, 1.4
S3	2, 0, 0.0	2, 0, 0.2	2, 0, 0.8	2, 0, 1.0
$l_i = -1.0$, $u_i = 0.5$ for i is even, $l_i = -0.5$, $u_i = 1.0$ for i is odd.				
n	50	100	150	180
S1	14, 32, 0.7	18, 58, 3.0	25, 108, 26.4	26, 104, 17
S2	14, 30, 0.3	17, 53, 1.7	24, 101, 21.4	7, 10, 3.4
S3	14, 32, 0.5	18, 58, 2.2	25, 108, 19.2	45, 307, 24.4

The three density functions have a common property: $\rho(-s) = \rho(s)$. By (2.6), this implies that for any $x \in \Re^n$, the derivatives $f_x(x, \varepsilon)$ of the three smooth approximation functions have the same limit $f^0(x)$ as $\varepsilon \downarrow 0$, where

$$f_i^0(x) = \begin{cases} M_i & \text{if } z_i \in (l_i, u_i), \\ I_i & \text{if } z_i \notin [l_i, u_i], \\ 1/2(I_i + M_i) & \text{otherwise.} \end{cases} \tag{4.1}$$

Furthermore, the three density functions satisfy (2.1) and (2.2).

We chose $x^0 = e$, $\rho = 0.75$, $\alpha = 0.56$, $\eta = 0.87$, and $\sigma = 0.2$ in Algorithm 3.1. The stopping criterion was $\|F(x^k)\| \le 10^{-8}$. Numerical results were obtained using MATLAB 4.2c on a Sun 2000 workstation. These testing problems are constructed from some testing problems for LCP and NCP. We added different lower and upper bounds on the variables. We report the iterations k, the total numbers of line search steps $\sum_{i=1}^{k} m_i$, and CPU time in Tables 1–5 for solving these problems with different dimensions in Examples 4.1–4.5. We report the iterations k, the numbers of iterations where the Newton step in step 1 was accepted, and the number of iterations where the smooth step and line search were taken in step 2 as well as the function values at the final step $\|F(x^k)\|$ in Tables 6 and 7 for solving these problems in Examples 4.6 and 4.7. The three smooth approximation functions perform similarly in our numerical test.

Table 2

Example 4.2*, iterations, total number of line search steps, CPU* (*sec.*)*.*

$l = 0$, $u = 10^7 e$				
n	100	200	300	400
S1	8, 21, 1.1	9, 34, 7.7	8, 19, 24.5	9, 22, 81.0
S2	8, 21, 0.9	9, 34, 7.8	8, 19, 26.0	9, 22, 89.7
S3	8, 21, 0.8	9, 34, 5.3	8, 19, 12.9	9, 22, 40.9
$l = 0.5e$, $u = e$				
n	100	200	300	400
S1	2, 0, 0.2	2, 0, 1.4	2, 0, 5.6	2, 0, 16.2
S2	2, 0, 0.2	2, 0, 1.3	2, 0, 5.4	2, 0, 16.6
S3	2, 0, 0.2	2, 0, 1.0	2, 0, 3.0	2, 0, 9.2
$l = -e$, $u = 0$				
n	100	200	300	400
S1	3, 0, 0.3	3, 0, 2.2	3, 0, 8.4	3, 0, 24.2
S2	3, 0, 0.2	3, 0, 2.0	3, 0, 8.1	3, 0, 25.5
S3	3, 0, 0.2	3, 0, 1.4	3, 0, 4.5	3, 0, 13.4
$l = -10e$, $u = 5$				
n	64	128	256	320
S1	17, 69, 1.4	26, 147, 10.3	15, 207, 43.1	26, 187, 140.1
S2	12, 57, 0.5	11, 108, 3.2	25, 164, 66.3	20, 153, 100.4
S3	17, 71, 1.0	10, 101, 2.8	25, 180, 41.5	16, 236, 43.5
$l = 0$, $u_i = 1$ for i is even, $u_i = 10^7$ for i is odd.				
n	32	64	128	256
S1	5, 3, 0.1	7, 13, 0.4	6, 4, 2.0	8, 14, 19.1
S2	5, 3, 0.1	7, 13, 0.3	6, 4, 1.9	8, 14, 20.6
S3	5, 3, 0.1	7, 13, 0.3	6, 4, 1.3	8, 14, 10.1
$u = 0$, $l_i = -1.0$ for i is even, $l_i = -10^7$ for i is odd.				
n	32	64	128	256
S1	4, 0, 0.1	4, 0, 0.2	6, 0, 2.1	4, 0, 10.8
S2	4, 0, 0.0	4, 0, 0.2	6, 0, 1.9	4, 0, 10.4
S3	4, 0, 0.1	4, 0, 0.2	6, 0, 1.6	4, 0, 6.7

Example 4.1 (see Murty [32]).

$$M = \begin{pmatrix} 1 & 2 & 2 & \dots & 2 \\ 0 & 1 & 2 & \dots & 2 \\ 0 & 0 & 1 & \dots & 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}, \qquad q = -e.$$

Example 4.2 (see Fathi [13]).

$$M = \begin{pmatrix} 1 & 2 & 2 & \dots & 2 \\ 2 & 5 & 6 & \dots & 6 \\ 2 & 6 & 9 & \dots & 10 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 2 & 6 & 10 & \dots & 4(n-1)+1 \end{pmatrix}, \qquad q = -e.$$

Table 3

Example 4.3*, iterations, total number of line search steps, CPU* (*sec.*)*.*

$l = 0$, $u = 10^7 e$				
n	100	200	300	400
S1	3, 0, 0.3	3, 0, 2.2	3, 0, 8.8	3, 0, 24.7
S2	3, 0, 0.3	3, 0, 2.1	3, 0, 8.5	3, 0, 25.3
S3	3, 0, 0.2	3, 0, 1.4	3, 0, 4.6	3, 0, 14.8
$l = 0$, $u = e$				
n	100	200	300	400
S1	4, 2, 0.5	4, 2, 3.1	4, 2, 11.7	4, 2, 33.2
S2	4, 2, 0.3	4, 2, 2.8	4, 2, 11.3	4, 2, 33.0
S3	4, 2, 0.3	4, 2, 2.0	4, 2, 6.2	4, 2, 20.9
$l = -e$, $u = e$				
n	100	200	300	400
S1	2, 0, 0.2	2, 0, 1.3	2, 0, 5.7	2, 0, 16.2
S2	2, 0, 0.2	2, 0, 1.3	2, 0, 5.6	2, 0, 16.3
S3	2, 0, 0.1	2, 0, 0.9	2, 0, 3.0	2, 0, 8.6
$l = -e$, $u = 0$				
n	100	200	300	400
S1	3, 0, 0.3	3, 0, 2.2	3, 0, 8.6	3, 0, 24.0
S2	3, 0, 0.3	3, 0, 2.1	3, 0, 8.4	3, 0, 24.5
S3	3, 0, 0.2	3, 0, 1.5	3, 0, 4.6	3, 0, 13.8
$l_i = -1$, $u_i = 1$ for i is even, $l_i = 0$, $u_i = 10^7$ for i is odd.				
n	100	200	300	400
S1	2, 0, 0.2	2, 0, 1.6	2, 0, 6.0	2, 0, 22.3
S2	2, 0, 0.2	2, 0, 1.4	2, 0, 5.8	2, 0, 21.5
S3	2, 0, 0.2	2, 0, 1.0	2, 0, 3.1	2, 0, 12.9
$l_i = -1.0$, $u_i = 0.5$ for i is even, $l_i = -0.5$, $u_i = 1.0$ for i is odd.				
n	100	200	300	400
S1	3, 0, 0.3	3, 0, 2.3	3, 0, 8.6	3, 0, 30.6
S2	3, 0, 0.3	3, 0, 2.1	3, 0, 10.6	3, 0, 33.3
S3	3, 0, 0.2	3, 0, 1.4	3, 0, 4.6	3, 0, 18.6

Example 4.3 (see Ahn [1]).

$$M = \begin{pmatrix} 4 & -2 & 0 & 0 & \dots & 0 \\ 1 & 4 & -2 & 0 & \dots & 0 \\ 0 & 1 & 4 & -2 & 0\dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 1 & 4 \end{pmatrix}, \qquad q = -4e.$$

Example 4.4. *We randomly generated* 100% *dense* $A \in \Re^{n\times n}$ *and* $q \in \Re^n$ *whose elements distributed in* $(-5, 5)$*. We used the QR decomposition of* A *to get a upper triangular matrix* N*. Then we replaced the diagonal elements of* N *by their absolute values and obtained a triangular matrix* M *with positive diagonal elements. The matrix* M *is* P*-matrix.*

Example 4.5 (see Harker and Pang [21]). *We randomly generated* 100% *dense* $A \in \Re^{n\times n}$, $B \in \Re^{\frac{n}{2}\times\frac{n}{2}}$, $d \in \Re^n$, *and* $q \in \Re^n$, *where* $a_{ij} \in (-5, 5)$, $b_{i,j} \in (-5, 5)$, $d_i \in (0, 0.3)$, *and* $q_i \in (-500, 500)$*. We define a* P*-matrix as*

$$M = A^T A + \begin{pmatrix} 0 & B^T \\ -B & 0 \end{pmatrix} + \mathrm{diag}(d).$$

Table 4
Example 4.4, *iterations, total number of line search steps, CPU* (*sec.*).

$l = 0$, $u = 10^7 e$					
n		100	200	300	400
	Max	8, 12, 1.1	8, 3, 8.3	7, 0, 28.1	7, 0, 69.1
S1	Avg	7, 2, 0.8	7, 0, 5.2	7, 0, 27.7	7, 0, 67.1
	Min	5, 1, 0.6	6, 2, 6.4	6, 0, 24.8	6, 0, 62.5
	Max	9, 5, 0.8	9, 7, 8.9	8, 4, 31.4	7, 0, 67.6
S2	Avg	7, 5, 0.7	8, 6 6.6	7, 0, 28.6	7, 0, 67.5
	Min	6, 1, 0.6	7, 0, 4.9	6, 0, 23.9	6, 0, 60.0
	Max	8, 12, 0.8	12, 17, 10.4	7, 1, 21.1	7, 0, 39.1
S3	Avg	7, 4, 0.7	7, 2, 6.7	7, 0, 19.1	7, 0, 38.9
	Min	5, 1, 0.5	7, 0, 3.6	6, 0, 18.4	6, 0, 38.7
$l_i = -1, u_i = 1.0$ for i is even, $l_i = 0, u_i = 10^7$ for i is odd					
n		100	200	300	400
	Max	24, 78, 4.0	33, 180, 33.7	30, 133, 115.7	42, 202, 421.3
S1	Avg	22, 87, 3.9	32, 126, 31.4	28, 119, 109.2	41, 200, 342.1
	Min	22, 53, 3.4	28, 113, 26.8	27, 119, 103.5	41, 196, 337.1
	Max	20, 67, 1.8	35, 160, 26.4	42, 188, 157.4	41, 196, 398.7
S2	Avg	19, 58, 1.8	32, 135, 24.2	34, 146, 126.1	38, 193, 300.0
	Min	15, 30, 1.3	26, 101, 20.6	32, 145, 117.3	31, 147, 243.8
	Max	24, 90, 2.8	32, 141, 21.2	33, 149, 85.4	43, 203, 338.8
S3	Avg	19, 75, 2.4	27, 100, 16.8	32, 138, 82.4	40, 183, 206.1
	Min	14, 37, 1.6	25, 99, 15.9	27, 113, 70.2	37, 174, 190.1
$l_i = -100, u_i = 0.0$ for i is even, $l_i = 0, u_i = 100$ for i is odd					
n		100	200	300	400
	Max	9, 7, 1.2	10, 2, 7.3	13, 16, 40.4	16, 32, 132.2
S1	Avg	7, 0, 0.8	8, 0, 5.7	12, 14, 35.4	12, 18, 100.0
	Min	6, 0, 0.7	7, 1, 5.1	10, 14, 30.3	10, 4, 82.4
	Max	8, 5, 0.7	9, 4, 6.0	11, 8, 31.4	12, 18, 95.3
S2	Avg	7, 0, 0.6	8, 0, 5.4	10, 7, 29.4	12, 7, 93.5
	Min	6, 0, 0.5	7, 0, 4.8	10, 3, 27.9	11, 2, 88.2
	Max	7, 2, 0.8	10, 6, 5.5	15, 27, 27.1	16, 33, 69.6
S3	Avg	7, 1, 0.7	8, 1, 4.4	12, 18, 21.4	15, 31, 67.6
	Min	6, 0, 0.5	7, 1, 3.9	12, 16, 21.4	13, 14, 62.0

Example 4.6. *We randomly generated* 100% *dense* $A \in \Re^{n\times n}$ *and* $q \in \Re^n$ *whose elements distributed in* $(-5, 5)$. *We used the QR decomposition of* A *to get an orthogonal matrix* Q. *We randomly generated a diagonal matrix* $D \in \Re^{n\times n}$ *whose diagonal elements distributed in* $[0, 1)$ *with* 10% *zero elements. We set* $M = QDQ^T$. *Then the matrix* M *is positive semidefinite with* $n \times 10\%$ *zero eigenvalues.*

Example 4.7 (see the Kojima-Shindo NCP test problem [31]).

$$g(x) = \begin{pmatrix} 3x_1^2 + 2x_1x_2 + 2x_2^2 + x_3 + 3x_4 - 6 \\ 2x_1^2 + x_1 + x_2^2 + 10x_3 + 2x_4 - 2 \\ 3x_1^2 + x_1x_2 + 2x_2^2 + 2x_3 + 9x_4 - 9 \\ x_1^2 + 3x_2^2 + 2x_3 + 3x_4 - 3 \end{pmatrix}.$$

Remark 4.1. Both Lemke's complementarity pivot algorithm and the Cottle and Dantzig principal pivoting method are known to run in exponential time for Examples 4.1 and 4.2 with $l_i = 0, u_i = \infty$ (i.e., the LCP). Tables 1 and 2 show that the hybrid Newton-smoothing method takes only a few iterations for the LCP. It is notable that the LCP is relatively easier than the problem having different lower and upper bounds. To see the reason, we consider Example 4.1. The solution for the LCP is

$$x^* = (0, 0, \ldots, 0, 1)^T$$

Table 5

Example 4.5*, iterations, total number of line search steps, CPU* (*sec.*)*.*

$l = 0$, $u = 10^7 e$					
n		50	100	200	300
S1	Max	15, 114, 1.3	19, 132, 4.4	7, 0, 6.8	7, 0, 30.7
	Avg	9, 33, 0.6	9, 25, 2.1	6, 0, 6.7	7, 0, 29.3
	Min	5, 1, 0.2	7, 0, 1.2	6, 0, 5.9	6, 0, 24.8
S2	Max	31, 271, 1.0	48, 534, 7.1	7, 0, 6.4	7, 0, 30.2
	Avg	22, 210, 1.0	36, 295, 6.1	6, 0, 6.0	7, 0, 29.0
	Min	19, 144, 0.6	7, 0, 1.1	6, 0, 5.7	6, 0, 25.5
S3	Max	14, 66, 0.9	15, 67, 2.8	7, 0, 5.4	7, 0, 22.2
	Avg	6, 1, 0.2	10, 7, 2.1	6, 0, 4.9	7, 0, 20.2
	Min	5, 0, 0.2	7, 0, 1.1	6, 0, 4.7	6, 0, 16.4

$l = -10^7$, $u = 0$					
n		100	200	300	400
S1	Max	23, 171, 6.0	8, 0, 10.7	10, 19, 36.0	10, 21, 101.3
	Avg	14, 60, 3.0	8, 0, 10.0	7, 0, 24.4	8, 0, 77.3
	Min	7, 0, 0.9	7, 0, 9.4	7, 0, 23.7	7, 0, 23.7
S2	Max	18, 57, 24.5	8, 0, 9.9	10, 10, 34.5	10, 3, 99.3
	Avg	12, 26, 1.3	8, 0, 9.5	7, 0, 23.6	8, 0, 72.9
	Min	7, 0, 0.7	7, 0, 8.4	7, 0, 23.6	8, 0, 72.7
S3	Max	23, 145, 4.5	8, 0, 10.1	10, 19, 22.5	10, 21, 58.8
	Avg	14, 55, 2.2	8, 0, 9.8	7, 0, 15.0	8, 0, 45.5
	Min	7, 0, 0.8	7, 0, 9.2	7, 0, 14.6	8, 0, 42.3

$l_i = 0$, $u_i = 10^7$ for i is even, $l_i = -10^7$, $u_i = 0$ for i is odd.					
n		100	200	300	400
S1	Max	7, 0, 0.9	7, 0, 8.9	7, 0, 24.2	8, 0, 78.4
	Avg	6, 0, 0.8	7, 0, 8.8	7, 0, 23.9	8, 0, 77.2
	Min	6, 0, 0.7	6, 0, 7.5	6, 0, 20.4	7, 0, 68.2
S2	Max	7, 0, 0.8	7, 0, 8.3	7, 0, 23.5	8, 0, 73.7
	Avg	6, 0, 0.7	7, 0, 8.2	7, 0, 22.9	8, 0, 73.1
	Min	6, 0, 0.6	6, 0, 7.5	6, 0, 20.1	7, 0, 61.4
S3	Max	7, 0, 0.8	7, 0, 8.9	7, 0, 15.7	8, 0, 46.8
	Avg	6, 0, 0.7	7, 0, 8.7	7, 0, 14.4	8, 0, 45.4
	Min	6, 0, 0.6	6, 0, 7.1	6, 0, 12.7	7, 0, 40.4

Table 6

Example 4.6*, iterations, number of iterations accepted in step* 1*, and number of iterations used in step* 2 *with line search steps (LS).*

$l = 0$, $u = 10^{14}$, $n = 200$						
	k	Step 1	Step 2	LS	$\|F(x^k)\|$	CPU
S1	4	3	1	1	1.5×10^{-14}	3.2
S2	4	3	1	1	1.3×10^{-14}	3.0
S3	4	3	1	1	1.3×10^{-14}	2.9
$l = -e$, $u = e$, $n = 300$						
S1	9	2	7	8	1.7×10^{-14}	27.6
S2	8	3	5	5	1.4×10^{-14}	23.2
S3	10	5	5	6	1.6×10^{-14}	28.7
$l = -10^{14}$, $u = e$, $n = 400$						
S1	9	9	0	0	4.6×10^{-14}	73.4
S2	9	9	0	0	4.6×10^{-14}	73.4
S3	9	9	0	0	4.6×10^{-14}	77.2

TABLE 7

Example 4.7, *iterations, number of iterations accepted in step* 1, *and number of iterations used in step* 2 *with line search steps (LS).*

$l = 0$, $u = 10^{14}e$						
	k	Step 1	Step 2	LS	$\|F(x^k)\|$	CPU
S1	5	4	1	3	2.1×10^{-10}	0.03
S2	6	4	2	4	1.1×10^{-13}	0.03
S3	5	4	1	3	2.1×10^{-10}	0.02
$l = (0, -1, -2, -3)$, $u = 2e$						
S1	10	4	6	20	3.3×10^{-9}	0.06
S2	9	4	5	17	7.5×10^{-11}	0.06
S3	9	4	5	20	0	0.06
$l = (0, -10, -200, -3000)$, $u = (2, 20, 200, 3000)$						
S1	7	5	2	10	2.2×10^{-11}	0.05
S2	9	6	3	8	9.9×10^{-16}	0.05
S3	7	5	2	10	1.0×10^{-9}	0.05
$l = (-1, 1, -2, 3)$, $u = (3, 2, 1, 4)$						
S1	2	2	0	0	0	0.03
S2	2	2	0	0	0	0.03
S3	2	2	0	0	0	0.03

and

$$Mx^* + q = (1, 1, \dots, 1, 0)^T.$$

In this case $q_i(x) + p_i(x) > 0$, $i = 1, 2, \dots, n$, i.e., x^* is a strictly complementary solution. The function F is differentiable at x^*.

For $0 < l < u = 1$, the solution for $\text{VI}(l, u, M, q)$ is

$$x^* = (l_1, \dots, l_{n-1}, u)$$

and

$$(Mx^* + q)_i > 0, \quad i = 1, \dots, n-1, \quad (Mx^* + q)_n = 0.$$

In this case, $q_i(x) + p_i(x) > 0$, $i = 1, 2, \dots, n-1$, and $q_n(x) + p_n(x) = 0$.

This problem is relatively easy to solve, if $l \geq 0$ or $u \leq 0$. However, this problem becomes hard if $l < 0$ but $u > 0$. For example, if $l_i = -1$ and $u_i = 1$ for each $i \in N$, the solution is

$$x^* = (1, -1, 1, -1, \dots, 1) \quad \text{if } n \text{ is odd}$$

or

$$x^* = (-1, 1, -1, 1, \dots, 1) \quad \text{if } n \text{ is even},$$

and

$$Mx^* + q = 0.$$

This means $p_i(x) + q_i(x) = 0$ for $i = 1, \dots, n$.

Acknowledgments. The authors would like to thank J.S. Pang for his comment on limiting behavior of trajectories, D. Sun for his helpful comments, and the referees for their constructive suggestions.

REFERENCES

[1] B.H. Ahn, *Iterative methods for linear complementarity problem with upperbounds and lowerbounds*, Math. Programming, 26 (1983), pp. 265–315.

[2] J.V. Burke and S. Xu, *The global linear convergence of a non-interior path-following algorithm for linear complementarity problems*, Math. Oper. Res., to appear.

[3] B. Chen, X. Chen, and C. Kanzow, *A Penalized Fischer-Burmeister NCP-Function: Theoretical Investigation and Numerical Results*, Department of Management and Systems, Washington State University, Pullman, WA, 1997.

[4] B. Chen and P.T. Harker, *A non-interior-point continuation method for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1168–1190.

[5] B. Chen and P.T. Harker, *Smooth approximations to nonlinear complementarity problems*, SIAM J. Optim., 7 (1997), pp. 403–420.

[6] C. Chen and O.L. Mangasarian, *A class of smoothing functions for nonlinear and mixed complementarity problems*, Comput. Optim. Appl., 5 (1996), pp. 97–138.

[7] X. Chen, *Superlinear convergence of smoothing quasi-Newton methods for nonsmooth equations*, J. Comput. Appl. Math., 80 (1997), pp. 105–126.

[8] X. Chen and L. Qi, *A parameterized Newton method and a Broyden-like method for solving nonsmooth equations*, Comput. Optim. Appl., 3 (1994), pp. 157–179.

[9] X. Chen, L. Qi, and D. Sun, *Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities*, Math. Comp., 67 (1998), pp. 519–540.

[10] X. Chen and T. Yamamoto, *Convergence domains of certain iterative methods for solving nonlinear equations*, Numer. Funct. Anal. Optim., 10 (1989), pp. 37–48.

[11] S.P. Dirkse and M.C. Ferris, *The PATH solver: A non-monotone stabilization scheme for mixed complementarity problems*, Optim. Methods Softw., 5 (1995), pp. 123–156.

[12] B.C. Eaves, *On the basic theorem of complementarity*, Math. Programming, 1 (1971), pp. 68–75.

[13] Y. Fathi, *Computational complexity of LCPs associated with positive definite matrices*, Math. Programming, 17 (1979), pp. 335–344.

[14] F. Facchinei, *Structural and stability properties of P_0 nonlinear complementarity problems*, Math. Oper. Res, to appear.

[15] F. Facchinei, A. Fischer, and C. Kanzow, *A semismooth Newton method for variational inequalities: The case of box constraints*, in Complementarity and Variational Problems: State of the Art, M.C. Ferris and J.S. Pang, eds., SIAM, Philadelphia, PA, 1997, pp. 76–90.

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

[17] M.C. Ferris and J.S. Pang, *Engineering and economic applications of complementarity problems*, SIAM Rev., 39 (1997), pp. 669–713.

[18] S.A. Gabriel, *A hybrid smoothing method for mixed nonlinear complementarity problems*, Comput. Optim. Appl., 9 (1998), pp. 153–173.

[19] S.A. Gabriel and J.J. Moré, *Smoothing of mixed complementarity problems*, in Complementarity and Variational Problems: State of the Art, M.C. Ferris and J.S. Pang, eds., SIAM, Philadelphia, PA, 1997, pp. 105–116.

[20] P.T. Harker and J.-S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications*, Math. Programming, 48 (1990), pp. 161–220.

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

[22] K. Hotta and A. Yoshise, *Global Convergence of a Class of Non-Interior-Point Algorithms Using Chen-Harker-Kanzow Functions for Nonlinear Complementarity Problems*, Discussion Paper Series 708, Institute of Policy and Planning Sciences, University of Tsukuba, Tsukuba, Ibaraki 305, Japan, December 1996.

[23] C. Kanzow, *Some noninterior continuation methods for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 851–868.

[24] C. Kanzow and M. Fukushima, *Theoretical and numerical investigation of the D-gap function*

for box constrained variational inequalities, Math. Programming, to appear.

[25] C. Kanzow, N. Yamashita, and M. Fukushima, *New NCP-functions and their properties*, J. Optim. Theory Appl., 94 (1997), pp. 115–135.

[26] C.T. Kelley, *Identification of the support of nonsmoothness*, in Large Scale Optimization: State of the Art, W.W. Hager, D.W. Hearn, and P.M. Pardalos, eds., Kluwer Academic Publishers B.V., Boston, 1993, pp. 192–205.

[27] C.T. Kelley and E.W. Sachs, *Multilevel algorithms for constrained compact fixed point problems*, SIAM J. Sci. Comput., 15 (1994), pp. 645–667.

[28] M. Kojima, N. Megiddo, and T. Noma, *Homotopy continuation methods for nonlinear complementarity problems*, Math. Oper. Res., 16 (1991), pp. 754–774.

[29] M. Kojima, N. Megiddo, T. Noma, and A. Yoshise, *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*, Lecture Notes in Comput. Sci. 538, Springer-Verlag, New York, 1991.

[30] M. Kojima, N. Megiddo, and Y. Ye, *An interior point potential reduction algorithm for the linear complementarity problem*, Math. Programming, 54 (1992), pp. 267–279.

[31] M. Kojima and S. Shindo, *Extensions of Newton and quasi-Newton methods to systems of PC^1 equations*, J. Oper. Res. Soc. Japan, 29 (1986), pp. 352–374.

[32] K.G. Murty, *Linear Complementarity, Linear and Nonlinear Programming*, Sigma Ser. Appl. Math. 3, Heldermann-Verlag, Berlin, 1988.

[33] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[34] F. Potra and Y. Ye, *Interior-point methods for nonlinear complementarity problems*, J. Optim. Theory Appl., 88 (1996), pp. 617–642.

[35] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.

[36] L. Qi and X. Chen, *A globally convergent successive approximation method for severely nonsmooth equations*, SIAM J. Control Optim., 33 (1995), pp. 402–418.

[37] G. Ravindran and M.S. Gowda, *Regularization of P_0-Functions in Box Variational Inequality Problems*, Tech. report, Department of Mathematics and Statistics, University of Maryland, Baltimore, MD, 1997.

[38] S.M. Robinson, *Newton's method for a class of nonsmooth functions*, Set-Valued Anal., 2 (1994), pp. 291–305.

[39] P. Tseng, *Simplified analysis of an $o(nL)$-iteration infeasible predictor-corrector path-following method for monotone LCP*, in Recent Trends in Optimization Theory and Applications, R.P. Agarwal, ed., World Scientific Press, Singapore, 1994, pp. 423–434.

[40] S. Xu and J.V. Burke, *A Polynomial Time Interior-Point Path-Following Algorithm for LCP Based on Chen-Harker-Kanzow Smoothing Techniques,* preprint, Department of Mathematics, University of Washington, Seattle, WA, 1996.

[41] Y. Ye and K. Anstreicher, *On quadratic and $O(\sqrt{N}L)$ convergence of a predictor-corrector algorithm for LCP*, Math. Programming, 62 (1993), pp. 537–552.

[42] Y. Ye, *On the finite convergence of interior-point algorithms for linear programming*, Math. Programming, 57 (1992), pp. 325–335.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 2, pp. 617–652

EXACT PENALIZATION OF MATHEMATICAL PROGRAMS WITH EQUILIBRIUM CONSTRAINTS[*]

STEFAN SCHOLTES[†] AND MICHAEL STÖHR[‡]

Abstract. We study theoretical and computational aspects of an exact penalization approach to mathematical programs with equilibrium constraints (MPECs). In the first part, we prove that a Mangasarian–Fromovitz-type condition ensures the existence of a stable local error bound at the root of a real-valued nonnegative piecewise smooth function. A specification to nonsmooth formulations of equilibrium constraints, e.g., complementarity conditions or normal equations, provides conditions which guarantee the existence of a nonsmooth exact penalty function for MPECs. In the second part, we study a trust region minimization method for a class of composite nonsmooth functions which comprises exact penalty functions arising from MPECs. We prove a global convergence result for the general method and incorporate a penalty update rule. A further specification results in an SQP trust region method for MPECs based on an ℓ_1 penalty function.

Key words. bilevel program, error bound, exact penalization, ℓ_1 penalty function, Mangasarian–Fromovitz constraint qualification, mathematical program with equilibrium constraints (MPEC), piecewise smooth, sequential quadratic programming, trust region

AMS subject classifications. 90C30, 90C33

PII. S0363012996306121

1. Introduction.

Mathematical programs with equilibrium constraints (MPECs) are mathematical programs which involve, besides the usual differentiable equality and inequality constraints, a so-called equilibrium constraint. Such MPECs arise naturally in different areas, such as economics, computational mechanics, neural network training, and traffic control, and have been the subject of a number of recent studies (e.g., [1, 19, 20, 21, 22, 24, 25, 26, 34]). An equilibrium constraint requires that some of the variables, the state variables, satisfy an equilibrium condition which, however, changes with the remaining so-called design variables. The equilibrium condition is usually formulated in terms of a variational inequality induced by a parametric vector field $F(.,y)$ over a set $S(y)$, where y is the design variable. Recall that a vector x satisfies the variational inequality for a fixed design vector y if $x \in S(y)$ and $F(x,y)^\top(z-x) \geq 0$ for every $z \in S(y)$. If $S(y)$ has a functional representation

$$S(y) = \{x \in \mathbb{R}^n \mid h(x,y) = 0, g(x,y) \leq 0\},$$

where g, h are smooth vector valued functions, then the variational inequality is often replaced by the corresponding stationarity condition

$$\begin{aligned} F(x,y) + \nabla_x h(x,y)^\top \mu + \nabla_x g(x,y)^\top \lambda &= 0, \\ h(x,y) &= 0, \\ g(x,y) &\leq 0, \\ \lambda &\geq 0, \\ g(x,y)^\top \lambda &= 0. \end{aligned} \tag{1.1}$$

[*] Received by the editors July 3, 1996; accepted for publication (in revised form) April 20, 1998; published electronically January 28, 1999.
http://www.siam.org/journals/sicon/37-2/30612.html

[†] Department of Engineering and Judge Institute of Management Studies, University of Cambridge, Cambridge CB2 1PZ, UK.

[‡] Universität Karlsruhe, Institut für Statistik und Mathematische Wirtschaftstheorie, 76128 Karlsruhe, Germany.

Two special cases are of particular interest. If, on the one hand, $F(.,y)$ is the gradient with respect to the state variable x of a function $f(.,y)$, then condition (1.1) corresponds to the Karush–Kuhn–Tucker (KKT) conditions of the parametric mathematical program $\min\{f(x,y) \mid x \in S(y)\}$. If, on the other hand, the function h is absent and $g(x,y) = -x$, then the multipliers λ and μ can be eliminated and (1.1) turns into the parametric nonlinear complementarity problem

$$\begin{aligned} F(x,y) &\geq 0, \\ x &\geq 0, \\ x^\top F(x,y) &= 0. \end{aligned} \tag{1.2}$$

We shall assume throughout that the functions F, g, and h are smooth so that (1.1) and (1.2) are systems of smooth equations and inequalities. If all further constraints of the MPEC are smooth equations or inequalities, then the problem turns into a smooth nonlinear program and one may want to apply the well-developed methods for such problems. Unfortunately, any set of constraints which involves a complementarity condition of the general form

$$\Phi(z) \geq 0, \quad \Psi(z) \geq 0, \quad \Phi(z)^\top \Psi(z) = 0 \tag{1.3}$$

is inherently unstable so that standard methods are likely to fail in the presence of round-off errors. The instability is due to the fact that the Mangasarian–Fromovitz constraint qualification, which is necessary for the stability of a nonlinear program [14, 35], is violated at any feasible point of the constraint set [4].

It is possible to circumvent these difficulties by reformulating system (1.1) as a system of nonsmooth equations

$$\begin{aligned} F(x,y) + \nabla_x h(x,y)^\top \mu + \nabla_x g(x,y)^\top \lambda &= 0, \\ h(x,y) &= 0, \\ \min\{-g(x,y), \lambda\} &= 0. \end{aligned} \tag{1.4}$$

From the viewpoint of nonsmooth optimization, there is no severe problem with this set of equations, i.e., constraint qualifications are satisfied under reasonable conditions [18].

If the constraint set S for the variational inequality is closed, convex, and independent of y, one may avoid the explicit use of the multipliers λ and μ in (1.1) and instead use an alternative nonsmooth formulation of the variational inequality either as a direct normal equation

$$\Pi_S(x - F(x,y)) - x = 0$$

or as Robinson's normal equation [38]

$$\Pi_S(z) - F(\Pi_S(z), y) - z = 0.$$

The solutions of the two equations are related via $x = \Pi_S(z)$ and $z = x - F(x,y)$.

Our aim in this paper is to study the classical nonsmooth exact penalization technique for optimization problems which involve nonsmooth formulations of equilibrium conditions. In the next section, we recall some results from the theory of piecewise smooth functions which we shall employ in the subsequent sections, and we give a short review of the principle of exact penalization. In section 3 we derive conditions for the existence of local error bounds for piecewise smooth constraints. By specifying these conditions to piecewise smooth formulations of equilibrium constraints we obtain exact penalization results for MPECs. In section 4 we present a globally convergent trust region method for the minimization of the type of composite piecewise smooth functions which arise, e.g., as exact penalty functions for MPECs. We include a penalty update rule and specify the method to MPECs with complementarity constraints in section 5. The study is complemented by some preliminary numerical results.

2. Preliminaries.

2.1. Piecewise smooth functions. In this first section we review some properties of piecewise differentiable functions which we shall use in subsequent sections. For a detailed account we refer to [41]. A continuous function $f\colon U \to \mathbb{R}^m$ defined on an open set $U \subseteq \mathbb{R}^n$ is said to be a PC^r-function at $x^0 \in U$ if there exists an open neighborhood $V \subseteq U$ of x^0 and a finite family of C^r-functions $f_1, \ldots, f_l\colon V \to \mathbb{R}^m$ such that

$$f(x) \in \{f_1(x), \ldots, f_l(x)\} \qquad \forall x \in V. \tag{2.1}$$

The function f is called piecewise affine (piecewise linear) if $U = V = \mathbb{R}^n$ and (2.1) holds for affine (linear) functions $f_1, \ldots, f_l$. A collection of functions $f_1, \ldots, f_l$ satisfying (2.1) is called a collection of *selection functions* of f at x^0 and f is said to be a continuous selection of the functions $f_1, \ldots, f_l$ on V. A selection function f_i is called *essentially active* at x^0 if $x^0 \in \mathrm{clint}\{x \in V \mid f(x) = f_i(x)\}$, where $\mathrm{cl}S$ and $\mathrm{int}S$ denote the closure and interior, respectively, of a set S. The subcollection of essentially active selection functions is still a collection of selection functions at x^0, i.e., f is a continuous selection of the essentially active selection functions on a possibly smaller open neighborhood $\tilde{V}$ of x^0 [41].

The function f is called a PC^r-function if it is a PC^r-function at every point of its domain. When we use the term PC^r-function, we tacitly assume that $r \geq 1$. The term *piecewise differentiable function* refers to a PC^1-function. Some properties of piecewise differentiable functions are summed up in the following proposition.

Proposition 2.1. *Let $U \subseteq \mathbb{R}^n$ be open and $f : U \to \mathbb{R}^m$ be a piecewise differentiable function with C^1-selection functions $f_1, \ldots, f_p\colon V \to \mathbb{R}^m$ in a neighborhood $V \subseteq U$ of $x^0 \in U$.*

1. *f is Lipschitzian and B-differentiable in the sense of* [37] *in a neighborhood of x^0 and its B-derivative $f'(x^0;.)$ is a continuous selection of the F-derivatives of the essentially active selection functions at x^0.*
2. *f is semismooth in the sense of* [23, 30].
3. *The set of all Jacobians of the essentially active selection functions of f at x coincides with the set*

$$\{M \in \mathbb{R}^{m\times n} \quad | \quad \exists x^k \to x \text{ such that } f \text{ is F-differentiable at } x^k \text{ and } \nabla f(x^k) \to M\},$$

the convex hull of which is Clarke's generalized Jacobian [5].

A proof of the first statement can be found in [41]. The second statement has been proved in [3] (cf. also [30, Cor. 2.4]). The final statement is an immediate consequence of the first statement and the definitions involved.

The notion of coherent orientation plays an important role in connection with inverse and implicit function theorems for piecewise differentiable functions. We slightly extend the standard definition and call a collection of $p \times q$ matrices $M_1, \ldots, M_k$ with $p \leq q$ *coherently oriented* if there exists a $(q-p) \times q$ matrix A such that all matrices

$$\begin{bmatrix} M_i \\ A \end{bmatrix}, \qquad i = 1, \ldots, k, \tag{2.2}$$

have the same nonvanishing determinantal sign. In the case $p = q$ the matrix A is superfluous. We shall employ the following implicit function theorem from [33] in the sequel.

Theorem 2.2. *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be a PC^r function with C^r-selection functions $f_1, \ldots, f_p : U \times V \to \mathbb{R}^n$ in a neighborhood of $(x^0, y^0) \in U \times V$ and let $z^0 = f(x^0, y^0)$. Then the following two statements are equivalent:*

1. *There exist neighborhoods $\tilde{U}, \tilde{V}, \tilde{W}$ of x^0, y^0, and z^0, respectively, such that the equation*

$$f(x, y) = z$$

has a unique solution $x(y, z)$ in $\tilde{U}$ for every $(y, z) \in \tilde{V} \times \tilde{W}$.

2. *The partial Jacobians with respect to x of the essentially active selection functions of f at (x^0, y^0) are coherently oriented and the piecewise linear equation*

$$f'((x^0, y^0); (u, v)) = w$$

has a unique solution $u(v, w)$ for every $(v, w) \in \mathbb{R}^m \times \mathbb{R}^n$.

Moreover, if either of the statements is true, then the solution function $x(.,.)$ is a PC^r function and $x'((y^0, z^0); (v, w)) = u(v, w)$.

The following result is an immediate consequence of the foregoing theorem and Corollary 19 of [33].

Corollary 2.3. *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be a PC^r function with C^r-selection functions $f_1, \ldots, f_p : U \times V \to \mathbb{R}^n$ in a neighborhood of $(x^0, y^0) \in U \times V$ and let $z^0 = f(x^0, y^0)$. Let $\mathcal{M}(x^0, y^0)$ be the collection of all matrices whose jth row coincides with the jth row of the partial Jacobian $\nabla_x f_i(x^0, y^0)$ of some essentially active selection function f_i of f at (x^0, y^0). If the matrices in $\mathcal{M}(x^0, y^0)$ have the same nonvanishing determinantal sign, then statement* 1 *of Theorem* 2.2 *holds.*

2.2. The principle of exact penalization.

The exact penalization approach toward constrained optimization problems

$$\min\{f(x) \mid x \in C\}$$

goes back to Eremin [8] and Zangwill [48]. It aims at replacing the constrained problem by an equivalent unconstrained problem by augmenting the objective function f through the addition of a term which penalizes infeasibility. From a geometric point of view, infeasibility is most naturally measured in terms of the distance

$$d_C(y) = \min\{\|x - y\| \mid x \in C\}$$

of the point y to the closed set C. The following seminal result of Clarke shows that the distance function d_C is indeed an appropriate tool for exact penalization. Recall that a function $f : S \to \mathbb{R}$ is said to be Lipschitz of rank K on S if

$$|f(y) - f(z)| \leq K\|y - z\|$$

for all $y, z \in S$.

THEOREM 2.4. [5] *Let $x \in S \subseteq \mathbb{R}^n$ and let $C \subseteq S$ be nonempty and closed. Suppose $f : S \to \mathbb{R}$ is Lipschitz of rank K on S and let $\hat{K} > K$. Then x is a global minimizer of f over C if and only if x is a global minimizer of the function $f + \hat{K} d_C$ over S.*

A result analogous to Theorem 2.4 for local minima is readily obtained using the following well-known properties of the distance function which are direct consequences of the triangle inequality.

LEMMA 2.5. *Let $C \subseteq \mathbb{R}^n$ be nonempty and closed.*

1. *If $d_C(y) = \|\bar{x} - y\|$ for some $\bar{x} \in C$, then $d_C(\alpha y + (1-\alpha)\bar{x}) = \alpha d_C(y)$ for every $\alpha \in [0, 1]$.*
2. *If $x \in C$, $B(x, \epsilon)$ is a closed ball around x with positive radius ϵ and $y \in B(x, \frac{\epsilon}{2})$, then $d_C(y) = d_{C \cap B(x,\epsilon)}(y)$.*

COROLLARY 2.6. *Let $x \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz of rank K in a ball around x which intersects a closed and nonempty set $C \subseteq \mathbb{R}^n$. If $\hat{K} > K$, then x is a local minimizer of f over C if and only if x is an unconstrained local minimizer of $f + \hat{K} d_C$.*

Proof. First we show that every local minimizer of $f + \hat{K} d_C$ is contained in C. Suppose x is a local minimizer of $f + \hat{K} d_C$ and let $\bar{x} \in C$ be such that $\|x - \bar{x}\| = d_C(x)$. Then there exists a scalar $\alpha \in [0, 1)$ such that

$$\begin{aligned} f(x) + \hat{K} d_C(x) &\le f(\alpha x + (1-\alpha)\bar{x}) + \hat{K} d_C(\alpha x + (1-\alpha)\bar{x}) \\ &= f(\alpha x + (1-\alpha)\bar{x}) + \hat{K} \alpha d_C(x), \end{aligned}$$

where the last equation is a consequence of the first part of Lemma 2.5. Since f is Lipschitz of rank K in a ball around x which intersects C and thus contains $\bar{x}$, we further obtain

$$f(\alpha x + (1-\alpha)\bar{x}) - f(x) \le (1-\alpha) K \|x - \bar{x}\|.$$

Putting the foregoing two inequalities together, we arrive at

$$(1-\alpha) K \|x - \bar{x}\| \ge f(\alpha x + (1-\alpha)\bar{x}) - f(x) \ge (1-\alpha) \hat{K} \|x - \bar{x}\|.$$

Since $\hat{K} > K$ and $\alpha < 1$ we conclude that $x = \bar{x} \in C$. Hence if x is a local minimizer of $f + \hat{K} d_C$, then it is contained in C and, since f and $f + \hat{K} d_C$ coincide on C, it is a local minimizer of f over C.

To see the reverse implication, suppose that x is a local minimizer of f over C. Then there exists a closed ball $B(x, \epsilon)$ around x with positive radius ϵ such that x is a global minimizer of f over $C \cap B(x, \epsilon)$ and f is Lipschitz of rank K on $B(x, \epsilon)$. Hence Theorem 2.4 implies that x is a global minimizer of $\tilde{g} = f + \hat{K} d_{C \cap B(x,\epsilon)}$ on $B(x, \epsilon)$ and thus an application of the second part of Lemma 2.5 shows that x is a local minimizer of $f + \hat{K} d_C$. □

Although the foregoing results are theoretically very appealing, they are only of limited practical value since the mere evaluation of the penalty function involves the solution of a constrained optimization problem. Thus, nothing is won in passing from the constrained problem $\min\{f(x) \mid x \in C\}$ to the unconstrained problem $\min\{(f + \hat{K} d_C)(x) \mid x \in \mathbb{R}^n\}$. One is therefore interested in finding upper bounds for the distance function in terms of functions which are easier to evaluate. Such majorants can again be used as penalization terms as pointed out in the following corollary. In [20] this fact is called the *principle of exact penalization.*

COROLLARY 2.7. *If the assumptions of Theorem* 2.4 *hold and if* $\psi: S \to \mathbb{R}$ *is a function such that*

1. $\psi(y) \geq d_C(y)$ *for every* $y \in S$,
2. $\psi(y) = d_C(y)$ *for every* $y \in C$,

then x *is a global minimizer of* f *over* C *if and only if* x *is a global minimizer of* $f + \hat{K}\psi$ *over* S.

The foregoing statement is indeed an immediate corollary of Theorem 2.4 because the functions $f + \hat{K}d_C$ and $f + \hat{K}\psi$ coincide on C and the latter function majorizes the former over S, so that, by Theorem 2.4, the sets of global minimizers of both functions coincide. The foregoing principle is the basis for many exact penalty results. We refer to the surveys in [2, 20] and the references therein. Notice that a function ψ which satisfies the majorization assumptions

1. $\psi(y) \geq d_C(y)$ for every $y \in S$,
2. $\psi(y) = d_C(y)$ for every $y \in C$,

where C is a closed set contained in the interior of a set S, is necessarily nonsmooth at every $x \in C$ which is a distance minimizer of some point $y \neq x$. In fact, if $d_C(y) = \|y-x\|$ for some $y \neq x$, then Lemma 2.5 shows that $d_C(x+\lambda(y-x)) = \lambda d_C(y)$ for every $\lambda \in [0,1]$ and hence the function d_C has positive slope $d_C(y)$ at x in the direction $y-x$, while a smooth majorant ψ would have zero slope at x in all directions because x is a minimizer of ψ.

Unfortunately, a direct pendant of the local exact penalization result of Corollary 2.6 does not hold with d_C being replaced by a majorant ψ. In fact, the simple one-dimensional example $f(x) = 0$, $C = \{0\}$ and $\psi(x) = |x| + |2x^2 \sin 1/x|$ shows that the function $f + \hat{K}\psi$ may have strict local minima outside C that are arbitrarily close to a global minimum in C. However, the following result, which requires the local minimizer x to be contained in C, holds without additional assumptions on ψ (cf., e.g., [6, 20]).

COROLLARY 2.8. *If the assumptions of Corollary* 2.6 *hold and if* $\psi: \mathbb{R}^n \to \mathbb{R}$ *is a function such that for some neighborhood* U *of* x

1. $\psi(y) \geq d_C(y)$ *for every* $y \in U$,
2. $\psi(y) = d_C(y)$ *for every* $y \in C \cap U$,

then x *is a local minimizer of* f *over* C *if and only if* x *is an unconstrained local minimizer of* $f + \hat{K}\psi$ *and* $x \in C$.

Proof. Since f and $f + \hat{K}\psi$ coincide on $C \cap U$, x is a local minimizer of f over C, provided it is a local minimizer of $f + \hat{K}\psi$. Conversely, if x is a local minimizer of f over C, then it is a local minimizer of $f + \hat{K}d_C$, by Corollary 2.6, and thus it is a local minimizer of $f + \hat{K}\psi$ which coincides with $f + \hat{K}d_C$ at $x \in C$ and majorizes the latter function in a neighborhood of x. □

A pendant of the result of Corollary 2.6 with d_C replaced by a majorant ψ holds under the additional, rather strong assumption that

$$\liminf_{z \to y} \frac{\psi(z) - \psi(y)}{\|z - y\|} \leq -1 \tag{2.3}$$

for every y not contained in C. In fact, if under the assumptions of Corollary 2.6 the point x is a local minimizer of $f + \hat{K}\psi$, then $f(x) + \hat{K}\psi(x) \leq f(z) + \hat{K}\psi(z)$ for every z close to x. Hence if f is Lipschitz of rank $K < \hat{K}$, then

$$\frac{\psi(z) - \psi(x)}{\|z - x\|} \geq \frac{f(x) - f(z)}{\hat{K}\|x - z\|} \geq \frac{-K\|x - z\|}{\hat{K}\|x - z\|} > -1.$$

Since (2.3) holds if x is not contained in C we conclude that $x \in C$. Note that, in view of the first part of Lemma 2.5, $\psi = d_C$ satisfies (2.3) for every $y \notin C$.

3. Local error bounds. In this section we focus attention on sets of the form

$$C = \{z \in \mathbb{R}^n \mid \phi(z) = 0\},$$

where ϕ is a continuous nonnegative and real-valued function. Any closed set C can be represented in this way with $\phi = d_C$. However, the function ϕ in this representation may be much easier to evaluate than the distance function. If, e.g., C is of the form

$$C = \{z \in \mathbb{R}^n \quad | \quad \begin{array}{ll} h_i(z) = 0, & i = 1, \dots, m, \\ g_j(z) \leq, 0 & j = 1, \dots, k\} \end{array} \tag{3.1}$$

for continuous functions h_i and g_j, then C is the set of roots of the continuous nonnegative function

$$\phi(z) = \max\{|h_1(z)|, \dots, |h_m(z)|, g_1(z), \dots, g_k(z)\}. \tag{3.2}$$

If the function ϕ is easy to evaluate, then a positive multiple $\gamma\phi$ is a desirable candidate for the penalization of infeasibility. The principle of exact penalization implies that $\gamma\phi$ is an admissible penalty term in a neighborhood of a root z^0 if

$$d_{\phi^{-1}(0)}(z) \leq \gamma\phi(z)$$

for all z in a neighborhood of z^0. If that is the case for a scalar γ, then the function ϕ is said to admit a *local error bound* at the root z^0. A continuous parametric family of nonnegative functions $\phi_t = \phi(.,t)$ is said to admit a *stable local error bound* at the root (z^0, t^0) if there exists a scalar γ such that

$$d_{\phi_t^{-1}(0)}(z) \leq \gamma\phi(z,t)$$

for every (z,t) in a neighborhood of (z^0, t^0). For the sake of simplicity we shall focus on finite dimensional parameters t. Robinson [35] has shown that the function ϕ in (3.2) admits a local error bound at a root z^0 if the functions h_i, g_j are smooth and the constraint set (3.1) satisfies the Mangasarian–Fromovitz constraint qualification at z^0. Moreover, this local error bound is stable if the functions h_i and g_j are embedded in Lipschitzian families of functions.

Notice, however, that Robinson's result does not apply to MPECs with complementarity constraints since, as mentioned before, any set of constraints that includes a complementarity constraint necessarily violates the Mangasarian–Fromovitz constraint qualification at any feasible point. In [20, Chap. 2] Luo, Pang, and Ralph give conditions which ensure the existence of error bounds for complementarity constraints. In particular, they give conditions which ensure that the function

$$\phi(x,y) = \| \min\{F(x,y), x\} \|$$

admits an error bound. However, since they obtain bounds which are global with respect to x, their conditions have to be rather restrictive. In fact, their conditions imply that the solution of the complementarity problem is a single valued Lipschitz continuous function of the parameters y in a given compact set (cf. [20, Lem. 2.3.15]). The aim of this section is to develop conditions which ensure the existence of a local error bound without an implicit function assumption. The conditions to be developed will be related to the Mangasarian–Fromovitz-type condition of [18] and to the piecewise Mangasarian–Fromovitz constraint qualification studied in [20, 39]. To this end, we shall first extend Robinson's result for functions ϕ of the form (3.2) to arbitrary piecewise differentiable functions ϕ. This will then enable us to derive local error bound results for constraint sets involving piecewise smooth formulations of equilibrium conditions and, a fortiori, exact penalization results for MPECs.

3.1. Local error bounds for piecewise smooth functions. The starting point for our discussion is the following lemma, which is a direct consequence of more general results in [36].

LEMMA 3.1. *If $\phi : \mathbb{R}^n \to \mathbb{R}$ is a nonnegative piecewise affine function, then it admits a local error bound at every root.*

The trivial example $\phi(z) = |z|$ versus $\tilde{\phi}(z) = |z^3|$ with $z^0 = 0$ shows that the local error bound property is not a topological property, i.e., it depends on the chosen local coordinate system. However, the property is preserved under local Lipschitzian coordinate changes. This fact has been implicitly used in the proof of Theorem 1 of [11].

LEMMA 3.2. *Let $\phi : \mathbb{R}^n \to \mathbb{R}$ be a nonnegative continuous function, $W \subseteq \mathbb{R}^n$, and γ a positive scalar, and suppose that*

$$d_{\phi^{-1}(0)}(\zeta) \leq \gamma\phi(\zeta) \quad \forall\zeta \in W.$$

If $\Psi : W \to \mathbb{R}^n$ is Lipschitzian of rank L on W and maps W homeomorphically onto U, then

$$d_{(\phi\circ\Psi^{-1})^{-1}(0)}(z) \leq \gamma L(\phi \circ \Psi^{-1})(z) \quad \forall z \in U.$$

Proof. Let $z \in U$, $\zeta = \Psi^{-1}(z) \in W$, and $\xi \in \phi^{-1}(0)$ with $d_{\phi^{-1}(0)}(\zeta) = \|\zeta - \xi\|$. Then $x = \Psi(\xi)$ is a root of $\phi \circ \Psi^{-1}$ and hence

$$\begin{aligned} d_{(\phi\circ\Psi^{-1})^{-1}(0)}(z) &\leq \|z - \Psi(\xi)\| \\ &= \|\Psi(\zeta) - \Psi(\xi)\| \\ &\leq L\|\zeta - \xi\| \\ &= L d_{\phi^{-1}(0)}(\zeta) \\ &\leq \gamma L\phi(\zeta) \\ &= \gamma L(\phi \circ \Psi^{-1})(z). \qquad \square \end{aligned}$$

Clarke's inverse function theorem [5] can be used in connection with Lemmas 3.1 and 3.2 to prove the following stable local error bound result, which is the main result of this section. The main idea of the proof is borrowed from [13], where a similar argument is used in a different context.

THEOREM 3.3. *Let $\phi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be a nonnegative continuous function and let $(z^0, t^0) \in \mathbb{R}^n \times \mathbb{R}^m$ be a root of ϕ. Suppose that at every point (z, t) in a neighborhood of (z^0, t^0) the function ϕ coincides with at least one of the functions*

$$s_i(z, t) = \sum_{j=1}^{q} \alpha_{ij} h_j(z, t) + \beta_i g(z, t) + \rho_i, \qquad i = 1, \ldots, l,$$

where α_{ij}, β_i, and ρ_i are scalars, $h_1, \ldots, h_q$ are C^1 functions, and g is a locally Lipschitz continuous function. Suppose the following conditions hold:

1. *The gradients $\nabla_z h_1(z^0, t^0), \ldots, \nabla_z h_q(z^0, t^0)$ are linearly independent.*
2. *There exists a vector $v \in \mathbb{R}^n$ with $\nabla_z h_1(z^0, t^0)v = \cdots = \nabla_z h_q(z^0, t^0)v = 0$ and $w^\top v < 0$ for every $w \in \partial_z g(z^0, t^0)$, where ∂_z denotes Clarke's partial subgradient operator with respect to z [5].*

Then $\phi(., t)$ admits a stable local error bound at (z^0, t^0).

Proof. Let $h = (h_1, \ldots, h_q)$ and define

$$\Phi(z, t) = \big(h(z, t), g(z, t), Az, t\big)^\top,$$

where A is a matrix whose rows complement the gradients $\nabla_z h_i(z^0, t^0)$ to a basis of the orthogonal complement of v. Let $(\zeta^0, t^0) = \Phi(z^0, t^0)$. We shall use Clarke's inverse function theorem [5] to show that Φ is a Lipschitzian homeomorphism at (z^0, t^0). Let I_m denote the $m \times m$ identity matrix and consider the $(n + m) \times (n + m)$ matrix

$$M = \begin{bmatrix} \nabla_z h(z^0, t^0) & \nabla_t h(z^0, t^0) \\ -v & 0 \\ A & 0 \\ 0 & \mathrm{I}_m \end{bmatrix}.$$

By our choice of A the matrix M is nonsingular. Now suppose the row $(-v, 0)$ is replaced by an element (w, u) of $\partial g(z^0, t^0)$, where $w \in \partial_z g(z^0, t^0)$. If the new matrix is singular, then w is contained in the span of the vectors $\nabla_z h_i(z^0, t^0)$ and the rows of A and is thus orthogonal to v which, however, contradicts the assumption that $w^\top v < 0$. Hence the matrix M remains nonsingular if the row $(-v, 0)$ is replaced by a gradient $(w, u) \in \partial g(z^0, t^0)$. Thus the generalized Jacobian $\partial\Phi(z^0, t^0)$ is regular and Clarke's inverse function theorem [5] shows that Φ is a local Lipschitzian homeomorphism mapping an open neighborhood S of (z^0, t^0) onto an open neighborhood T of (ζ^0, t^0).

Let $B(x, \epsilon)$ denote the ball around x of radius ϵ and let $\epsilon \geq \delta > 0$ be such that

$$\begin{aligned} &B(\zeta^0, \epsilon) \times B(t^0, \epsilon) \subseteq T, \\ &B(z^0, \delta) \times B(t^0, \delta) \subseteq \Phi^{-1}(B(\zeta^0, \epsilon) \times B(t^0, \epsilon)). \end{aligned}$$

Let $\Psi(\zeta, t)$ be the vector of the first n components of $\Phi^{-1}(\zeta, t)$, i.e., $z = \Psi(\zeta, t)$ is the unique solution of the equation

$$(h(z, t), g(z, t), Az) = \zeta$$

for $(\zeta, t) \in T$. If L is a Lipschitz constant of Ψ on $B(\zeta^0, \epsilon) \times B(t^0, \epsilon)$, then

$$\|\Psi(\zeta, t) - \Psi(\tilde{\zeta}, t)\| \leq L\|\zeta - \tilde{\zeta}\|$$

for every $\zeta, \tilde{\zeta} \in B(\zeta^0, \epsilon)$ and every $t \in B(t^0, \epsilon)$, i.e., for every $t \in B(t^0, \epsilon)$ the mapping $\Psi_t = \Psi(., t)$ is Lipschitzian of rank L in $B(\zeta^0, \epsilon)$. Moreover, the mappings Ψ_t map the balls $B(\zeta^0, \epsilon)$ homeomorphically onto $\Psi_t(B(\zeta^0, \epsilon))$. Note that $B(z^0, \delta) \subseteq \Psi_t(B(\zeta^0, \epsilon))$ for every $t \in B(t^0, \delta)$.

Reducing ϵ if necessary, we may assume that ϕ is a continuous selection of the functions

$$s_i(z, t) = \sum_{j=1}^{q} \alpha_{ij} h_j(z, t) + \beta_i g(z, t) + \rho_i$$

on $\Phi^{-1}(B(\zeta^0,\epsilon)\times B(t^0,\epsilon))$. By the definition of Ψ_t we obtain

$$s_i(\Psi_t(\zeta),t)=\sum_{j=1}^{q}\alpha_{ij}\zeta_j+\beta_i\zeta_{q+1}+\rho_i$$

for every $t\in B(t^0,\epsilon)$ and $\zeta\in B(\zeta^0,\epsilon)$. Hence the function

$$\tilde{\phi}(\zeta,t)=\phi(\Psi_t(\zeta),t)$$

is a continuous selection of the latter affine functions in the ball $B(\zeta^0,\epsilon)$ and, moreover, is independent of t for $t\in B(t^0,\epsilon)$. Further reducing ϵ if necessary, we may thus assume, in view of Lemma 3.1, that there exists a positive constant γ such that

$$d_{\tilde{\phi}_t^{-1}(0)}(\zeta)\le\gamma\tilde{\phi}(\zeta,t)$$

for every $\zeta\in B(\zeta^0,\epsilon)$, where $\tilde{\phi}_t=\tilde{\phi}(.,t)$. Note that

$$\phi_t(z):=\phi(z,t)=\tilde{\phi}(\Psi_t^{-1}(z),t)=\tilde{\phi}_t(\Psi_t^{-1}(z)).$$

Recall that Ψ_t is Lipschitzian of rank L on $B(\zeta^0,\epsilon)\supseteq\Psi_t^{-1}(B(z^0,\delta))$ and that the latter set is mapped homeomorphically onto $B(z^0,\delta)$ by Ψ_t. Hence Lemma 3.2 shows that

$$d_{\phi_t^{-1}(0)}(z)\le\gamma L\phi(z,t)$$

for every $z\in B(z^0,\delta)$ and every $t\in B(t^0,\delta)$, which proves the assertion. □

The relation of the assumptions of Theorem 3.3 to the classical Mangasarian–Fromovitz constraint qualification is obvious in view of part 3 of Proposition 2.1.

COROLLARY 3.4. *If in the setting of Theorem* 3.3 *the function g is piecewise smooth in a neighborhood of (z^0,t^0) with selection functions $g_1,\dots,g_p$, then $\phi(.,t)$ admits a stable local error bound at a root (z^0,t^0), provided the following conditions hold:*

1. *The gradients $\nabla_z h_1(z^0,t^0),\dots,\nabla_z h_q(z^0,t^0)$ are linearly independent.*
2. *There exists a vector $v\in\mathbb{R}^n$ with $\nabla_z h_1(z^0,t^0)v=\cdots=\nabla_z h_q(z^0,t^0)v=0$ and $\nabla_z g_j(z^0,t^0)^\top v<0$ for every $j\in\{1,\dots,p\}$ with $g_j(z^0,t^0)=g(z^0,t^0)$.*

The foregoing result is a generalization of the aforementioned result of Robinson [35] which applies to sets of the form

$$C(t)=\{z\mid h_i(z,t)=0,\quad i=1,\dots,q,\quad g_j(z,t)\le 0,\quad j=1,\dots,p\}.$$

Indeed, for $g(z,t)=\max\{g_1(z,t),\dots,g_p(z,t)\}$ the set $C(t)$ is the set of roots of the nonnegative piecewise smooth function

$$\phi(z,t)=\max\{g(z,t),|h_1(z,t)|,\dots,|h_q(z,t)|\}$$

and the assumptions of Corollary 3.4 hold at (z^0,t^0) if and only if the Mangasarian–Fromovitz constraint qualification is satisfied at $z^0\in C(t^0)$. Note that g is not needed as a local selection function if $g(z^0,t^0)<0$. A linear independence–type result can be readily obtained from Corollary 3.4 in view of the following well-known fact.

LEMMA 3.5. *If the gradients $\nabla_z g_i(z^0,t^0)$, $i=1,\dots,p$, and $\nabla_z h_j(z^0,t^0)$, $j=1,\dots,q$, are linearly independent, then assumptions* 1 *and* 2 *of Corollary* 3.4 *hold.*

Nonsmooth formulations of equilibrium constraints often involve composite equations of the form

$$G \circ H(z) = 0, \tag{3.3}$$

where the function G is nonsmooth and is thought of as a structure function determining a problem class, while H is smooth and represents the data of a problem instance [33]. A typical example is the equation

$$\min\{x, F(x,y)\} = 0,$$

which reformulates a parametric nonlinear complementarity problem. If we apply Corollary 3.4 to constraints involving composite equations, we obtain the following corollary.

COROLLARY 3.6. *Consider the set*

$$\Omega(t) = \{z \in \mathbb{R}^n \ \mid \ G(H(z,t)) = 0, \quad g(z,t) \leq 0, \quad h(z,t) = 0\}$$

with $t \in \mathbb{R}^m$*,* $G(H(z,t)) \in \mathbb{R}^r$*,* $g(z,t) \in \mathbb{R}$*,* $h(z,t) \in \mathbb{R}^q$ *and suppose* H *and* h *are smooth,* g *is piecewise smooth, and* G *is piecewise affine. Let* $z^0 \in \Omega(t^0)$ *and let* $g_1, \ldots, g_l$ *be a collection of smooth selection functions for* g *at* (z^0, t^0)*. If*

1. *the matrix* $[\nabla_z H(z^0,t^0)^\top, \nabla_z h(z^0,t^0)^\top]$ *has full column rank,*
2. *there exists a vector* $v \in \mathbb{R}^n$ *such that*

$$\nabla_z H(z^0,t^0)v = 0, \quad \nabla_z h(z^0,t^0)v = 0, \quad \nabla_z g_i(z^0,t^0)v < 0 \quad \forall i : g_i(z^0,t^0) = 0,$$

then there exist neighborhoods U *of* z^0 *and* V *of* t^0 *and a constant* γ *such that*

$$d_{\Omega(t)}(z) \leq \gamma \phi(z,t) \quad \forall (z,t) \in U \times V,$$

where

$$\phi(z,t) = \max\{\|G(H(z,t))\|_\infty, \|h(z,t)\|_\infty, g(z,t)\}.$$

Proof. Locally around (z^0,t^0) the function ϕ has the selection functions g_i, $i : g_i(z^0,t^0) = 0$, $\pm h_i$, and

$$\pm \sum_{j=1}^{p} \alpha_{ij} H_j(z,t) + \rho_i,$$

where $\alpha_{ij} v_j + \rho_i$ is a selection function of the piecewise affine function G_i. Hence, replacing the function h in the statement of Corollary 3.4 by the function (H,h), we obtain the result. □

3.2. Local error bounds for MPECs.

3.2.1. Stationarity constraints. We shall next apply the foregoing results to constraint sets which involve equilibrium constraints in the form of stationarity conditions for variational inequalities. In order not to overload the exposition with too many technicalities, we confine ourselves to variational inequalities over inequality constrained sets and assume that all further constraints of the MPEC are expressed as a single piecewise smooth inequality. More precisely, we assume that the constraint set of the MPEC is of the form

$$\Omega(t) = \{(x,y,\lambda) \mid \begin{array}{l} F(x,y,t) + \nabla_x\Psi(x,y,t)^\top\lambda = 0, \\ \min\{-\Psi(x,y,t),\lambda\} = 0, \\ g(x,y,t) \le 0\} \end{array} \tag{3.4}$$

for some fixed perturbation parameter t, where F is a C^1-function, Ψ is a C^2-function, and g is a real-valued PC^1-function. The set $\Omega(t)$ is the set of roots of the function

$$\begin{aligned}\phi(x,y,\lambda,t) = \quad & \max\{\|F(x,y,t) + \nabla_x\Psi(x,y,t)^\top\lambda\|_\infty, \\ & \qquad \|\min\{-\Psi(x,y,t),\lambda\}\|_\infty, g(x,y,t)\}.\end{aligned} \tag{3.5}$$

Note that ϕ coincides in a neighborhood of a root (x^0,y^0,λ^0,t^0) with the function

$$\begin{aligned}\tilde\phi(x,y,\lambda,t) = \max\{&\|F(x,y,t) + \nabla_x\Psi(x,y,t)^\top\lambda\|_\infty, \\ &\|\Psi_I(x,y,t)\|_\infty, \|\lambda_K\|_\infty, \\ &\|\min\{-\Psi_J(x,y,t),\lambda_J\}\|_\infty, g(x,y,t)\},\end{aligned}$$

where

$$\begin{array}{rcl} I &=& \{i \mid \Psi_i(x^0,y^0,t^0) = 0 < \lambda_i^0\}, \\ J &=& \{j \mid \Psi_j(x^0,y^0,t^0) = 0 = \lambda_j^0\}, \\ K &=& \{k \mid \Psi_k(x^0,y^0,t^0) < 0 = \lambda_k^0\}. \end{array} \tag{3.6}$$

Given a vector z and an index set I we have used the notation z_I to denote the vector with components z_i, $i \in I$. In view of the above observation, Corollary 3.6 yields the following stable local error bound conditions.

Corollary 3.7. *Let (x^0,y^0,λ^0,t^0) be a root of the function ϕ defined by* (3.5)*, let the index sets I and J be defined by* (3.6)*, and let*

$$M_{x\lambda} = \begin{pmatrix} \nabla_x F(x^0,y^0,t^0) + \sum_{i\in I}\lambda_i^0\nabla_{xx}^2\Psi_i(x^0,y^0,t^0) & \nabla_x\Psi_I(x^0,y^0,t^0)^\top \\ -\nabla_x\Psi_{I\cup J}(x^0,y^0,t^0) & 0 \end{pmatrix},$$

$$M_y = \begin{pmatrix} \nabla_y F(x^0,y^0,t^0) + \sum_{i\in I}\lambda_i^0\nabla_{xy}^2\Psi_i(x^0,y^0,t^0) \\ -\nabla_y\Psi_{I\cup J}(x^0,y^0,t^0) \end{pmatrix}.$$

Then $\phi(.,t)$ admits a stable local error bound at (x^0,y^0,λ^0,t^0) if the following conditions hold:

1. *The matrix $M = (M_{x\lambda}, M_y)$ has full row rank.*
2. *If $g(x^0,y^0,t^0) = 0$, then there exists a collection of C^1 selection functions $g_1,\dots,g_p$ of g at (x^0,y^0,t^0) and vectors u,v,w such that $M_{x\lambda}(u,v)^\top + M_y w = 0$ and*

   $$\nabla_x g_k(x^0,y^0,t^0)u + \nabla_y g_k(x^0,y^0,t^0)w < 0$$

   for every $k = 1,\dots,p$.

Lemma 3.5 shows that the assumptions of the corollary are particularly satisfied if the gradients of the active constraint functions are linearly independent, i.e., if the matrix

$$\left(\begin{array}{cc} M_{x\lambda} & M_y \\ (\nabla_x g_1(x^0,y^0,t^0),0) & \nabla_y g_1 L(x^0,y^0,t^0) \\ \vdots & \vdots \\ (\nabla_x g_p(x^0,y^0,t^0),0) & \nabla_y g_p L(x^0,y^0,t^0) \end{array}\right)$$

has full row rank. The first condition of the corollary can be viewed as a regularity condition for the variational inequality at the solution point x^0 for the given parameter vector y^0. Recall that the global error bound conditions of [20], which are concerned with the problem for fixed parameter t^0, imply that the stationary solution x of the variational inequality is a Lipschitz continuous function of y. If t^0 is fixed and the linear independence constraint qualification and strict complementarity, i.e., $J = \emptyset$, hold for the lower level problem, then the Lipschitz continuity of x as a function of y is equivalent to the nonsingularity of the matrix

$$\begin{pmatrix} \nabla_x F(x^0,y^0,t^0) + \sum\limits_{i\in I} \lambda_i^0 \nabla_{xx}^2 \Psi_i(x^0,y^0,t^0) & \nabla \Psi_I(x^0,y^0,t^0)^\top \\ -\nabla \Psi_I(x^0,y^0,t^0) & 0 \end{pmatrix},$$

which implies that the first condition of the above corollary holds. However, the latter condition is considerably weaker. In particular it does not imply Lipschitzian behavior of the stationary point x as a function of y. In fact, if g is of the type

$$g(x,y) = \max\{g_1(x,y),\ldots,g_p(x,y)\}$$

with smooth functions g_i and strict complementarity holds at (x^0,y^0,λ^0,t^0), i.e., $J = \emptyset$, then the condition of the corollary is equivalent to the Mangasarian–Fromovitz condition for the set

$$\begin{array}{lcl} \{(x,y,\lambda) & | & F(x,y,t) + \nabla_x \Psi(x,y,t)^\top \lambda = 0, \\ & & \Psi_I(x,y,t) = 0, \\ & & \lambda_K = 0, \\ & & g(x,y,t) \leq 0\}, \end{array}$$

which coincides with $\Omega(t)$ in a neighborhood of (x^0,y^0,λ^0,t^0) and the assertion follows from Robinson's result [35] in this special case.

It follows from the results of [40] that the conditions of Corollary 3.7 serve as a constraint qualification for MPECs in the sense that they guarantee that a local minimizer is a stationary point. Alternative constraint qualifications based on piecewise smooth formulations of MPECs have been suggested [18, 20]. In the first reference condition 1 of Corollary 3.7 is relaxed to the requirement that either M has full row rank or every square submatrix of M of maximal dimension has full row rank. This relaxation allows for more active constraints than variables and is in the spirit of a constant rank assumption. In [20, 39], a so-called piecewise Mangasarian–Fromovitz constraint qualification was suggested. Here, we confine ourselves to the setting of [20] which assumes that

$$g(x,y) = \max\{g_1(x,y),\ldots,g_p(x,y)\}$$

for smooth functions g_i. The approach is based on the observation that the constraint set (3.4) is in a neighborhood of $(x^0, y^0, \lambda^0, t^0)$ locally representable as the finite union of the constraint sets

$$\begin{aligned}\Omega_\alpha(t) = \{(x, y, \lambda) \quad | \quad & F(x, y, t) + \nabla_x \Psi(x, y, t)^\top \lambda = 0,\\ & \Psi_{I\cup\alpha}(x, y, t) = 0,\\ & \lambda_{I\cup\alpha} \geq 0,\\ & \Psi_{K\cup\bar{\alpha}}(x, y, t) \leq 0,\\ & \lambda_{K\cup\bar{\alpha}}(x, y, t) = 0,\\ & g_1(x, y, t) \leq 0,\\ & \vdots\\ & g_p(x, y, t) \leq 0\},\end{aligned}$$

where I, J, K are as defined in (3.6), $\alpha \subseteq J$, and $\bar{\alpha}$ is the complement of α in J. The piecewise Mangasarian–Fromovitz constraint qualification requires that all sets $\Omega_\alpha(t^0)$ with $\alpha \subseteq J$ satisfy the smooth Mangasarian–Fromovitz constraint qualification at (x^0, y^0, λ^0). This condition is weaker than the condition of Corollary 3.7 which is equivalent to the requirement that the set

$$\begin{aligned}\Omega^*(t) = \{(x, y, \lambda) \quad | \quad & F(x, y, t) + \nabla_x \Psi(x, y, t)^\top \lambda = 0,\\ & \Psi_{I\cup J}(x, y, t) = 0,\\ & \lambda_{K\cup J} = 0,\\ & g_1(x, y, t) \leq 0,\\ & \vdots\\ & g_p(x, y, t) \leq 0\}\end{aligned}$$

satisfies the smooth Mangasarian–Fromovitz constraint qualification. The weaker piecewise Mangasarian–Fromovitz constraint qualification still gives rise to local error bounds by means of the maximum of the local error bounds for the pieces $\Omega_\alpha(t)$, since the distance to $\Omega(t)$ is locally the minimum of the distances to $\Omega_\alpha(t)$ and ϕ coincides locally with the minimum of the functions ϕ_α corresponding to the sets $\Omega_\alpha(t)$.

3.2.2. Complementarity constraints. If $\Psi(x, y, t) = -x$, then the multiplier λ of the stationarity condition for the variational inequality can be eliminated and the stationarity condition turns into a nonlinear complementarity problem. The constraint set of the MPEC is then of the form

$$\Omega(t) = \{(x, y) \quad | \quad \min\{x, F(x, y, t)\} = 0, g(x, y, t) \leq 0\} \tag{3.7}$$

for some parameter t. This feasible set is the set of roots of the nonnegative function

$$\phi(x, y, t) = \max\{\|\min\{x, F(x, y, t)\}\|_\infty, g(x, y, t)\} \tag{3.8}$$

and we obtain the following conditions for a stable local error bound of $\phi(., t)$ at a root (x^0, y^0, t^0).

COROLLARY 3.8. *Let (x^0, y^0, t^0) be a root of the function ϕ defined by* (3.8) *and let*

$$I = \{i \mid F_i(x^0, y^0, t^0) = 0\}, \quad J = \{i \mid x_i^0 > 0\}.$$

Then $\phi(., t)$ admits a stable local error bound at (x^0, y^0, t^0) if the following conditions hold:

1. *The matrix*

$$\big(\nabla_{x_J} F_I(x^0, y^0, t^0), \quad \nabla_y F_I(x^0, y^0, t^0)\big)$$

has full row rank.

2. *If* $g(x^0, y^0, t^0) = 0$*, then there exists a collection of* C^1 *selection functions* $g_1, \dots, g_p$ *of* g *at* (x^0, y^0, t^0) *and vectors* u, v *such that*

$$\nabla_{x_J} F_I(x^0, y^0, t^0)u + \nabla_y F_I(x^0, y^0, t^0)v = 0$$

and

$$\nabla_{x_J} g_k(x^0, y^0, t^0)u + \nabla_y g_k(x^0, y^0, t^0)v < 0$$

for every $k = 1, \dots, p$.

3.2.3. Normal equation constraints. Next we derive error bound conditions for constraints involving normal equations. We assume that P is a closed convex set and $F(.,.)$ is a parametric vector field. Recall that ξ satisfies the variational inequality induced by $F(., y)$ over P if and only if $\xi = \Pi_P(x)$ and x solves Robinson's normal equation [38]

$$H(x, y) := \Pi_P(x) - F(\Pi_P(x), y) - x = 0. \tag{3.9}$$

Recall that H is locally Lipschitzian, provided F is locally Lipschitzian. We can thus apply the following result, which is a direct consequence of Lemmas 3.1 and 3.2.

LEMMA 3.9. *Let*

$$\Omega(t) = \{z \in \mathbb{R}^n \mid g(z, t) \le 0, h(z, t) = 0\},$$

where $g: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ *and* $h: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$ *are locally Lipschitzian, let* $z^0 \in \Omega(t^0)$*, let* $f: O \to \mathbb{R}^{n-p-1}$ *be a locally Lipschitzian function defined in a neighborhood* O *of* (z^0, t^0)*, and let* $\Psi: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ *be defined by*

$$\Psi(z, t) = \big(g(z, t), h(z, t), f(z, t)\big).$$

If the equation $\Psi(z, t) = u$ *has a locally unique and Lipschitzian solution* $z(u, t)$ *in a neighborhood of* $(z^0, t^0, \Psi(z^0, t^0))$*, then there exists a constant* γ *and neighborhoods* U *of* z^0 *and* V *of* t^0 *such that*

$$d_{\Omega(t)}(z) \le \gamma \phi(z, t) \quad \forall (z, t) \in U \times V,$$

where

$$\phi(z, t) = \max\{\|h(z, t)\|_\infty, g(z, t)\}.$$

Proof. Let $\Psi_t^{-1}(u)$ be the locally unique solution z of $\Psi(z, t) = u$. If $u = (\alpha, v, w)$ with $\alpha \in \mathbb{R}$, $v \in \mathbb{R}^p$, and $w \in \mathbb{R}^{n-p-1}$, then $\phi \circ \Psi_t^{-1}(u) = \max\{\|v\|_\infty, \alpha\}$ which is piecewise linear. Hence the Lipschitz continuity of the solution function $z(.,.)$ implies the result in view of the Lemmas 3.1 and 3.2. □

The foregoing result can be used to derive stable local error bound conditions for normal maps if Π_P is piecewise differentiable and F is differentiable. Sufficient conditions for the piecewise smoothness of Π_P have been derived in [17, 28]. In fact, if $P = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0\}$ with convex C^{r+1}-functions g_i and Janin's constant rank condition [12] holds at $\Pi_P(x) \in P$, then Π_P is a PC^r function in a neighborhood of x.

PROPOSITION 3.10. *Let*

$$\begin{aligned}\Omega(t) = \{(x,y) \in \mathbb{R}^p \times \mathbb{R}^q \quad | \quad & \Pi_P(x) - F(\Pi_P(x), y, t) - x = 0,\\ & h(\Pi_P(x), y, t) = 0,\\ & g(\Pi_P(x), y, t) \leq 0\},\end{aligned}$$

where $F: \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^m \to \mathbb{R}^p$, $h: \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^m \to \mathbb{R}^s$ *and* $g: \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^m \to \mathbb{R}^r$ *are* C^1 *functions, and* $P \subseteq \mathbb{R}^p$ *is a closed convex set. Let* $(x^0, y^0) \in \Omega(t^0)$ *and let* $J = \{j \mid g_j(\Pi_P(x^0), y^0, t^0) = 0\}$. *Suppose* Π_P *is* PC^1 *with selection functions* $P_1, \ldots, P_l$ *in a neighborhood of* x^0. *If the matrices*

$$M_i = \left[\begin{array}{cc} \nabla P_i(x^0) - \nabla_x F(\Pi_P(x^0), y^0, t^0)\nabla P_i(x^0) - \mathrm{I} & -\nabla_y F(\Pi_P(x^0), y^0, t^0) \\ \nabla_x h(\Pi_P(x^0), y^0, t^0)\nabla P_i(x^0) & \nabla_y h(\Pi_P(x^0), y^0, t^0) \\ \nabla_x g_J(\Pi_P(x^0), y^0, t^0)\nabla P_i(x^0) & \nabla_y g_J(\Pi_P(x^0), y^0, t^0) \end{array}\right]$$

with $i \in \{1, \ldots, l\}$ *are coherently oriented* (*cf.* (2.2)), *then there exist neighborhoods* U *of* (x^0, y^0) *and* V *of* t^0 *and a constant* γ *such that*

$$d_{\Omega(t)}(x,y) \leq \gamma\phi(x,y,t) \quad \forall (x,y,t) \in U \times V,$$

where

$$\begin{aligned}\phi(x,y,t) = \max\{&\|\Pi_P(x) - F(\Pi_P(x), y, t) - x\|_\infty, \|h(\Pi_P(x), y, t)\|_\infty,\\ & g_1(\Pi_P(x), y, t), \ldots, g_r(\Pi_P(x), y, t)\}.\end{aligned}$$

Proof. Since $\phi(x,y,t)$ locally coincides with the function

$$\tilde{\phi}(x,y,t) = \max_{j \in J}\{\|\Pi_P(x) - F(\Pi_P(x), y, t) - x\|_\infty, \|h(\Pi_P(x), y, t)\|_\infty, g_j(\Pi_P(x), y, t)\},$$

we may neglect nonactive inequalities and assume for simplicity that $J = \{1, \ldots, r\}$. Recall that coherent orientation requires the existence of a matrix $A = (A_x, A_y)$ such that all matrices

$$\begin{pmatrix} M_i \\ A \end{pmatrix}$$

have the same nonvanishing determinantal sign. The foregoing matrices are in fact the partial Jacobians with respect to (x,y) of the selection functions of the mapping

$$\Psi(x,y,t) = \left[\begin{array}{c} \Pi_P(x) - F(\Pi_P(x), y, t) - x \\ h(\Pi_P(x), y, t) \\ g(\Pi_P(x), y, t) \\ A_x x + A_y y \end{array}\right]$$

at (x^0, y^0, t^0). To apply Theorem 2.2, it thus remains to show that the first-order approximation of the equation $\Psi(x,y,t) = s$ defines a unique solution function. The B-derivative of Ψ is given by

$$\Psi'((x^0, y^0, t^0); (u,v,w)) = \left[\begin{array}{l} \Pi_P'(x^0; u) - \nabla_x F(\Pi_P(x^0), y^0, t^0)\Pi_P'(x^0; u) - \\ \qquad \nabla_y F(\Pi_P(x^0), y^0, t^0)v - \nabla_t F(\Pi_P(x^0), y^0, t^0)w - u \\ \nabla h(\Pi_P(x^0), y^0, t^0)(\Pi_P'(x^0; u), v, w)^\top \\ \nabla g(\Pi_P(x^0), y^0, t^0)(\Pi_P'(x^0; u), v, w)^\top \\ A_x u + A_y v \end{array}\right].$$

Recall that the B-derivative of a piecewise differentiable function is piecewise linear. In fact, the B-derivative $\Pi'_P(x^0;.)$ is the projection onto a polyhedral cone (cf., e.g., [28]). Hence the branching number of the underlying subdivision is 4 (cf. [32, 42]) and thus the fact that the block matrices with rows M_i and A have the same nonvanishing determinantal sign ensures that the equation $\Psi'((x^0, y^0, t^0); (u, v, w)) = r$ has a unique solution $u(w, r), v(w, r)$ [16]. Hence we can apply Theorem 2.2 to deduce that the equation $\Psi(x, y, t) = s$ has a locally unique solution $x(t, s), y(t, s)$. An application of Lemma 3.9 thus proves the assertion. □

To check the assumption of the foregoing proposition one has to calculate a collection of selection functions for Π_P at the point x^0. This has been done in [17, 28] for sets of the form $P = \{z \in \mathbb{R}^n \mid q_i(z) \leq 0, i = 1, \ldots, m\}$ with convex C^2-functions q_i under certain constraint qualifications. For the sake of completeness we repeat the arguments for the most important case that the linear independence constraint qualification holds at a point $p^0 = \Pi_P(x^0)$. The linear independence assumption implies that for every index set $I \subseteq I^0 = \{i \mid q_i(x^0) = 0\}$ there exists a unique solution $p_I(x), \lambda_I(x)$ of the stationary point equation

$$\begin{aligned} p + \nabla q_I(p)^\top \lambda_I &= x, \\ q_I(p) &= 0 \end{aligned}$$

in a neighborhood of $(p^0 + \nabla q_I(p^0)^\top \lambda_I^0, p^0, \lambda_I^0)$ and furthermore that the point $p_I(x^0)$ coincides with the point p^0 if and only if $I^+ \subseteq I \subseteq I^0$, where $I^+ = \{i \mid \lambda_i^0 > 0\}$. Hence, the functions $p_I(.)$, $I^+ \subseteq I \subseteq I^0$ constitute a collection of selection functions for Π_P in a neighborhood of x^0. The inverse function theorem yields the gradients of the functions p_I at x^0.

In [19, 20] it has been shown that a parametric normal equation $H(x, y) = t$ admits a local error bound at a solution, provided the solution x of the normal equation is locally a Lipschitz continuous function of the parameters y and t. This result is encompassed by the above proposition if one chooses the matrix $A = (A_x, A_y) = (0, \mathrm{I})$ to augment the matrices in the coherency condition. Our result, however, is in the spirit of a full rank assumption rather than an implicit function assumption. In fact, if specialized to the case $P = \mathbb{R}^n$, in which case Π_P is the identity mapping and the normal equation turns into the equation $-F(x, y) = 0$, then for smooth F the condition in [19] requires that $\nabla_x F(x^0, y^0, t^0)$ is nonsingular, while our condition requires that the matrix

$$\big(\nabla_x F(x^0, y^0, t^0), \nabla_y F(x^0, y^0, t^0)\big)$$

has full row rank.

The proof of Proposition 3.10 can mutatis mutandis be adopted to prove an analogous result for constraint sets involving the direct normal equation

$$\Pi_P(x - F(x, y)) - x = 0.$$

We leave the details to the reader.

The results of this section hold not only for the specified functions ϕ but in fact for any other nonnegative piecewise affine composition of the constraint functions of the MPEC; e.g., the function

$$\phi(x, y, t) = \sum |\min\{x_i, F_i(x, y, t)\}| + \sum \max\{g_j(x, y, t), 0\}$$

admits a local error bound at a point (x^0, y^0, t^0) if the conditions of Corollary 3.8 hold. We shall make use of a penalty function of the latter type in section 5.

The stable error bound results of this section can be refined and extended to infinite dimensional perturbation parameters by using Theorem 11 and Corollary 21 of [33] instead of Clarke's inverse function theorem [5], Theorem 2.2, or Corollary 2.3. The former results allow for general perturbations in the C^1 topology rather than merely for parametric perturbations.

In view of the principle of exact penalization, the foregoing local error bound results yield the following concluding exact penalty result for MPECs.

THEOREM 3.11. *Suppose the assumptions of Corollary* 3.7*, Corollary* 3.8*, or Proposition* 3.10 *hold and let* $f(x, y)$ *be a locally Lipschitz continuous function. Then there exists a neighborhood* U *of* (x^0, y^0, t^0) *and a positive scalar* γ^0 *such that for all* $(x, y, t) \in U$ *with* $(x, y) \in \Omega(t)$ *the following two statements are equivalent:*

1. (x, y) *is a local minimizer of the* MPEC $\min\{f(\xi, \eta) \mid (\xi, \eta) \in \Omega(t)\}$,
2. (x, y) *is a local minimizer of the function* $f(.,.) + \gamma\phi(.,.,t)$ *for every* $\gamma \geq \gamma^0$.

4. A trust region method.

The MPEC penalty functions that we encountered in the last section are of the type

$$f(x) = g(x) + p(h(x)),$$

where g is the C^1 objective function, p is piecewise affine, and h is a vector valued C^1-function. Such functions are locally Lipschitz and B-differentiable and one can thus associate two different stationarity concepts with them. We call a point x a *Bouligand stationary* (B-stationary) point of f if $f'(x; .) \geq 0$, and we call it *Clarke stationary* (C-stationary) if $0 \in \partial f(x)$, where ∂f denotes Clarke's subdifferential [5]. Since the B-derivative is always majorized by Clarke's derivative, a B-stationary point is C-stationary as well. The reverse statement holds generally only if f is regular at x, i.e., its B-derivative at x coincides with Clarke's derivative at this point.

A suitable tool for the minimization of the penalty functions f is the bundle-trust-region method of Schramm and Zowe [43]. However, their method is designed to find a C-stationary point and, in fact, their convergence theory states that under suitable assumptions at least one of the accumulation points of the computed sequence will be C-stationary. To apply the method it is necessary to be able to compute a Clarke subgradient of the penalty function at a given point x. Since the B-derivative is majorized by the Clarke derivative, the subdifferential of the first-order approximation of f is a subset of the subdifferential of f. It thus suffices to calculate a subgradient of the piecewise linear B-derivative

$$f'(x; y) = \nabla g(x)y + p'(h(x); \nabla h(x)y)$$

at $y = 0$. Recall that the Clarke subdifferential of a piecewise differentiable function is the convex hull of the collection of all gradients of its essentially active selection functions. So it suffices to determine an essentially active selection function of the piecewise linear function $f'(x; .)$. However, this can be a nontrivial combinatorial problem and the best way to do it depends heavily on the function p. We will not go into the details of the bundle-trust-region approach here, but instead propose a different trust region method which is applicable to the nonsmooth MPEC penalty functions and is designed to find a B-stationary point.

4.1. The trust region framework.

In this section, we present a general trust region framework for the minimization of locally Lipschitz continuous B-differentiable functions $f : \mathbb{R}^n \to \mathbb{R}$.

Given a point x^k, a trust region method employs a local model $m_k(.)$ of the function $f(x^k + .)$ which is assumed to be sufficiently accurate within the so-called trust radius δ_k about x^k. In order to determine a better point x^{k+1}, one determines an approximation s^k of the optimal solution of the problem

$$\min_{\|s\| \le \delta_k} m_k(s). \tag{4.1}$$

Having determined s^k, one checks whether the model $m_k(.)$ of the function $f(x^k + .)$ is indeed sufficiently accurate at the point s^k by comparing the actual reduction $f(x^k) - f(x^k + s^k)$ of the function value with the predicted reduction $m_k(0) - m_k(s^k)$ given by the model $m_k(.)$. If, on the one hand, the actual reduction does not exceed a fixed fraction of the predicted reduction, then the model is considered to be inadequate in the ball around x^k with radius δ_k; the method will reduce the trust radius and possibly improve the model, but it will not move to the point $x^k + s^k$. If, on the other hand, the actual reduction exceeds a certain fixed fraction of the predicted reduction, then x^k is updated to $x^{k+1} = x^k + s^k$ and a new model m_{k+1} along with a suitable update $\delta_{k+1} \ge \delta_k$ of the trust radius is determined. We shall use the following setup for the trust region method.

Trust Region Method. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function and let $0 < c_0 < c_1 < 1$ be fixed parameters.

```
kth iteration: Let $x^k \in \mathbb{R}^n$, $\delta_k > 0$ and $m_k : \mathbb{R}^n \to \mathbb{R}$ be given.

    Determine a vector $s^k$ with $\|s^k\| \le \delta_k$ and $m_k(s^k) < m_k(0)$. If no such
    vector exists, then stop.

    Set $r_k := \frac{f(x^k) - f(x^k + s^k)}{m_k(0) - m_k(s^k)}$.

    If $r_k \le c_0$, then set $x^{k+1} := x^k, \delta_{k+1} := Reduce(\delta_k)$,
    else if $c_0 < r^k \le c_1$, then set $x^{k+1} := x^k + s^k, \delta_{k+1} := \delta_k$,
        else set $x^{k+1} := x^k + s^k$, $\delta_{k+1} := Increase(\delta_k)$.

    Determine $m_{k+1}$.
```

A sequence $\{x^k\}_{k \in \mathbb{N}}$ is called a trust region sequence if it can be generated by the above procedure from some initial vector x^0. The above method is clearly a descent method, i.e., $f(x^{k+1}) \le f(x^k)$ for every $k \in \mathbb{N}$. This framework can thus be employed for the minimization of an arbitrary function f. However, to obtain convergence results, one has to impose a number of assumptions. Our first assumption concerns the function class for which we analyse the method.

(A1) *The function $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous and B-differentiable.*

In order to convert the trust region framework into an algorithm, one has to specify the model function m_k, a method for the determination of s_k, as well as the trust radius adjustments *Reduce*(.) and *Increase*(.). We will not give specific recipes for these selections but instead impose certain restrictions on such recipes. The assumptions that we use here are not the most general convergence assumptions, but they are general enough to cover the case of MPEC penalty functions and at the same time allow a fairly straightforward convergence proof. They are not implied by the assumptions made in [7, 31]. A convergence proof for the method under slightly more general assumptions as well as a comparison of the assumptions with the convergence conditions used in [7, 27, 31] is given in [44].

We first assume that s^k is chosen in such a way that a certain fixed fraction of the optimal model reduction can be guaranteed in each step.

(A2) *There is a number* $\tau \in (0,1]$ *(independent of* k*) such that*

$$m_k(0) - m_k(s^k) \geq \tau(m_k(0) - m_k^*), \tag{4.2}$$

where $m_k^* = \min_{\|s\| \leq \delta_k} m_k(s)$*.*

Our next assumption concerns the functions *Reduce*$(.)$ and *Increase*$(.)$.

(A3) *For each* $\delta_0 > 0$ *the assignment* $\delta_k := \mathit{Reduce}(\delta_{k-1})$ *generates a strictly decreasing positive nullsequence and there exists* $c_R \in (0,1)$ *such that*

$$\mathit{Reduce}(\delta) \geq c_R \delta.$$

For each $\delta_0 > 0$ *the assignment* $\delta_k := \mathit{Increase}(\delta_{k-1})$ *generates an increasing sequence and there exists* $c_I > 1$ *and* $\delta_{\max} > 0$ *such that*

$$\mathit{Increase}(\delta) \geq c_I \delta$$

for every $\delta \leq \delta_{\max}$*.*

The final set of assumptions concerns the choice of the model function and its approximation properties.

(A4) *The model function* m_k *is of the form*

$$m_k(s) = \Phi(x^k; s) + \frac{1}{2} s^T B_k s, \tag{4.3}$$

where B_k *is an* $n \times n$*-matrix and* $\Phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *is a function which does not depend on* k *and has the following properties:*

(A4.1) *For every* $x \in \mathbb{R}^n$ *the function* $\Phi_x = \Phi(x;.) : \mathbb{R}^n \to \mathbb{R}$ *is locally Lipschitz continuous and B-differentiable.*

(A4.2) $\Phi_x(0) = f(x)$ *and* $\Phi_x'(0;s) = f'(x;s)$*.*

(A4.3) *If* $\bar{x}$ *is an accumulation point of the trust region sequence, then*

1. $\displaystyle\lim_{(x,s)\to(\bar{x},0)} \frac{f(x+s) - \Phi(x;s)}{\|s\|} = 0,$
2. *there exists a positive number* ϵ *such that for every fixed* s *with* $\|s\| < \epsilon$ *the function* $x \to \Phi_s(x) := \Phi(x;s) : \mathbb{R}^n \to \mathbb{R}$ *is upper semicontinuous at* $\bar{x}$*.*

(A4.4) *There exists a number* M *such that* $\|B_k\| \leq M$ *for every* k*.*

assumption (A4.3) is rather restrictive. In general the first-order approximation $f(x) + f'(x;s)$ does not satisfy this assumption. We have seen, however, that a typical MPEC penalty function has the form

$$f(x) = g(x) + p(h(x)), \tag{4.4}$$

where $g : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^m$ are differentiable and $p : \mathbb{R}^m \to \mathbb{R}$ is piecewise linear. A natural approximation of such functions f is of the form

$$\Phi(x;s) = g(x) + \nabla g(x)s + p(h(x) + \nabla h(x)s). \tag{4.5}$$

The following theorem shows that this approximation function satisfies assumption (A4).

THEOREM 4.1. *Let $g:\mathbb{R}^n \to \mathbb{R}$ and $h:\mathbb{R}^n \to \mathbb{R}^m$ be C^1-functions and $p:\mathbb{R}^m \to \mathbb{R}$ be locally Lipschitz continuous and B-differentiable. If f has the form* (4.4) *and Φ is defined by* (4.5)*, then Φ satisfies the conditions of assumption* (A4).

Proof. The function Φ_x is certainly locally Lipschitz continuous and B-differentiable. Moreover, $\Phi_x(0) = f(x)$ and

$$\begin{aligned}\Phi'_x(0; s) &= \nabla g(x)s + p'(h(x); \nabla h(x)s)\\ &= f'(x; s),\end{aligned}$$

in view of the chain rule for the B-derivative [35]. Since the function Φ_s is continuous for every $s \in \mathbb{R}^n$, it remains to be shown that

$$\lim_{(x,s)\to(\bar{x},0)} \frac{f(x+s) - \Phi(x;s)}{\|s\|} = 0. \tag{4.6}$$

Note first of all that

$$\lim_{(x,s)\to(\bar{x},0)} \frac{g(x+s) - g(x) - \nabla g(x)s}{\|s\|} = 0 \tag{4.7}$$

since g is a C^1-function. Moreover, if L is a Lipschitz constant of p, then

$$\frac{\|p(h(x+s)) - p(h(x) + \nabla h(x)s)\|}{\|s\|} \leq L \frac{\|h(x+s) - h(x) - \nabla h(x)s\|}{\|s\|}.$$

Since h is a C^1-function, the right-hand side of the inequality tends to zero, provided x tends to $\bar{x}$ and s tends to zero. This proves (4.6) in view of (4.7). □

Note that the iteration function $\Phi(x; s)$ does not satisfy the assumptions of the trust region methods found in [7, 31] since it is generally neither regular nor subhomogeneous in s.

4.2. Convergence analysis. Having set out the method along with some general assumptions, we shall next study properties of the accumulation points of the trust region sequence. The first rather trivial observation tells us that the trust region method will not stop unless a Bouligand stationary point of the function f is found.

PROPOSITION 4.2. *Suppose assumptions* (A1), (A4.1)*, and* (A4.2) *are satisfied. If the trust region method stops at iteration k and $\delta_k > 0$, then x^k is a Bouligand stationary point of f.*

Proof. If $\delta_k > 0$, then the trust region method stops at iteration k if and only if the origin is a local minimizer and a fortiori a Bouligand stationary point of $m_k(.)$, i.e., $\Phi'_{x^k}(0; .) \geq 0$. Assumption (A4) thus implies that $f'(x^k; .) \geq 0$, i.e., x^k is a Bouligand stationary point of f. □

Next, we show that if x^k is not B-stationary, then the trust region sequence will eventually move to a new point.

PROPOSITION 4.3. *Suppose assumptions* (A1)–(A4) *hold. If x^k is not a B-stationary point of f, then there exists a number $r \in \mathbb{N}$ such that $x^{k+r} \neq x^k$.*

Proof. Since x^k is not a B-stationary point of f, the method cannot terminate at the point x^k in view of Proposition 4.2. It thus either moves to a new point at some iteration or produces a stationary sequence. Let us assume the latter, i.e., $x^k = x^{k+r}$ for every $r \geq 1$. Then the trust radius is reduced at every iteration and thus assumption (A3) implies that δ_k is a nullsequence. Since x^k is not a B-stationary point of f, there exists a direction s with $\|s\| = 1$ and a real number $\alpha > 0$ such that $f'(x^k; s) = -\alpha$. Since we have assumed that the sequence is stationary, the point x^k is an accumulation point of the trust region sequence and hence assumption (A4) implies that

$$m_k'(0; s) = \Phi'_{x^k}(0; s) = f'(x^k; s) = -\alpha.$$

Thus, there exists a number $\bar{\delta} > 0$ such that

$$m_k(\delta s) - m_k(0) \le \frac{-\alpha\delta}{2}$$

for every $\delta \in [0, \bar{\delta}]$. Set $m_k^* = \min\{m_k(s) \mid \|s\| \le \delta_k\}$. If $0 < \delta_k \le \bar{\delta}$, which is eventually the case since δ_k is a positive nullsequence, then

$$m_k(0) - m_k^* \ge m_k(0) - m_k(\delta_k s) \ge \frac{\alpha\delta_k}{2} > 0. \tag{4.8}$$

Now we expand

$$\begin{aligned}\frac{f(x^k) - f(x^k + s^k)}{m_k(0) - m_k(s^k)} &= \frac{m_k(0) - m_k(s^k) - f(x^k + s^k) + m_k(s^k)}{m_k(0) - m_k(s^k)}\\ &= 1 - \frac{[f(x^k + s^k) - m_k(s^k)]}{m_k(0) - m_k(s^k)}\\ &= 1 - \frac{[m_k(0) - m_k^*]}{[m_k(0) - m_k(s^k)]}\,\frac{\|s^k\|}{[m_k(0) - m_k^*]}\,\frac{[f(x^k + s^k) - m_k(s^k)]}{\|s^k\|}.\end{aligned}$$

First, assumption (A2) ensures that the positive number $\frac{m_k(0) - m_k^*}{m_k(0) - m_k(s^k)}$ is bounded above by $\frac{1}{\tau}$. Second, since $\|s^k\| \le \delta_k$, inequality (4.8) implies

$$\frac{\|s^k\|}{m_k(0) - m_k^*} \le \frac{2\delta_k}{\alpha\delta_k} = \frac{2}{\alpha}.$$

Finally, since δ_k tends to zero and x^k is an accumulation point of the stationary sequence, assumption (A4) implies that

$$\lim_{k\to\infty} \frac{f(x^k + s^k) - m_k(s^k)}{\|s^k\|} = 0.$$

Hence $(f(x^k) - f(x^k + s^k))/(m_k(0) - m_k(s^k))$ tends to unity and thus $x^k + s^k$ will eventually be accepted by the method. This contradicts the stationarity of the sequence. □

The next proposition shows that an accumulation point of a trust region sequence is B-stationary, provided the corresponding sequence of trust radii does not tend to zero.

PROPOSITION 4.4. *Let $\{x^k\}_{k\in\mathbb{N}}$ be an infinite trust region sequence and let x^* be an accumulation point of this sequence. Suppose assumptions* (A1), (A2), *and* (A4) *hold. If x^* is not a B-stationary point of f, then there exists a subsequence $\{x^{k_i}\}_{i\in\mathbb{N}}$ tending to x^* such that the sequence of trust radii δ_{k_i} tends to zero.*

Proof. Since x^* is an accumulation point of the trust region sequence, there exists a subsequence $\{x^{k_i}\}_{i\in\mathbb{N}}$ converging to x^*. In view of Propositions 4.2 and 4.3, we may assume that $x^{k_i} \neq x^{k_i+1}$, i.e., the step $x^{k_i+1} := x^{k_i} + s^{k_i}$ was accepted at each iteration k_i, $i \in \mathbb{N}$. Hence

$$f(x^{k_i}) - f(x^{k_i+1}) > c_0[m_{k_i}(0) - m_{k_i}(s^{k_i})]. \tag{4.9}$$

By assumption (A4.4) we may pass to a subsequence with the property that the sequence of matrices B_{k_i} converges to a matrix B^*. Since x^* is not a B-stationary point of f, assumption (A4.2) shows that it is not a B-stationary point of the function $m^*(.)$ defined by $m^*(s) = \Phi(x^*, s) + \frac{1}{2}s^T B^* s$. Hence, for an arbitrarily chosen $\epsilon > 0$ we can find a point $s^* \in \mathbb{R}^n$ with $\|s^*\| \leq \epsilon$ such that

$$m^*(s^*) < m^*(0). \tag{4.10}$$

Now suppose the statement of the proposition is false, i.e., there exists a scalar $\mu > 0$ such that $\delta_{k_i} \geq \mu$ for every $i \in \mathbb{N}$. Choose $\epsilon \leq \mu$ in such a way that $\Phi(., s^*)$ is upper semicontinuous at x^*, which is possible in view of assumption (A4.3). Hence

$$\limsup_{i \to \infty} m_{k_i}(s^*) \leq m^*(s^*), \tag{4.11}$$

and since $\|s^*\| \leq \epsilon \leq \mu \leq \delta_{k_i}$ for every $i \in \mathbb{N}$, we further obtain

$$m_{k_i}(0) - \min_{\|s\| \leq \delta_{k_i}} m_{k_i}(s) \geq m_{k_i}(0) - m_{k_i}(s^*). \tag{4.12}$$

In view of assumptions (A2) and (A4.2), the inequalities (4.11) and (4.12) imply that

$$\begin{array}{rcl} m_{k_i}(0) - m_{k_i}(s^{k_i}) & \geq & \tau[m_{k_i}(0) - m_{k_i}(s^*)] \\ & \geq & \frac{\tau}{2}[m^*(0) - m^*(s^*)] \end{array} \tag{4.13}$$

for every sufficiently large $i \in \mathbb{N}$. Hence, in view of (4.9), the inequality

$$f(x^{k_i}) - f(x^{k_i+1}) > \frac{c_0 \tau}{2}[m^*(0) - m^*(s^*)]$$

holds for every sufficiently large $i \in \mathbb{N}$. On the one hand, by (4.10), the term on the right-hand side of the foregoing inequality is positive, and hence the series $\sum_{i=1}^{\infty}[f(x^{k_i}) - f(x^{k_i+1})]$ tends to infinity. On the other hand, $f(x^{k_i+1}) \geq f(x^{k_{i+1}})$ since the trust region method is a descent method, and hence

$$\sum_{i=1}^{\infty}[f(x^{k_i}) - f(x^{k_i+1})] \leq \sum_{i=1}^{\infty}[f(x^{k_i}) - f(x^{k_{i+1}})] = f(x^{k_1}) - f(x^*).$$

Thus the assumption that $\delta_{k_i} \geq \mu > 0$ for every $i \in \mathbb{N}$ leads to a contradiction. □

The final proposition shows that an accumulation point of the trust region method is a C-stationary point of the function if it is the limit point of a subsequence x^{k_i} whose corresponding sequence of trust radii δ_{k_i} tends to zero.

PROPOSITION 4.5. *Let $\{x^{k_i}\}_{i \in \mathbb{N}}$ be a convergent subsequence of an infinite trust region sequence $\{x^k\}_{k \in \mathbb{N}}$. Suppose assumptions* (A1)–(A4) *hold. If the sequence of trust radii δ_{k_i} tends to zero, then the limit point x^* of the sequence $\{x^{k_i}\}_{i \in \mathbb{N}}$ is a C-stationary point.*

Proof. Suppose x^* is not a C-stationary point. Then there exists a direction $\bar{y} \in \mathbb{R}^n$ with $\|\bar{y}\| = 1$ and $f^\circ(x^*; \bar{y}) < 0$, where $f^\circ(.;.)$ denotes Clarke's generalized derivative [5]. The definition of $f^\circ(x^*;.)$ yields the existence of positive numbers $\bar{\nu}, \bar{\epsilon}$, and $\bar{\delta}$ such that

$$f(x^k) - f(x^k + \delta_k \bar{y}) \geq \bar{\nu}\delta_k$$

for every $k \in \mathbb{N}$ satisfying

$$\begin{aligned} \|x^k - x^*\| &\leq \bar{\epsilon}, \\ \delta_k &\leq \bar{\delta}. \end{aligned} \tag{4.14}$$

If we set $m_k^* = \min\{m_k(s) \mid \|s\| \leq \delta_k\}$, then we obtain

$$\frac{f(x^k) - m_k^*}{\delta_k} \geq \frac{f(x^k) - m_k(\delta_k \bar{y})}{\delta_k} \geq \bar{\nu} + \frac{f(x^k + \delta_k \bar{y}) - m_k(\delta_k \bar{y})}{\delta_k}$$

for every k satisfying (4.14). Assumptions (A4.3) and (A4.4) imply that

$$\frac{f(x^k) - m_k^*}{\delta_k} \geq \frac{\bar{\nu}}{2} \tag{4.15}$$

for every k satisfying (4.14), provided we have chosen $\bar{\epsilon}$ and $\bar{\delta}$ small enough. Now we show that the ratio r_k is arbitrarily close to 1, provided $\bar{\epsilon}$ and $\bar{\delta}$ are chosen sufficiently small. Expand

$$\begin{aligned} r_k &= \frac{f(x^k) - f(x^k + s^k)}{f(x^k) - m_k(s^k)} \\ &= 1 - \frac{f(x^k + s^k) - m_k(s^k)}{f(x^k) - m_k(s^k)} \\ &= 1 - \frac{f(x^k + s^k) - m_k(s^k)}{\|s^k\|} \frac{\|s^k\|}{\delta_k} \frac{\delta_k}{f(x^k) - m_k^*} \frac{f(x^k) - m_k^*}{f(x^k) - m_k(s^k)}. \qquad (4.16) \end{aligned}$$

Note that

$$\begin{aligned} \frac{\|s^k\|}{\delta_k} &\leq 1, \\ \frac{\delta_k}{f(x^k) - m_k^*} &\leq \frac{2}{\bar{\nu}}, \\ \frac{f(x^k) - m_k^*}{f(x^k) - m_k(s^k)} &\leq \frac{1}{\tau}, \end{aligned}$$

where the second inequality follows from (4.15) and the last inequality follows from assumption (A2) since $m_k(0) = f(x^k)$. Assumptions (A4.3) and (A4.4) ensure that the term

$$\frac{f(x^k + s^k) - m_k(s^k)}{\|s^k\|}$$

will be arbitrarily small, provided k satisfies (4.14) and $\bar{\epsilon} > 0$ and $\bar{\delta} > 0$ are chosen small enough. Hence we conclude that for sufficiently small positive numbers $\bar{\epsilon}$ and $\bar{\delta}$ the inequality

$$r_k > c_1, \ c_1 \in [c_0, 1) \tag{4.17}$$

holds for every k satisfying (4.14). Now consider the subsequence $\{x^{k_i}\}_{i\in\mathbb{N}}$. Reducing $\bar{\delta}$ if necessary, we may assume that $\bar{\delta} \le \delta_{\max}$, where the latter constant is given in (A3). Thus the trust radius will always be increased for sufficiently large k_i, i.e., $\delta_{k_i+1} \ge c_I\delta_{k_i}$, where $c_I > 1$ is the constant given in assumption (A3). Now let k_i be such that $\delta_{k_i} < c_R\bar{\delta}$, where c_R is again given in assumption (A3) and $\|x^{k_i} - x^*\| \le \bar{\epsilon}/2$. Let $j_i \le k_i$ be the smallest index such that all ratios $r_{j_i+1}, \dots, r_{k_i}$ exceed c_1. Let us first show that

$$\|x^{j_i} - x^*\| > \bar{\epsilon}. \tag{4.18}$$

We may assume without loss of generality that $\bar{\epsilon}$ has been chosen small enough so that $\|x^0 - x^*\| > \bar{\epsilon}$ since we have assumed that x^* is not C-stationary and, a fortiori, not B-stationary and thus, in view of Proposition 4.3, $x^* \neq x^0$. Hence (4.18) holds if $j_i = 0$. To see that the inequality also holds for $j_i \ge 1$, note that $\delta_{j_i} \le \bar{\delta}$ since the trust region bound has been increased in the steps $j_i + 1, \dots, k_i$; hence $\delta_{j_i+1} \le \delta_{k_i} \le c_R\bar{\delta}$ and, in view of assumption (A3), $c_R\delta_{j_i} \le \mathit{Reduce}(\delta_{j_i}) \le \delta_{j_i+1}$. Since $r_{j_i} \le c_1$ and (4.17) holds for every k satisfying (4.14), we conclude that (4.18) holds. Since $\|x^{k_i} - x^*\| \le \bar{\epsilon}/2$, we obtain from (4.18) and assumption (A3) that

$$\frac{\bar{\epsilon}}{2} \le \|x^{k_i} - x^{j_i}\| \le \sum_{m=0}^{k_i-j_i-1} \|s^{j_i+m}\| \le \delta_{j_i} + \sum_{\mu=1}^{k_i-j_i-1} \delta_{k_i-\mu}.$$

Note that, in view of assumption (A3), $\delta_{k_i-\mu} \le \frac{\delta_{k_i}}{c_I^\mu}$ for $\mu = 1, \dots, k_i - j_i - 1$. Moreover, again due to (A3), $\delta_{j_i} \le \delta_{k_i}/c_R$. Hence

$$\frac{\bar{\epsilon}}{2} \le \frac{\delta_{k_i}}{c_R} + \sum_{\mu=1}^{\infty} \frac{\delta_{k_i}}{c_I^\mu} = [\frac{1}{c_R} + \frac{1}{c_I - 1}]\delta_{k_i}.$$

Since, by (A3), $c_R > 0$ and $c_I > 1$, we conclude that δ_{k_i} is bounded away from zero which, however, leads to a contradiction as we have assumed that δ_{k_i} tends to zero. Thus x^* is a C-stationary point. □

4.3. A penalty update rule. In order to apply the trust region method to nonsmooth penalty functions $g + \alpha p \circ h$ arising from constrained optimization problems

$$\min\{g(x) \mid p(h(x)) = 0\}$$

with nonnegative function $p \circ h$, one has to specify a penalty parameter α. Rather than determining this parameter in advance, one may update it in the course of the method. To allow for penalty parameter updates, we consider the model functions

$$m_k(s) = g(x^k) + \nabla g(x^k)s + \alpha_k p(h(x^k) + \nabla h(x^k)s) + \frac{1}{2}s^\top B_k s.$$

In this section we show how the penalty update rule suggested by Yuan [47] can be incorporated into our method. Yuan's method determines the sequence of penalty parameters α_k by means of a control sequence β_k which is also determined during the course of the method. In fact, starting from some initial positive parameters α_0 and β_0, the method generates the sequences α_k and β_k which are updated after each accepted step s^k using the following rule.

Penalty update rule for iteration k.

Let $\bar{\alpha}, \bar{\beta}$ be constants independent of k with $0 < \bar{\beta} < 1 < \bar{\alpha}$.

If

$$m_k(0) - m_k(s^k) < \beta_k \alpha_k \min\{\delta_k, p(h(x^k))\},$$

then

$$\alpha_{k+1} = \bar{\alpha}\alpha_k \text{ and } \beta_{k+1} = \bar{\beta}\beta_k,$$

else

$$\alpha_{k+1} = \alpha_k \text{ and } \beta_{k+1} = \beta_k.$$

Theorem 4.6. *Let* $\{x^k\}$ *be a bounded trust region sequence and suppose* (A1), (A2), *and* (A4) *hold.*

1. *If* $\alpha_k \overset{k\to\infty}{\longrightarrow} \infty$, *then there exists a subsequence which converges to a C-stationary point of* $p \circ h$.
2. *If* $\lim_{k\to\infty} \alpha_k < \infty$ *and* (A3) *holds, then every accumulation point of the trust region sequence is a feasible Clarke stationary point of the function* $g + \alpha(p \circ h)$ *for every* $\alpha \geq \lim_{k\to\infty} \alpha_k$.

Proof. Define $\rho_k = \min\{\delta_k, p(h(x^k))\}$.

1. Since $\alpha_k \overset{k\to\infty}{\longrightarrow} \infty$ and $\{x^k\}$ is bounded, there exists a subsequence x^{k_i} such that $\alpha_{k_i+1} > \alpha_{k_i}$ and $x^{k_i} \overset{i\to\infty}{\longrightarrow} x^*$. We first prove that

$$\lim_{i\to\infty} \frac{p(h(x^{k_i})) - \min\limits_{||d||\leq\rho_{k_i}} p(h(x^{k_i}) + \nabla h(x^{k_i})d)}{\rho_{k_i}} = 0. \tag{4.19}$$

Note that $p(h(x^{k_i})) > 0 \;\forall\; i$. Let d^{k_i} be a solution of

$$\min_{||d||\leq\rho_{k_i}} p(h(x^{k_i}) + \nabla h(x^{k_i})d).$$

Since $\alpha_{k_i+1} > \alpha_{k_i}$, we obtain

$$\begin{aligned} \alpha_{k_i}\beta_{k_i}\rho_{k_i} &> m_{k_i}(0) - m_{k_i}(s^{k_i}) \\ &\geq \tau[m_{k_i}(0) - m_{k_i}(s^*)] \\ &\geq \tau[m_{k_i}(0) - m_{k_i}(d^{k_i})] \\ &\geq \tau\Big[- \nabla g(x^{k_i})d^{k_i} - \frac{1}{2}(d^{k_i})^\top B_{k_i} d^{k_i} \\ &\qquad + \alpha_{k_i}\left(p(h(x^{k_i})) - p(h(x^{k_i}) + \nabla h(x^{k_i})d^{k_i})\right)\Big]. \end{aligned} \tag{4.20}$$

Dividing the inequalities by $\alpha_{k_i}\rho_{k_i}$ we obtain

$$\begin{aligned} \beta_{k_i} \quad \geq \quad \tau\Bigg[&\frac{-2\nabla g(x^{k_i})d^{k_i} - (d^{k_i})^\top B_{k_i} d^{k_i}}{2\alpha_{k_i}\rho_{k_i}} \\ &+ \frac{p(h(x^{k_i})) - p(h(x^{k_i}) + \nabla h(x^{k_i})d^{k_i})}{\rho_{k_i}}\Bigg]. \end{aligned} \tag{4.21}$$

Note that

$$\left|\frac{-2\nabla g(x^{k_i})d^{k_i} - (d^{k_i})^\top B_{k_i} d^{k_i}}{2\alpha_{k_i}\rho_{k_i}}\right| \leq \frac{2\|\nabla g(x^{k_i})\|_2 \|d^{k_i}\|_2 + \|d^{k_i}\|_2 \|B_{k_i}\|_2^* \|d^{k_i}\|_2}{\alpha_{k_i}\rho_{k_i}}$$
$$\leq \frac{2\|\nabla g(x^{k_i})\|_2 + \|d^{k_i}\|_2 \|B_{k_i}\|_2^*}{\alpha_{k_i}},$$

where $\|.\|_2$ and $\|.\|_2^*$ denote the ℓ_2 vector and operator norm, respectively. Since B_k is bounded, x^{k_i} converges to x^*, and $\alpha_{k_i} \overset{i\to\infty}{\longrightarrow} \infty$, we conclude that the first term in the right-hand bracket of (4.21) tends to zero. Note that $\beta_{k_i} \overset{i\to\infty}{\longrightarrow} 0$ since $\alpha_{k_i} \overset{i\to\infty}{\longrightarrow} \infty$; hence we conclude that the nonnegative term

$$\frac{p(h(x^{k_i})) - p(h(x^{k_i}) + \nabla h(x^{k_i})d^{k_i})}{\rho_{k_i}}$$

tends to 0 as i tends to ∞, thus proving (4.19).

Now suppose x^* is not a Clarke stationary point of $p(h(.))$, i.e., there exists a descent direction $d \in \mathbb{R}^n : ||d|| = 1$ in x^* of Clarke's generalized derivative [5]. In other words, there exist positive constants $\epsilon, \tilde{\delta}, \gamma$ such that

$$p(h(x)) - p(h(x + \delta d)) > \gamma\delta \quad \forall ||x - x^*|| \leq \epsilon, 0 < \delta \leq \tilde{\delta}$$

and hence, reducing ϵ and $\tilde{\delta}$ if necessary, that

$$p(h(x)) - p(h(x) + \nabla h(x)\delta d)) > \frac{1}{2}\gamma\delta \quad \forall ||x - x^*|| \leq \epsilon, 0 < \delta \leq \tilde{\delta}.$$

Now suppose, on the one hand, that $\rho_{k_i} \leq \tilde{\delta}$. Then

$$\frac{p(h(x^{k_i})) - p(h(x^{k_i}) + \nabla h(x^{k_i})d^{k_i})}{\rho_{k_i}} \geq \frac{p(h(x^{k_i})) - p(h(x^{k_i}) + \nabla h(x^{k_i})\rho_{k_i} d)}{\rho_{k_i}}$$
$$\geq \frac{1}{2}\gamma.$$

If, on the other hand, $\rho_{k_i} > \tilde{\delta}$, then

$$\frac{p(h(x^{k_i})) - p(h(x^{k_i}) + \nabla h(x^{k_i})d^{k_i})}{\rho_{k_i}} \geq \frac{p(h(x^{k_i})) - p(h(x^{k_i}) + \nabla h(x^{k_i})\tilde{\delta} d)}{\rho_{k_i}}$$
$$\geq \frac{1}{2}\gamma\frac{\tilde{\delta}}{\rho_{k_i}}.$$

Note that $\limsup \rho_{k_i}$ is bounded above by $p(h(x^*))$. Hence both inequalities taken together constitute a contradiction to (4.19).

2. Because of the assumption $\lim_{k\to\infty} \alpha_k < \infty$ we know that the penalty parameter sequence remains constant for sufficiently large k, i.e., $\alpha_k = \alpha$ and $\beta_k = \beta$ for every $k > k_0$. Therefore we can apply the convergence results of our trust region method. Hence every accumulation point of the trust region sequence is stationary in the sense of Clarke. We have only to show that these points are also feasible.

Let x^* be an arbitrary accumulation point and (x^{k_i}) a subsequence of successful steps which converges to x^*. Suppose x^* is not a feasible point. Then there exists a positive number $\hat{\delta}$ such that

$$p(h(x^{k_i})) \geq \hat{\delta} \ \forall \ k_i \in \mathbb{N} \text{ large enough.}$$

Hence the definition of the penalty update rule yields

$$m_{k_i}(0) - m_{k_i}(s^{k_i}) \geq \alpha\beta\rho_{k_i} \geq \alpha\beta \min\{\delta_{k_i}, \hat{\delta}\}$$

for every sufficiently large k_i and thus, since the left-hand side of this inequality tends to zero, there exists $\tilde{c} > 0$ such that

$$m_{k_i}(0) - m_{k_i}(s^{k_i}) \geq \tilde{c}\delta_{k_i}. \tag{4.22}$$

We have

$$\infty > f(x^{k_1}) - f(x^*) = \sum_{i=1}^{\infty}[f(x^{k_i}) - f(x^{k_{i+1}})] \geq \sum_{i=1}^{\infty}[f(x^{k_i}) - f(x^{k_i+1})] \geq \sum_{i=1}^{\infty} c_0\tilde{c}\delta_{k_i}$$

and conclude that (δ_{k_i}) is a nullsequence. Expanding r_{k_i} as in (4.16), we conclude from (4.22) that $\lim_{i\to\infty} r_{k_i} = 1$. Now we can apply similar arguments as in the proof of Proposition 4.5 to verify that the corresponding trust radius sequence (δ_{k_i}) is bounded away from zero and thus obtain a contradiction to our assumptions. □

5. An ℓ_1 penalty function for MPECs. In order to apply the proposed trust region method to MPECs, we have to specify the matrices B_k along with a method for the determination of an approximate solution s^k of the subproblem

$$\begin{array}{ll} \min & \nabla g(x^k)s + p(h(x^k) + \nabla h(x^k)s) + \frac{1}{2}s^\top B_k s \\ \text{s.t.} & \|s\| \leq \delta_k. \end{array} \tag{5.1}$$

We shall use a specification which turns into the $S\ell_1QP$ method of Fletcher [10, Chap. 12] if no complementarity constraints are present. We choose the ℓ_∞ norm for the trust region and consider ℓ_1 penalty terms of the form $\phi = p \circ h$, where

$$p(z) = \sum_{i=1}^{m} |\min\{z_i, z_{m+i}\}| + \sum_{v=2m+1}^{2m+p} \max\{z_v, 0\} + \sum_{w=2m+p+1}^{2m+p+q} |z_w|. \tag{5.2}$$

Such penalty terms arise from MPECs with complementarity constraints. The term $|\min\{z_i, z_{m+i}\}|$ is used to penalize the violation of the complementarity conditions $z_i \geq 0$, $z_{m+i} \geq 0$, $z_i z_{m+i} = 0$. Notice that, by our choice of p, the subproblem (5.1) is nonconvex, even if the matrix B_k is positive semidefinite. To overcome the inherent combinatorial difficulties, we suggest solving a related subproblem which is obtained by replacing the nonconvex function p by a suitable convex majorant

$$\begin{aligned} p_I(z) = & \sum_{\substack{i=1 \\ i\in I}}^{m} \max\{|z_i|, -z_{m+i}\} + \sum_{\substack{j=1 \\ m+j\in I}}^{m} \max\{|z_{m+j}|, -z_j\} \\ & + \sum_{v=2m+1}^{2m+p} \max\{z_v, 0\} + \sum_{w=2m+p+1}^{2m+p+q} |z_w|, \end{aligned} \tag{5.3}$$

where $I \subseteq \{1, \ldots, 2m\}$ and for every $i \in \{1, \ldots, m\}$ either $i \in I$ or $m + i \in I$, but not both. Notice that the function p is the pointwise minimum over all functions p_I and if

$$\begin{array}{rcl} \{i \mid h_i(x) < h_{m+i}(x)\} \cup \{m + i \mid h_{m+i}(x) < h_i(x)\} & \subseteq & I, \\ \{i \mid h_i(x) \leq h_{m+i}(x)\} \cup \{m + i \mid h_{m+i}(x) \leq h_i(x)\} & \supseteq & I, \end{array} \tag{5.4}$$

then $p_I(h(x)) = p(h(x))$. Thus, if x is a B-stationary point of $g + \alpha(p \circ h)$, then it is a B-stationary point of $g + \alpha(p_I \circ h)$ for every I satisfying (5.4). Initially we set $I_0 = \{i \mid h_i(x_0) \leq h_{m+i}(x_0)\} \cup \{m+j \mid h_{m+j}(x_0) < h_j(x_0)\}$, where x_0 is the starting point.

To incorporate second-order information in the model we make use of multiplier vectors $\lambda^k \in \mathbb{R}^{2m+p+q}$ which correspond to the Lagrangian

$$L(x, \lambda) = g(x) + h(x)^\top \lambda.$$

These multipliers will be obtained by employing the above majorant of the model function. Initially we set $\lambda^0 = 0$.

At iteration k we are given the current iterate x^k, a multiplier vector λ^k, a trust radius δ_k, a penalty parameter α_k, a matrix B_k, and an index set I_k. We determine s^k as a locally optimal solution of the program

$$(5.5) \qquad \begin{array}{ll} \min & \nabla g(x^k)s + \alpha_k p_{I_k}(h(x^k) + \nabla h(x^k)s) + \frac{1}{2}s^\top B_k s \\ \text{s.t.} & -\delta_k \leq s_t \leq \delta_k, \quad t = 1, \dots, n, \end{array}$$

where $B_k = \nabla^2_{xx} L(x^k, \lambda^k)$. If s^k is a local minimizer of (5.5), then it satisfies Clarke's stationarity condition [2, 5], i.e., there exist multipliers (μ^k, r_k, ξ^k) such that

$$(5.6) \qquad \begin{array}{rcl} \nabla g(x^k)^\top + B_k s^k + \nabla h(x^k)^\top \mu^k + r_k \xi^k & = & 0, \\ \mu^k & \in & \alpha_k \partial p_{I_k}(h(x^k) + \nabla h(x^k)s^k), \\ \xi^k & \in & \partial \|s^k\|_\infty, \\ r_k & \geq & 0, \\ r_k(\|s^k\| - \delta_k) & = & 0. \end{array}$$

We solve (5.5) by rewriting it as the quadratic program

$$\begin{array}{llll} \min & \multicolumn{3}{l}{\nabla g(x^k)s + \alpha_k[\sum_{i=1}^m \gamma_i + \sum_{v=1}^p \rho_v + \sum_{w=1}^q \sigma_w] + \frac{1}{2}s^\top B_k s} \\ \text{s.t.} & & & \\ -\gamma_i \leq & h_i(x^k) + \nabla h_i(x^k)s & \leq \gamma_i, & i \in I_k \cap \{1, \dots, m\}, \\ -\gamma_i \leq & h_{m+i}(x^k) + \nabla h_{m+i}(x^k)s, & & i \in I_k \cap \{1, \dots, m\}, \\ -\gamma_j \leq & h_{m+j}(x^k) + \nabla h_{m+j}(x^k)s & \leq \gamma_j, & m+j \in I_k \cap \{m+1, \dots, 2m\}, \\ -\gamma_j \leq & h_j(x^k) + \nabla h_j(x^k)s, & & m+j \in I_k \cap \{m+1, \dots, 2m\}, \\ & h_{2m+v}(x^k) + \nabla h_{2m+v}(x^k)s & \leq \rho_v, & v = 1, \dots, p, \\ & 0 & \leq \rho_v, & v = 1, \dots, p, \\ -\sigma_w \leq & h_{2m+p+w}(x^k) + \nabla h_{2m+p+w}(x^k)s & \leq \sigma_w, & w = 1, \dots, q, \\ -\delta_k \leq & s_t & \leq \delta_k, & t = 1, \dots, n. \end{array}$$

Given the Lagrange multipliers for this quadratic program at a local minimizer s^k, we define a vector $(\mu^k, \nu^k) \in \mathbb{R}^{2m+p+q} \times \mathbb{R}^n$ as follows. Let $1 \leq i \leq 2m+p+q$. If the term $h_i(x^k) + \nabla h_i(x^k)s$ is constrained from two sides, e.g., $-\gamma_i \leq h_i(x^k) + \nabla h_i(x^k)s \leq \gamma_i$, and $\mu^k_{i,+} \geq 0$ and $\mu^k_{i,-} \leq 0$ are the multipliers corresponding to the right-hand and left-hand sides, respectively, of the two-sided inequality, then we set $\mu^k_i = \mu^k_{i,+} + \mu^k_{i,-}$. Otherwise, we let μ^k_i be the multiplier corresponding to the one-sided inequality constraining the term $h_i(x^k) + \nabla h_i(x^k)s$. The multipliers corresponding to $0 \leq \rho_v$ are neglected and we finally set $\nu^k_t = \nu^k_{t,+} + \nu^k_{t,-}$, where $\nu^k_{t,+} \geq 0$ and $\nu^k_{t,-} \leq 0$ are the multipliers corresponding to the right-hand and left-hand sides, respectively, of the two-sided inequality $-\delta_k \leq s_t \leq \delta_k$. A straightforward but rather lengthy comparison of the KKT conditions of the above quadratic program with the stationarity conditions

(5.6) shows that (μ^k, ν^k) is of the form $(\mu^k, r_k\xi^k)$, where (μ^k, r_k, ξ^k) satisfies (5.6). Note that this implies in particular that

$$\|\mu^k\|_\infty \leq \alpha_k, \tag{5.7}$$

since the subdifferential of every function p_I at every point is contained in the ℓ_∞ unit-ball.

If step s^k is rejected by the trust region method, then we set $\lambda^{k+1} := \lambda^k$ and $I_{k+1} := I_k$. Otherwise, we set $\lambda^{k+1} := \mu^k$ and determine the index set I_{k+1} by the following rule, which makes use of two predetermined positive tolerances τ_1, τ_2. For every $i = 1, \dots, m$ we have to decide whether to include i or $m + i$ in I_{k+1}:

If $\max\{|h_i(x^{k+1})|, |h_{m+i}(x^{k+1})|\} < \tau_1$ and $\lambda_i^{k+1} > \tau_2$, then $m + i \in I_{k+1}$,
else if $\max\{|h_i(x^{k+1})|, |h_{m+i}(x^{k+1})|\} < \tau_1$ and $\lambda_{m+i}^{k+1} > \tau_2$, then $i \in I_{k+1}$,
    else if $h_i(x^{k+1}) \leq h_{m+i}(x^{k+1})$, then $i \in I_{k+1}$,
        else $m + i \in I_{k+1}$.

The first part of this rule needs some explanation. Let us focus attention on the first if statement. Note first that the multiplier λ_i^{k+1} can be positive only if i is already contained in I_k, i.e., we had forced h_i to be close to zero at the last iteration and allowed the complementary function h_{m+i} to become positive. The first if statement, however, indicates that the objective function seems to force the function h_{m+i} to be close to zero anyway. Therefore, instead of allowing h_{m+i} to become positive and forcing h_i to be zero, we swap, i.e., in the next step we force h_{m+i} to be close to zero and allow h_i to become positive. This is done by including $m+i$ in I_{k+1}. The positive multiplier λ_i^{k+1} indicates that this swap seems to be favorable.

5.1. Relation to SQP methods. If the gradients $\nabla h_i(x^*)$ of the active functions h_i are linearly independent and x^* is a local minimizer of the MPEC, then there exists a uniquely determined Lagrange multiplier vector $\lambda^* \in \mathbb{R}^{2m+p+q}$ such that

(5.8)

$$\begin{aligned}
\nabla g(x^*)^\top + \nabla h(x^*)^\top \lambda^* &= 0, \\
h_i(x^*)\lambda_i^* &= 0 \quad \forall i = 1, \dots, 2m + p, \\
\lambda_i^*, \lambda_{m+i}^* &\leq 0 \quad \forall i \in \{1, \dots, m\} \text{ s.t. } h_i(x^*) = h_{m+i}(x^*) = 0, \\
\lambda_l^* &\geq 0 \quad \forall l = 2m + 1, \dots, 2m + p,
\end{aligned}$$

and this is equivalent to B-stationarity of x^* for $g + \alpha p \circ h$ for every sufficiently large α, [20, 40]. The next proposition gives conditions under which the estimates λ^k of the method converge to the Lagrange multipliers λ^*.

Proposition 5.1. *Suppose the sequence of matrices B_k is bounded, the sequence of iterates x^k tends to a feasible C-stationary point x^*, and the gradients of the active functions h_i at x^* are linearly independent. Then there exists a unique multiplier λ^* such that*

$$\begin{aligned}
\nabla g(x^*)^\top + \nabla h(x^*)^\top \lambda^* &= 0, \\
h_i(x^*)\lambda_i^* &= 0 \quad \forall i = 1, \dots, 2m + p + q, \\
\lambda_l^* &\geq 0 \quad \forall l = 2m + 1, \dots, 2m + p.
\end{aligned} \tag{5.9}$$

If the trust radius sequence δ_k is bounded away from zero, then there exists a positive scalar $\bar{\tau}_1$, such that for every choice of the tolerance $\tau \in (0, \bar{\tau}_1)$ the sequence λ^k

converges to λ^*. *Moreover, there exists a second positive scalar* $\bar\tau_2$ *such that*

$$\lambda_i^*, \lambda_{m+i}^* \le 0 \quad \forall i \in \{1,\dots,m\} \ \text{ s.t. } \ h_i(x^*) = h_{m+i}(x^*) = 0,$$

provided $\tau_2 \in [0, \bar\tau_2)$.

Proof. The existence and uniqueness of λ^* satisfying (5.9) follows from the definition of C-stationarity and the linear independence assumption. Let

$$\bar\tau_1 = \min\{h_j(x^*) \mid h_j(x^*) > 0, \ \ j = 1,\dots,2m\}$$

and suppose $0 < \tau_1 < \bar\tau_1$. Note first that for sufficiently large k the index set $I = I_k$ satisfies (5.4) with $x = x^*$ and hence $p_{I_k}(h(x^*)) = p(h(x^*))$. Indeed, for every $i = 1,\dots,m$ the equation $\min\{h_i(x^*), h_{m+i}(x^*)\} = 0$ holds since x^* is feasible. Hence, if, on the one hand, $0 = h_i(x^*) < h_{m+i}(x^*)$ and k is sufficiently large, then $\max\{|h_i(x^k)|, |h_{m+i}(x^k)|\} > \tau_1$ and thus $i \in I_k$. On the other hand, if $i \in I_k$, then $h_i(x^*) \le h_{m+i}(x^*)$ because, if $0 = h_{m+i}(x^*) < h_i(x^*)$, then again $\max\{|h_i(x^k)|, |h_{m+i}(x^k)|\} > \tau_1$ for sufficiently large k and hence $m + i \in I_k$ and $i \notin I_k$. The same argument holds with i and $m + i$ swapped.

Recall that $\lambda^{k+1} = \lambda^k$ if s^k is not accepted by the method. Let s^{k_i} be the subsequence of accepted steps. Since the steps s^{k_i} tend to zero and hence the corresponding trust radius is thus eventually inactive, (5.6) implies that λ^{k_i+1} satisfies

$$\begin{array}{rcl} 0 & = & \nabla g(x^{k_i})^\top + B_{k_i} s^{k_i} + \nabla h(x^{k_i})^\top \lambda^{k_i+1}, \\ \lambda^{k_i+1} & \in & \alpha_{k_i} \partial p_{I_{k_i}}(h(x^{k_i}) + \nabla h(x^{k_i}) s^{k_i}), \end{array}$$

provided k_i is sufficiently large. If $h_j(x^*) \ne p(h(x^*)) = 0$, then $h_j(x^*) \ne p_{I_{k_i}}(h(x^*)) = p(h(x^*))$; hence $h_j(x^{k_i}) + \nabla h_j(x^{k_i}) s^{k_i} \ne 0$ for sufficiently large k_i and thus $\lambda_j^{k_i+1} = 0$. Now consider the chain of inequalities

$$\begin{aligned} 0 &= \|\nabla g(x^{k_i})^\top + \nabla h(x^{k_i})^\top \lambda^{k_i+1} + B_{k_i} s^{k_i}\| \\ &= \|\nabla g(x^{k_i})^\top + \nabla h(x^{k_i})^\top \lambda^{k_i+1} + B_{k_i} s^{k_i} \\ &\quad \underbrace{-\nabla g(x^*)^\top - \nabla h(x^*)^\top \lambda^*}_{=0} + (\nabla h(x^*) - \nabla h(x^*))^\top \lambda^{k_i+1}\| \\ &\ge \|\nabla h(x^*)^\top (\lambda^{k_i+1} - \lambda^*)\| - \\ &\quad \underbrace{\|\nabla g(x^{k_i})^\top - \nabla g(x^*)^\top - (\nabla h(x^*) - \nabla h(x^{k_i}))^\top \lambda^{k_i+1} + B_{k_i} s^{k_i}\|}_{\to 0 \quad \text{if} \quad k_i \to \infty}. \end{aligned}$$

We conclude that

$$\lim_{k_i \to \infty} \|\nabla h(x^*)^\top (\lambda^{k_i+1} - \lambda^*)\| = 0. \tag{5.10}$$

Since $\lambda_j^{k_i+1} = 0$ if $h_j(x^*) \ne p(h(x^*)) = 0$, we obtain

$$\nabla h(x^*)^\top (\lambda^{k_i+1} - \lambda^*) = \nabla h_{I_0}(x^*)^\top (\lambda_{I_0}^{k_i+1} - \lambda_{I_0}^*), \tag{5.11}$$

where I_0 is the set of indices of those functions h_j which vanish at x^*. The convergence thus follows from (5.10), (5.11), and the full column rank of $\nabla h_{I_0}(x^*)^\top$.

Finally, suppose $h_i(x^*) = h_{m+i}(x^*)$ and $\lambda_i^* > 0$. Then there exists $\bar k$ such that $\lambda_i^k > 0$ for every $k \ge \bar k$; hence

$$i \in I_k \quad \forall k \ge \bar k. \tag{5.12}$$

Now let

$$\tau_2 < \min\{\lambda_i^* \mid \lambda_i^* > 0, \quad i = 1, \ldots, 2m\}.$$

Enlarging $\bar{k}$ if necessary, we may assume that $\max\{|h_i(x^k)|, |h_{m+i}(x^k)|\} < \tau_1$ for every $k \geq \bar{k}$. Since λ^k converges to λ^*, we may further assume that $\lambda_i^k > \tau_2$ for every $k \geq \bar{k}$. Now let $k \geq \bar{k}$ be such that s_k is accepted and hence $\lambda^{k+1} = \mu^k$. Since $\lambda^{k+1} > \tau_2$, the rule for the determination of I_{k+1} yields $m + i \in I_{k+1}$, contradicting (5.12). $\square$

Note that we solve in iteration k the subproblem of Fletcher's trust region SNQP method [10] applied to the nonlinear program

$$\begin{array}{lll} \min & g(x), & \\ \text{s.t.} & h_i(x) = 0, & i \in I_k, \\ & h_j(x) \geq 0, & j = \{1, \ldots, 2m\} \backslash I_k, \\ & h_l(x) \leq 0, & l = 2m+1, \ldots, 2m+p, \\ & h_r(x) = 0, & r = 2m+p+1, \ldots, 2m+p+q. \end{array} \tag{5.13}$$

If the trust region bound is eventually inactive and the sequence of iterates converges, then the trust region SNQP method turns into a local SNQP method. Theorem 14.4.1 of [10] gives conditions under which the nonglobalized SNQP method turns asymptotically into an SQP method and thus inherits the favorable local convergence properties. If these conditions hold, then

$$p_{I_k}(h(x^k) + \nabla h(x^k) s^k) = 0. \tag{5.14}$$

If, in addition, upper-level strict complementarity holds, i.e., $\lambda_i^* \lambda_{m+i}^* \neq 0$ for every i with $h_i(x^*) = h_{m+i}(x^*)$, then the first branch in the rule for determination of I_k will not be used for sufficiently large k and thus $p_{I_k}(h(x^k)) = p(h(x^k))$. This equation together with (5.14) implies that the step s^k of the SNQP method is accepted by our method, provided it is accepted by the trust region SNQP method [10], since the ratio of actual reduction by predicted reduction in our method is at least as large as the same ratio with p replaced by p_{I_k}. Note that our method may change the index set I_k after each iteration. However, a direct extension of the proof of [20, Thm. 6.4.3] shows that the convergence results for the SNQP method carry over to the present case.

The most severe assumption in the above arguments is the boundedness of $1/\delta_k$, [46]. In order to ensure this, Fletcher [10] suggests the use of second-order correction steps and, indeed, Yuan [45] shows that under mild assumptions the trust region radius is bounded away from zero if second-order correction steps are used. The incorporation of second-order correction steps into our method, however, would require a modification of the outer trust region framework and is beyond the scope of the present study.

6. Numerical experiments. We have tested the method on various small-scale examples of MPECs. The performance of the method was found to be comparable with the reported performance of the $S\ell_1QP$ method in the literature [10]. Occasionally, slow convergence due to the Maratos effect has been observed. As mentioned above, this phenomenon could possibly be prevented if second-order correction steps were included. A very important feature of the method is its apparent insensitivity to violated strict complementarity in the lower-level problem at the optimal solution.

We give some numerical results for two particular instances, a simple three-dimensional example and a five-dimensional example from the literature. For the

Table 6.1
$f(x,y) = (x_1+1)^2 + (x_2-2.5)^2 + (x_3+1)^2$.

Iteration	(x^k, y_1^k, y_2^k)	δ_k	α_k	PFV	FV
0	(1.0000,1.0000,1.0000)	1	4	28.9963	10.2500
1	(0.3673,1.9970,0.3673)	2	4	7.5552	3.9922
2	(0.0000,1.8270,0.0000)	4	4	3.8370	2.4529
3	(0.0000,2.0000,0.0000)	8	4	2.2500	2.2500
4	(0.0000,2.5000,0.0000)	16	4	2.0000	2.0000

Table 6.2
$f(x,y) = (x_1+1)^2 + x_2^2 + 10(x_3-1)^2$.

Iteration	(x^k, y_1^k, y_2^k)	δ_k	α_k	PFV	FV
0	(1.0000,1.0000,1.0000)	1	4	22.7463	5.0000
1	(0.0000,1.5632,0.5751)	1	4	10.1055	5.2493
2	(-0.7627,1.4775,0.7021)	2	8	11.1816	3.1268
3	(-0.3646,2.1074,0.6040)	4	16	13.0761	6.4132
4	(0.0000,2.6579,0.5389)	4	16	11.0891	10.1911
5	(0.0000,2.7101,0.5365)	8	16	10.4926	10.4924
6	(0.0000,2.7101,0.5365)	16	16	10.4925	10.4925
7	(0.0000,2.7101,0.5365)	32	16	10.4925	10.4925

experiments the method was implemented in MATLAB 5.0 and the built-in QP-Solver was used for the solution of the subproblems. In the examples below, we have used the following parameters: $\alpha_0 = 4, \beta_0 = 1, \bar{\alpha} = 2, \bar{\beta} = 0.25, c_0 = 0.25, c_1 = 0.75, \delta_0 = 1, \tau_1 = 10^{-6}, \tau_2 = 0$, *Reduce*$(\delta) = 0.25\delta$, *Increase*$(\delta) = 2\delta$. Stopping criterion for the method was that $\|\nabla g(x^k)^\top + \nabla h(x^k)^\top \lambda^k\| \leq 10^{-10}$ and that the multipliers corresponding to violated strict complementarity in the lower level are less than τ_1.

6.1. A three-dimensional example. We first tested the method on the problem

$$\begin{array}{ll} \min & f(x,y) \\ \text{s.t.} & \min\{x, -e^x + y_1 - e^{y_2}\} = 0, \\ & y_2 \geq 0 \end{array}$$

for various quadratic objective functions f. Tables 6.1 and 6.2 show two typical runs. The abbreviations FV and PFV refer to the values of $f((x,y)^k)$ and $f((x,y)^k) + \alpha_k p(h((x,y)^k))$, respectively. Note that the point $(0,2,0)$ is a Clarke stationary point and that this point is localized at iteration 3 of the first example. However, one of the corresponding Lagrange multipliers is positive and thus a change of the index set I_k occurs, which then allows the method to find the minimizer in iteration 4. Strict complementarity in the lower-level problem is violated at the optimal point in the second example, i.e., both functions in the min expression are active. This has no negative effect on the performance of the method.

6.2. Outrata's example. The following example is due to Outrata [25], who used it to illustrate an implicit function-based bundle trust region method for MPECs. The example also was applied in [9] to illustrate a homotopy method. The MPEC is

Table 6.3
Number of solved QPs in Outrata's example.

	f_1	f_2	f_3	f_4
$y^0 = 0$	20	27	14	11
$y^0 = 10$	9	9	10	10

of the form

$$\begin{array}{rl} \min & f(x,y), \\ \text{s.t.} & \min\{x_1, (1+0.2y)x_1 - (1+1.333y) - 0.333x_3 + 2x_1x_4\} = 0, \\ & \min\{x_2, (1+0.1y)x_2 - y + x_3 + 2x_2x_4\} = 0, \\ & \min\{x_3, 0.333x_1 - x_2 + 1 - 0.1y\} = 0, \\ & \min\{x_4, 9 + 0.1y - x^2 - x^2\} = 0, \\ & 0 \le y \le 10. \end{array}$$

The equilibrium constraints are stationary point conditions of a convex programming problem and are uniquely solvable for every feasible parameter value y. This fact has been exploited by Outrata [25] and is also important in the theoretical convergence analysis of [9]. Although our method does not make use of such information, it gives competitive results, as reported in Table 6.3. The method was started at the points $x^0 = (0,0,0,0)$ and $y^0 = 0$ as well as $y^0 = 10$ with the objective functions

$$\begin{aligned} f_1(x,y) &= \tfrac{1}{2}[(x_1-3)^2 + (x_2-4)^2], \\ f_2(x,y) &= \tfrac{1}{2}[(x_1-3)^2 + (x_2-4)^2 + (x_3-1)^2], \\ f_3(x,y) &= \tfrac{1}{2}[(x_1-3)^2 + (x_2-4)^2 + 10x_4^2], \\ f_4(x,y) &= \tfrac{1}{2}[(x_1-3)^2 + (x_2-4)^2 + (x_3-1)^2 + (x_4-1)^2 + y^2]. \end{aligned}$$

The method detected the optimal solutions reported in [25] and the numerical results indicate local quadratic convergence in all cases. Note that lower-level strict complementarity is violated at the optimal solution corresponding to the objective function f_1.

The method showed similar behavior on the small-scale test problems used in [9]. For all these problems, global convergence and local quadratic convergence to the local minimum reported in [9] was observed.

7. Conclusion. We have studied theoretical and computational aspects of the exact penalization approach to MPECs and more general composite piecewise smooth programs. In the first part, we have extended the local error bound result of Robinson [35] to piecewise smooth programs and then specified the result to MPECs. This yields a theoretical justification for the use of exact penalization methods. We then proposed a general trust region method for composite piecewise smooth functions which combines the approaches of Fletcher [10] and Dennis, Li, and Tapia [7] and includes the penalty update rule of Yuan [47]. We have proved a global convergence result. Finally, we have specified the method to make it applicable for MPECs with complementarity constraints. The resulting method extends Fletcher's $S\ell_1QP$ method to MPECs. Some preliminary numerical results are reported.

Acknowledgments. We thank Daniel Ralph for pointing out that the piecewise Mangasarian–Fromovitz constraint qualification discussed in section 3.2.1 gives rise to a local error bound. We also wish to thank two anonymous referees for their valuable remarks and suggestions.

REFERENCES

[1] G. Anandalingam and T. Friesz, eds., *Hierarchical optimization*, Ann. Oper. Res., 34 (1992).

[2] J. V. Burke, *An exact penalization viewpoint of constrained optimization*, SIAM J. Control Optim., 29 (1991), pp. 968–998.

[3] R. W. Chaney, *Piecewise C^k-functions in nonsmooth analysis*, Nonlinear Anal., 15 (1990), pp. 649–660.

[4] Y. Chen and M. Florian, *The nonlinear bilevel programming problem: Formulations, regularity, and optimality conditions*, Optimization, 32 (1995), pp. 193–209.

[5] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, Chichester, UK, 1983.

[6] J. P. Dedieu, *Penalty functions in subanalytic optimization*, Optimization, 26 (1992), pp. 27–32.

[7] J. E. Dennis, Jr., S. B. Li, and R. A. Tapia, *A unified approach to global convergence of trust region methods for nonsmooth optimization*, Math. Programming, 68 (1995), pp. 319–346.

[8] I. I. Eremin, *The penalty method in convex programming*, Soviet Math. Dokl., 8 (1966), pp. 459–462.

[9] F. Facchinei, H. Jiang, and L. Qi, *A Smoothing Method for Mathematical Programs with Equilibrium Constraints*, AMR 96/15, Applied Mathematics Report, University of New South Wales, Sydney, Australia, 1996, Math. Programming Ser. A, to appear.

[10] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., John Wiley & Sons, Chichester, UK, 1987.

[11] R. Henrion and D. Klatte, *Metric regularity of the feasible set mapping in semi-infinite optimization*, Appl. Math. Optim., 30 (1994), pp. 103–109.

[12] R. Janin, *Directional derivative of the marginal function in nonlinear programming*, Math. Programming Study, 21 (1984), pp. 110–126.

[13] H. T. Jongen and D. Pallaschke, *On linearization and continuous selections of functions*, Optimization, 19 (1988), pp. 343–353.

[14] H. T. Jongen and G. W. Weber, *Nonlinear optimization: Characterization of structural stability*, J. Global Optim., 1 (1991), pp. 47–64.

[15] M. Kojima, *Strongly stable stationary solutions in nonlinear programming*, in Analysis and Computation of Fixed Points, S.M. Robinson, ed., Academic Press, New York, 1980, pp. 93–138.

[16] D. Kuhn and R. Löwen, *Piecewise affine bijections of $\mathbb{R}^n$ and the equation $Sx^+ - Tx^- = y$*, Linear Algebra Appl., 96 (1987), pp. 109–129.

[17] L. Kuntz and S. Scholtes, *Structural analysis of nonsmooth mappings, inverse functions, and metric projections*, Math. Anal. Appl. 188 (1994), pp. 59–75.

[18] L. Kuntz and S. Scholtes, *A nonsmooth variant of the Mangasarian-Fromovitz constraint qualification*, J. Optim. Theory Appl., 82 (1994), pp. 59–75.

[19] Z. Q. Luo, J.-S. Pang, D. Ralph, and S. Q. Wu, *Exact penalization and stationarity conditions for mathematical programs with equilibrium constraints*, Math. Programming, 75 (1996), pp. 19–76.

[20] Z. Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, UK, 1996.

[21] Z. Q. Luo, J.-S. Pang, and D. Ralph, *Piecewise sequential quadratic programming for mathematical programs with nonlinear complementarity constraints*, in Multilevel Optimization: Algorithms and Applications, A. Migdalas et al., eds., Kluwer, Dordrecht, The Netherlands, 1998, pp. 209–229.

[22] P. Marcotte and D. L. Zhu, *Exact and inexact penalty methods for the generalized bilevel programming problem*, Math. Programming, 74 (1996), pp. 141–157.

[23] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.

[24] J. V. Outrata, *On the numerical solution of a class of Stackelberg problems*, ZOR—Meth. Model Oper. Res., 34 (1990), pp. 255–277.

[25] J. V. Outrata, *On Optimization Problems with Variational Inequality Constraints*, SIAM J. Optim., 4 (1994), pp. 340–357.

[26] J. V. Outrata and J. Zowe, *A numerical approach to optimization problems with variational inequality constraints*, Math. Programming, 68 (1995), pp. 105–130.

[27] J.-S. Pang, S.-P. Han, and N. Rangaraj, *Minimization of locally Lipschitzian functions*, SIAM J. Optim., 1 (1991), pp. 57–82.

[28] J.-S. Pang and D. Ralph, *Piecewise smoothness, local invertibility, and parametric analysis of normal maps*, Math. Oper. Res., 21 (1996), pp. 401–426.

[29] R. Poliquin and L. Qi, *Iteration functions in some nonsmooth optimization algorithms*, Math. Oper. Res., 20 (1995), pp. 479–496.

[30] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.

[31] L. Qi and J. Sun, *A trust region algorithm for minimization of locally Lipschitzian functions*, Math. Programming, 66 (1994), pp. 25–43.

[32] D. Ralph, *On branching numbers of normal manifolds*, Nonlinear Anal., 22 (1994), pp. 1041–1050.

[33] D. Ralph and S. Scholtes, *Sensitivity analysis of composite piecewise smooth equations*, Math. Programming, 76 (1997), pp. 593–614.

[34] D. Ralph, *Sequential quadratic programming for mathematical programs with linear complementarity constraints*, in Proceedings of the Seventh Conference on Computational Techniques and Applications, R. L. May and A. K. Easton, eds., Scientific Press, Singapore, 1996, pp. 663–668.

[35] S. M. Robinson, *Stability theory for systems of inequalities, Part* II*: Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.

[36] S. M. Robinson, *Some continuity properties of polyhedral multifunctions*, Math. Programming Study, 14 (1981), pp. 206–214.

[37] S. M. Robinson, *Local structure of feasible sets in nonlinear programming, Part* III, Math. Programming Study, 30 (1987), pp. 45–66.

[38] S. M. Robinson, *Normal maps induced by linear transformations*, Math. Oper. Res., 17 (1992), pp. 691–714.

[39] H. Scheel, *Ein Straffunktionsansatz für Optimierungsprobleme mit Gleichgewichtsrestriktionen*, diploma thesis, Institut für Statistik und Mathematische Wirtschaftstheorie, Universität Karlsruhe, Karlsruhe, Germany, 1995.

[40] H. Scheel and S. Scholtes, *Mathematical Programs with Complementarity Constraints: Stationarity, Optimality, and Sensitivity*, manuscript, Department of Engineering, University of Cambridge, Cambridge, UK, 1998.

[41] S. Scholtes, *Introduction to Piecewise Differentiable Equations*, habilitation thesis, Institut für Statistik und Mathematische Wirtschaftstheorie, Universität Karlsruhe, Karlsruhe, Germany, 1994.

[42] S. Scholtes, *A proof of the branching number bound for normal manifolds*, Linear Algebra Appl., 246 (1996), pp. 83–95.

[43] H. Schramm and J. Zowe, *A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results*, SIAM J. Optim., 2 (1992), pp. 121–152.

[44] M. Stöhr, *Global konvergente Straffunktionsmethoden zur Behandlung von Optimierungsproblemen mit Gleichgewichtsrestriktionen*, diploma thesis, Institut für Statistik und Mathematische Wirtschaftstheorie, Universität Karlsruhe, Karlsruhe, Germany, 1996.

[45] Y. Yuan, *On the superlinear convergence of a trust region algorithm for nonsmooth optimization*, Math. Programming, 31 (1985), pp. 269–285.

[46] Y. Yuan, *An example of only linear convergence of trust region algorithms for nonsmooth optimization*, IMA J. Numer. Anal., 4 (1984), pp. 327–335.

[47] Y. Yuan, *On the convergence of a new trust region algorithm*, Numer. Math., 70 (1995), pp. 515–539.

[48] W. I. Zangwill, *Nonlinear programming via penalty functions*, Management Sci., 13 (1967), pp. 344–358.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 2, pp. 653–671

ASYMPTOTIC ANALYSIS FOR PENALTY AND BARRIER METHODS IN VARIATIONAL INEQUALITIES[*]

A. AUSLENDER[†]

Abstract. We study the convergence of a class of penalty and barrier methods for solving monotone variational inequalities with constraints. This class of methods is an extension of penalty and barrier methods for convex optimization to the setting of variational inequalities. Primal convergence is established under weaker conditions than usual: the solution set is supposed either to be a compact set or, for the case of interior barrier methods, the sum of a compact set and a linear space. Dual convergence is also analyzed. The analysis strongly exploits a new formula related to recession calculus.

Key words. variational inequalities, maximal monotone operators, convex analysis, recession functions, penalty and barrier methods

AMS subject classifications. 90C25, 90C30, 90C31, 49J40, 47H05

PII. S0363012996310909

1. Introduction.

Let C be a nonempty closed convex set in $\mathbb{R}^N$ and let A be a multivalued map from $\mathbb{R}^N$ into $\mathbb{R}^N$ such that dom $A := \{x \in \mathbb{R}^N : A(x) \neq \phi\}$ contains C. In this paper we consider the variational inequality

$$\text{(VI)} \qquad \begin{array}{l} \text{“find } x^* \in C \text{ and } c^* \in A(x^*) \text{ satisfying} \\ \langle c^*, x - x^* \rangle \geq 0 \quad \forall x \in C\text{”} \end{array}$$

and we denote by S the set of such x^*, solutions of (VI).

All the notations and definitions used in this paper are standard in convex analysis and can be found in Rockafellar's book [18].

When A is a monotone operator, most of the methods for solving (VI) are extensions of well-known methods in optimization theory. In optimization, A is the subdifferential ∂f_0 of a closed proper convex function f_0, and (VI) consists of minimizing f_0 on C. The usual assumption which ensures convergence is that the optimal set S is a nonempty compact set, which can be expressed equivalently in many ways, in particular, in terms of recession analysis (see for example [18, Thm. 27.1]). These characterizations are, in general, useful to prove convergence of numerical methods. In contrast, when (VI) does not reduce to an optimization problem, most convergence results are based on a coercivity assumption stronger than the compactness of the solution set S.

For instance, it is often supposed that the "feasible set is compact" (see, for example, Fukushima [11]; Marcotte and Dussault [15]) or that there exists $v_0 \in C$ such that

$$\lim_{\|x\| \to +\infty} \left\langle c(x), \frac{x - v_0}{\|x - v_0\|} \right\rangle = +\infty \quad \text{with } c(x) \in A(x),$$

[*] Received by the editors October 21, 1996; accepted for publication (in revised form) June 26, 1998; published electronically February 2, 1999.
http://www.siam.org/journals/sicon/37-2/31090.html

[†] Laboratoire d'Econométrie, Ecole Polytechnique, 5 rue Descartes, Paris 75005, France and Université de Paris I, Pantheon Sorbonne, 90 rue de Tolbiac, Paris 75013, France (auslen@poly.polytechnique.fr).

a condition satisfied by strongly monotone operators (see for example Pang and Chan [17]).

If we consider the classical exterior penalty methods and interior barrier methods which were introduced more than twenty years ago, there have been no proofs of convergence under the weaker assumption that the solution set S is compact, as is the case for optimization problems. There is a real technical difficulty. So the first challenge of this paper is to find completely new techniques to overcome this difficulty. To do this, in section 2 we establish a simple but truly new formula (formula (2.11)) related to recession analysis for monotone operators.

This formula is of particular interest. It plays a fundamental role not only in proving convergence of algorithms but also in characterizing the compactness of the set of solutions S. Such a characterization has recently been established independently by Crouzeix [10] and by Rockafellar and Wets [19]. In [10], A is pseudomonotone, upper semicontinuous, and for each $x \in C$, $A(x)$ is compact (which is unnecessary in our case). In [19] the result is given for A assumed monotone, single valued, and continuous on C. But in fact it can be stated, without effort, for more general situations as an immediate consequence of formula (2.11). We shall do so in section 2.

When S is a nonempty compact set we say that (VI) is "coercive"; when S is a sum of a nonempty compact set and a linear space we say that (VI) is "weakly coercive." These properties are discussed in section 2, and then by exploiting the recession formula (2.11), the proof of convergence of the algorithms will be established, assuming only coercivity, and for interior methods weak coercivity.

Very recently Auslender, Cominetti, and Haddou [4] introduced a wide class of penalty and barrier methods for convex programming including most of the specific functions proposed in the literature as well as some new ones. They gave also a systematic way to generate penalty and barrier functions in this class. This clearly calls for an extension of the approach to variational inequalities. This is done in detail in section 3 and is the main purpose of this paper. The idea is the following: Suppose the feasible set C is defined by

$$C = \{x \in \mathbb{R}^N : f_i(x) \leq 0, \quad i = 1, 2 \dots m\}$$

where $f_i : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ are closed proper convex functions.
In order to solve the variational inequality (VI) we approximate it by solving a family of unconstrained generalized equations of the form

$(\mathrm{VI})_r$ find x_r satisfying

$$0 \in A(x_r) + F_r(x_r) \tag{1.1}$$

with

$$F_r(x) = \alpha(r)\partial g_r(x), \quad g_r(x) = \sum_{i=1}^{m} \theta\left(\frac{f_i(x)}{r}\right), \text{ if } x \in \bigcap_{i=1}^{m} \mathrm{dom}\ f_i \text{ , } +\infty \text{ otherwise,} \tag{1.2}$$

where $\partial g_r(x)$ is the subdifferential of g_r at x and $r > 0$ is a penalty parameter which will ultimately go to 0. The functions $\theta : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ and $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ are such that θ is closed, proper, convex, and nondecreasing with

$$\mathrm{dom}\theta := \{u : \theta(u) < +\infty\} =]-\infty, \eta[, \quad 0 \leq \eta \leq +\infty,$$

and

$$\lim_{r\to 0^+} \alpha(r) = 0, \qquad \liminf_{r\to 0^+} \frac{\alpha(r)}{r} > 0.$$

Additional assumptions concerning the recession function of θ (the same as in [4]) are needed for ensuring convergence when (VI) is coercive. Concrete examples concerning the functions θ (the same as in [4]) are given that will convince the reader of the breath of the framework. Primal convergence is proved for interior methods ($\eta = 0$) under a weaker assumption than in [4]: weak coercivity instead of coercivity.

Finally, section 4 concerns the convergence of dual paths corresponding to primal paths. In contrast to optimization problems, a complete and useful duality theory is not available for variational inequalities. Nevertheless, a dual variational inequality (VID) is associated with the primal variational inequality (VI). Roughly speaking, the dual variables correspond to Kuhn–Tucker multipliers. The properties relating (VI) and (VID) are outlined in [3], [5].

In this last section, these relations are improved and new results obtained. For example, we prove in Theorem 4.3 that under reasonable conditions (VID) is related to a maximal monotone operator which allows us to exploit the rich structure of this type of operator. In particular, it is possible to prove the convergence of the whole dual path to a single point for some methods.

2. Coercive and weakly coercive variational inequalities. For optimization problems, the fact that the optimal solution set S is a nonempty compact set can be written in terms of recession functions. Recall that for a set $\mathcal{Q} \subset \mathbb{R}^N$, its asymptotic or recession cone is denoted by $\mathcal{Q}_\infty$ or $0^+\mathcal{Q}$ and is defined by

$$\mathcal{Q}_\infty = \Big\{ y : \exists t_k \to +\infty, x_k \in \mathcal{Q}, \text{ with } y = \lim_{k\to\infty} \frac{x_k}{t_k} \Big\}.$$

Recall also that for a closed and proper function $f : \mathbb{R}^N \to \mathbb{R} \cup +\infty$, the recession function f_∞ of f is defined by

$$epi\ (f_\infty) = (epif)_\infty, \text{ where } epif = \big\{ (x,r) \in \mathbb{R}^N \times \mathbb{R} : f(x) \le r \big\}.$$

As a straightforward consequence of this definition we obtain

$$f_\infty(y) = \inf \Big\{ \liminf_{k\to\infty} \frac{f(t_k x_k)}{t_k} \Big| \ t_k \to +\infty, x_k \to y \Big\}, \tag{2.1}$$

where $\{t_k\}$ and $\{x_k\}$ are sequences in $\mathbb{R}$ and $\mathbb{R}^N$, respectively.

For closed proper convex functions, we also have

$$f_\infty(y) = \lim_{t\to+\infty} \frac{f(x+ty) - f(x)}{t} \quad \forall x \in \mathrm{dom} f,$$

and

$$f_\infty(y) = \sup\ \{ \langle y, d\rangle \mid d \in \partial f(x), x \in \ \mathrm{dom}\partial f \}. \tag{2.2}$$

Then, when $A = \partial f_0$, it is well known that the assumption "S is nonempty and compact" is equivalent to

$$(f_0)_\infty(d) > 0 \qquad \forall d \in C_\infty,\ d \ne 0. \tag{2.3}$$

Recall also that when C is defined by

$$C = \{x : f_i(x) \leq 0, \qquad i = 1, 2, ..., m\}, \tag{2.4}$$

where $f_i : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ are closed, proper, and convex functions, then C_∞ is given by

$$C_\infty = \{d : \ (f_i)_\infty(d) \leq 0, \qquad i = 1, ..., m\}. \tag{2.5}$$

Such an equivalence is very advantageous and has been largely used for proving convergence of algorithms in [4]. Can we extend such a characterization to variational inequalities? To answer this, we have to refer to the recession analysis for general maximal monotone operators.

The first attempts to define a notion of recession function for monotone operators, well suited to characterize existence of solutions of (VI), can be found in the seminal work of Brezis and Nirenberg [7], Lions [13], and Browder [8]. More recently, Attouch, Chbani, and Moudafi [1] have compared the various definitions and, for different reasons, have concluded that the best choice is the support function of the range of the operator. For the rest of this paper we shall denote, as in [1],

$$f_\infty^B \ (d) = \sup \{\langle c, d\rangle \mid c \in R(B)\}, \tag{2.6}$$

where B is a multivalued map from $\mathbb{R}^N$ into $\mathbb{R}^N$ and $R(B)$ is the range of the map, and, as in [1], we shall call this function the recession function of B. When B is a maximal monotone map it is worthwhile to note that it was proven by Lions [13] that the subdifferential of this recession function is a maximal monotone operator B_∞ which is the graphical limit of $B(t.)$ when $t \to +\infty$. Furthermore, thanks to Theorem 13.3 in [18], this definition coincides with the usual notion in the case of a convex function.

In order to illustrate the use of the recession function of an operator, consider the generalized equation

$$0 \in B(x), \tag{2.7}$$

where B is a maximal monotone operator, and let $T = B^{-1}(0)$ denote its solution set. Recall that $R(B) = \operatorname{dom} B^{-1}$ and that T is a nonempty compact set if and only if $0 \in \operatorname{int} R(B)$. This has been proved in [19, Thm. 12.35] and [14] but appears as an immediate consequence of the formula

$$B^{-1} \ (0)_\infty = N_{R(B)} \ (0)$$

given in [2, formula (2.2)] ($N_{R(B)}(0)$ denotes the usual normal cone of $R(B)$ at 0).

Since by Minty's theorem [16] int $R(B)$ is convex, it follows from [18, Thm. 13.1] that T is a nonempty compact set if and only if

$$f_\infty^B \ (d) > 0 \qquad \forall d \neq 0. \tag{2.8}$$

In such a case we shall say that B is coercive.

Let us now return to our initial variational inequality and let $N_C(x)$ denote the usual normal cone of C at x. Recall that if $\delta(. \mid C)$ is the indicator function of C, i.e.,

$$\delta(x \mid C) = 0 \text{ if } x \in C, \quad +\infty \text{ else},$$

then

$$N_C(x) = \partial\delta(x \mid C),$$

where ∂ is the subdifferential operator.

From the above formula it follows that N_C is a maximal monotone operator. Furthermore, it is well known that the set of solutions S of (VI) is given by

$$S = (A + N_C)^{-1}\ (0). \tag{2.9}$$

So, if we want to use criteria such as (2.8) we are led to suppose that $A + N_C$ is maximal monotone.

If E and F are maximal monotone operators, defined on $\mathbb{R}^N$, there exist many criteria which ensure that $E + F$ is maximal monotone (see, in particular, [19] and the references therein).
Special cases of interest are the following:

i) int dom $E \cap$ dom $F \neq \phi$,

ii) ri dom $E \cap$ ri dom $F \neq \phi$ (ri dom E denotes the relative interior of dom E and is nonempty and convex thanks to Minty's theorem).

iii) F is a single valued map defined on C, monotone and continuous on C and $E = N_C$.

If we consider criteria (2.3) we see that it is equivalent to saying that

$$[f_0 + \delta(|C)]_\infty(d) > 0 \quad \forall d \neq 0,$$

since we have

$$[f_0 + \delta(|C)]_\infty(d) = (f_0)_\infty(d) + \delta(d \mid C_\infty).$$

Unfortunately, such a formula cannot be extended to maximal monotone operators, as can be seen in the following counterexample given by Brezis and Haraux [6]: A is the rotation with angle $\frac{\pi}{2}$ in $\mathbb{R}^2$ and $C = \mathbb{R} \times 0$. In this case if we set $B = A + N_C$, then

$$f^B_\infty\ (d_1, d_2) = 0 \quad \text{if } d_2 = 0,\ +\infty \text{ else and } f^A_\infty\ (d) + f^{N_C}_\infty\ (d) = +\infty.$$

Thus, even when A and $A + N_C$ are maximal monotone, we cannot expect, in general, the formula

$$f^B_\infty(d) = f^A_\infty(d) + f^{N_C}_\infty\ (d) = f^A_\infty\ (d) + \delta(d|C_\infty)$$

with $B = A + N_C$.

Such a formula is valid only under strict conditions such as those given in [6, Thm. 3]. By avoiding such restrictive assumptions, we give, in the next proposition, a formula that is useful for proving convergence of algorithms and that is extensively used in section 3.

PROPOSITION 2.1. *Let $B = A + N_C$ and define $f^{A,C}_\infty$ by*

$$f^{A,C}_\infty(d) = \sup\big\{\langle c, d\rangle | c \in A(x), x \in C\big\} \textit{ if } d \in C_\infty, +\infty \textit{ else.} \tag{2.10}$$

Suppose that dom $A \supset C$ *and that C is a nonempty closed convex set in $\mathbb{R}^N$. Then*

$$f^B_\infty = f^{A,C}_\infty. \tag{2.11}$$

Furthermore, if $A + N_C$ is maximal monotone, then S is a nonempty compact set if and only if

$$f^{A,C}_\infty(d) > 0 \quad \forall d \neq 0. \tag{2.12}$$

Proof. Thanks to the above discussion, we have only to prove formula (2.11). Recall that the Fenchel conjugate $\delta^*(|C)$ of $\delta(|C)$ is the support functional of C and let us denote by $ba(C)$ the barrier cone of C. By definition $ba(C) = \text{dom } \delta^*(|C)$.

1) Let $d \notin C_\infty$. Then, since $\overline{ba(C)}$ is the polar cone of the recession cone [18, Cor. 14.2.1], there exist $\tilde{u} \in \overline{ba(C)}$ such that $\langle \tilde{u}, d\rangle > 0$.

Let us prove that there exists (x, u) such that

$$x \in C, u \in N_C(x) \text{ and } \langle u, d\rangle > 0. \tag{2.13}$$

Since $\tilde{u} \in \overline{ba(C)}$ we have $\tilde{u} = \lim_{n\to\infty} u_n$ with $u_n \in \text{ri } ba(C)$. Then since $u_n \in \text{ri } ba(C)$, $\delta^*(|C)$ is subdifferentiable at u_n and $v_n \in \partial\delta^*(u_n|C)$ if and only if $v_n \in C$ with $u_n \in N_C(v_n)$. As a consequence (2.13) is satisfied for $(u, x) = (u_n, v_n)$ and n sufficiently large.

Now let u be such that (2.13) holds. Since $\lambda u \in N_C(x)$ for all $\lambda > 0$, it follows that $f_\infty^B(d) = +\infty$ and then (2.11) holds for each $d \notin C_\infty$.

2) Now let $d \in C_\infty$. Since $\overline{ba(C)}$ is the polar cone for C_∞ and since $R[N_C] \subset ba(C)$, we have

$$\langle u, d\rangle \le 0 \qquad \forall u \in R[N_C]$$

and then $f_\infty^B(d) \le f_\infty^{A,C}(d)$. Furthermore, let $x_n \in C, c_n \in A(x_n)$ such that $f_\infty^{A,C}(d) = \lim_{n\to\infty} \langle c_n, d\rangle$. Then since $u_n = 0 \in N_C(x_n)$, we have $\langle c_n + u_n, d\rangle = \langle c_n, d\rangle \le f_\infty^B(d)$ and then $f_\infty^{A,C}(d) \le f_\infty^B(d)$. □

Now we want to weaken conditions (2.8) and (2.12) in order to consider variational inequalities with noncompact solution sets. For optimization problems, condition (2.8) corresponds to the assumption "$0 \in \text{int dom } f_0^*$," where f_0^* is the usual Fenchel conjugate, and it is well known that the assumption $0 \in \text{ri dom } f^*$ weakens the above and plays a fundamental role in convex analysis. Since $\text{ri dom } f_0^* = \text{ri dom } \partial f_0^*$, the natural extension will be

$$0 \in \text{r}i\ R(B). \tag{2.14}$$

In this case we shall say that B is weakly coercive.

From [18, Thm. 13.1], this is equivalent to saying that

$$f_\infty^B(d) \ge 0 \quad \forall d \text{ and } L_B := \left\{d : f_\infty^B(d) = 0\right\} \quad \text{is } a \textit{ linear space} \tag{2.15}$$

or equivalently

$$f_\infty^B(d) > 0 \quad \forall d \in L_B^\perp,\ d \neq 0 \tag{2.16}$$

(with L_B supposedly a linear space and $L_B^\perp$ the orthogonal of L_B).

Example. Suppose that B is coercive and let F be a matrix (m, N) with $m > N$ such that rank $F^t = N$. Then, by [19, Thm. 12.38], the operator $D = F \circ B \circ F^t$ is maximal monotone. Since B is coercive and F is of full rank, it follows in a straightforward way that D is weakly coercive and that

$$L_D = \text{k}er\ F^t. \tag{2.17}$$

Let us now characterize weakly coercive maximal monotone operators B via the structure of the set of solutions T of the generalized equation (2.7).

For this, as in [2, p. 167] we denote by E_t the affine hull of dom B^{-1}, E the parallel subspace to E_t, and Π_E the projector operator on E. Since $E_t + E^\perp = \mathbb{R}^N$, every $x \in \mathbb{R}^N$ can be uniquely decomposed as follows:

$$x = \Pi_{E_t} x + x_{E^\perp} \text{ with } x_{E^\perp} \in E^\perp.$$

Following [2, p. 168], we define the extension operator B_E^{-1} of B^{-1} by

$$B_E^{-1}(x) = B^{-1}\ (\Pi_{E_t} x) \cap E \quad \forall x \in \mathbb{R}^N. \tag{2.18}$$

It has been proved in [2, Prop. 2.4 and Cor. 2.5] that B_E^{-1} is a maximal monotone operator and that

$$B^{-1}(x) = B_E^{-1}(x) + E^\perp \text{ i}f\ x \in E_t, = \phi \text{ e}lse, \tag{2.19}$$

$$\text{i}nt\ dom\ B_E^{-1} = \text{ r}i\ dom\ B^{-1} + E^\perp \neq \phi, \tag{2.20}$$

$$d(x^*|B^{-1}(0)) = d(\Pi_E x^*|B_E^{-1}(0)) \quad \forall x^*, \tag{2.21}$$

with $d(.|B^{-1}(0))$ denoting the distance to the set $B^{-1}(0)$.

Furthermore, we have the following.

PROPOSITION 2.2. *Let B be a multivalued maximal monotone map. Then* $0 \in$ ri $R(B)$ *if and only if the solution set T of* (2.7) *is given by the formula*

$$T = B_E^{-1}(0) + E^\perp, \tag{2.22}$$

where $B_E^{-1}(0)$ is a nonempty compact set.

In addition, in this case we have

$$E^\perp = L_B = \{d : \langle c, d\rangle = 0 \quad \forall c \in R(B)\}.$$

Proof. a) Suppose that $0 \in$ ri $R(B) =$ ri dom B^{-1}. Then by [2, Cor. 3.2] (2.22) holds and $B_E^{-1}(0)$ is a nonempty compact set. Conversely, suppose that (2.22) holds and that $B_E^{-1}(0)$ is a nonempty compact set. Then from [2, formula (2.2)] it follows that $0 \in$ int domB_E^{-1}, and by (2.20) we deduce that $0 \in$ ri dom B^{-1}.

b) When $0 \in$ ri dom B^{-1}, then from the definition of f_∞^B, and since L_B defined in (2.15) is a linear space, it follows that

$$L_B = R(B)^\perp = (\text{dom } B^{-1})^\perp = E^\perp. \qquad \square$$

COROLLARY 2.3. *Let S be the solution set of* (VI). *Set $B = A + N_C$. Suppose* dom $A \supset C$, *C is a nonempty closed convex set in $\mathbb{R}^N$, and B is maximal monotone. Then $S = B_E^{-1}(0) + E^\perp$, where $E = L_B^\perp$ and $B_E^{-1}(0)$ is a nonempty compact set if and only if* $0 \in$ ri $R(B)$, *or equivalently, if and only if*

$$f_\infty^{A,C}(d) > 0 \quad \forall d \in L_B^\perp \quad d \neq 0 \tag{2.23}$$

with

$$L_B = \{d \in C_\infty \cap -C_\infty : \langle c, d\rangle = 0 \quad \forall c \in A(x),\ \forall x \in C\}. \tag{2.24}$$

Proof. From the definition of L_B (see formula (2.15)) and the fact that L_B is a linear space, it follows from Proposition 2.1 that

$$L_B = \{d : f_\infty^{A,C}(d) = f_\infty^{A,C}(-d) = 0\}.$$

Using the definition of $f_\infty^{A,C}$ (see formula (2.10)) we then get formula (2.24). Then, as it was recalled above, $0 \in \text{ri }(B)$ if and only if (2.16) holds. Using formula (2.11) this is equivalent to saying that $0 \in \text{ri }(B)$ if and only if (2.23) is satisfied. □

Remark. When C is defined by (2.4), then formula (2.24) becomes

(2.25)
$$L_B = \{d : (f_i)_\infty(d) = (f_i)_\infty(-d) = 0\, \forall\, i = 1, 2 \dots m, \langle c, d\rangle = 0 \quad \forall c \in A(u), \quad \forall u \in C\}.$$

Furthermore, let us denote by $\overline{x}$ the projection of x onto $L_B^\perp$. Since $x = \overline{x} + \Pi_{L_B}(x)$, and since for each $d \in C_\infty \cap -C_\infty$ we have $f_i(\overline{x} + d) = f_i(\overline{x})$, it follows that

$$f_i(x) = f_i(\overline{x}) \quad \forall i = 1, 2, \dots, m, \quad \langle c, x\rangle = \langle c, \overline{x}\rangle \quad \forall c \in A(u) \quad \forall u \in C. \tag{2.26}$$

This formula will be used for proving the convergence of the forthcoming algorithms.

3. Penalty and barrier methods for variational inequalities.

In this section we suppose that A is a maximal monotone operator with int dom $A \supset C$. This implies that [int dom A] $\cap$ dom N_C is nonempty and then, as it was recalled in section 2, it follows that $A + N_C$ is maximal monotone. Furthermore, suppose that C is now defined by

$$C = \bigl\{x \in \mathbb{R}^N : f_i(x) \le 0, \quad i = 1, \dots, m\bigr\},$$

where $f_i : \mathbb{R}^N \to \mathbb{R} \cup +\infty$ are closed proper convex functions. Recall that in this case C_∞ is given by (2.5).

In this section we propose to extend the methods introduced in [4] for optimization problems to variational inequalities. Therefore, and in order to solve (VI), we introduce a new class of penalty and barrier methods which consists of solving a family of unconstrained generalized equations of the following form.

$(\text{VI})_r$ find x_r satisfying

$$0 \in A(x_r) + F_r(x_r) \tag{3.1}$$

with

(3.2)
$$F_r(x) = \alpha(r)\partial g_r(x), \quad g_r(x) = \sum_{i=1}^m \theta\left(\frac{f_i(x)}{r}\right), \text{ if } x \in \bigcap_{i=1}^m \text{dom } f_i \;, \; +\infty \text{ otherwise,}$$

where $r > 0$ is a penalty parameter which will ultimately go to 0. The functions $\theta : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ and $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ are such that

$$(H_0) \qquad \begin{cases} \theta \text{ is closed proper, convex, and nondecreasing} \\ \text{with dom } \theta =]-\infty, \eta[\text{ and } 0 \le \eta \le +\infty, \\ \lim_{u\to\eta^-} \theta(u) = +\infty, \quad \theta_\infty(1) > 0, \\ \lim_{r\to 0^+} \alpha(r) = 0, \quad \liminf_{r\to 0^+} \alpha(r)/r > 0. \end{cases}$$

For the case $\eta = 0$, we assume Slater's condition

$$\text{“There exists } x_0 \in \bigcap_{i=1}^m \text{ri dom } f_i \text{ such that } f_i(x_0) < 0 \text{ for each } i\text{”}$$

and then it follows by [18, Thms. 6.5 and 7.6] that

$$\text{ri C} = \left\{ x \in \bigcap_{i=1}^{m} \text{ri dom } f_i : f_i(x) < 0, \quad \forall \text{i} = 1, \ldots, m \right\}$$

and every solution x_r of $(\text{VI})_r$ satisfies $f_i(x_r) < 0$ for each i.

For the case $\eta > 0$ we suppose that dom $A = \mathbb{R}^N$.

With these assumptions g_r is a closed convex proper function and dom $g_r \cap C$ is nonempty. Then dom $g_r \cap$ int dom A is nonempty and since ri dom $g_r =$ ri dom ∂g_r, it follows that ri dom $\partial g_r \cap$ int dom A is nonempty. As a consequence, as it was recalled in section 2, $A + F_r$ is maximal monotone.

We shall study two classes of methods. For the first class we suppose

$$(H_1) \qquad \begin{cases} \lim_{t \to -\infty} \theta(t) = 0, \\ \theta_\infty(1) < +\infty, \\ \lim_{r \to 0^+} \alpha(r)/r = +\infty, \end{cases}$$

while for the second class of methods we suppose

$$(H_2) \qquad \begin{cases} \theta_\infty(-1) = 0, \\ \theta_\infty(1) = +\infty, \\ \alpha(r) = r. \end{cases}$$

In [4] a systematic way to generate functions satisfying (H_1) and (H_2) is given. In fact, for functions θ satisfying (H_1) the method of generation was given by Chen and Mangasarian [9] and consists of integrating twice a function p which is a piecewise continuous probability density function with a finite number of pieces and with supp $(p) = \mathbb{R}$. In order to obtain functions satisfying (H_2), one can start from functions of class C^2 satisfying (H_1) and take a primitive of such functions as indicated in [4], but there are also many other ways to do it as described in [4].

Functions satisfying (H_2) can be divided into two subclasses: those for which $\eta > 0$ as penalty-barrier functions used in augmented Lagrangian methods, and those for which $\eta = 0$, which may be designated as *interior* methods.

Specific cases of interest for $\eta > 0$ are

$\theta_1(u) \quad = \exp\,(u)$ (exponential penalty),

$\theta_2(u) \quad = -ln(1-u)$ for $u < 1$ (modified barrier),

$$\theta_3(u) \quad = \begin{cases} u + \frac{1}{2}u^2 & \text{if } u \geq -\frac{1}{2} \\ -\frac{1}{4}\, ln(-2u) - \frac{3}{8} & \text{if } u \leq -\frac{1}{2} \end{cases} \quad \text{(quadratic logarithmic method).}$$

In this last method the second derivative θ_3'' is continuous and bounded, which is advantageous for algorithms based on Newton steps. Another example in the same class is given by

$$\theta_4(u) = u/(1-u) \quad \text{for } u < 1 \quad \text{(hyperbolic MBF (modified barrier formula)).}$$

The idea of combining barrier methods and quadratic penalty methods as in θ_6 can be extended to other types of barrier functions, for example:

$$\theta_5(u) \quad = \begin{cases} u/(1-u) & \text{if } u \leq -\eta, \\ au^2 + bu + c & \text{otherwise,} \end{cases}$$

with $a > 0, b, c$ and $\eta > -1$ conveniently chosen so that θ_5 is of class C^2.

For $\eta = 0$ specific cases of interest are

$$\begin{aligned} \theta_6(u) &= -ln(-u) \text{ (log barrier)}, \\ \theta_7(u) &= -(1/u) \text{ (inverse barrier method)}, \\ \theta_8(u) &= \begin{cases} -ln(-u) & \text{if } \delta \le u < 0, \\ a + b/u^2 - c/u & \text{if } u \le \delta, \end{cases} \end{aligned}$$

where $\delta < 0$, and the parameters $a, b > 0$, and $c > 0$ are chosen so that θ_8 is twice differentiable.

THEOREM 3.1. *Let $B = A + N_C$; suppose that the functions f_i are closed proper convex functions defined on $\mathbb{R}^N$, that A is maximal monotone with* int dom $A \supset C$, *and that* $0 \in$ ri $R(B)$. *Consider the case where θ and α satisfy* (H_0) *and* (H_2). *Suppose in addition that*

1) *when $\eta = 0, A$ is single valued on C and Slater's condition holds;*

2) *when $\eta > 0$,* dom $A = \mathbb{R}^N$ *and* int $R(B)$ *is nonempty.*

Then for each $r > 0$, the solution set S_r of $(\mathrm{VI})_r$ *defined by* (3.1) *is nonempty.*

Proof. Consider for r, n fixed the variational inequality

$(\mathrm{VI})_{r,n}$ find $x_n \in B_n := \{x : \|x\| \le n\}$ such that there exists $d_n \in (A + F_r)(x_n)$ with

$$\langle d_n, x - x_n\rangle \ge 0 \quad \forall x \in B_n.$$

Since $A + F_r$ is maximal monotone and since B_n is convex compact, then for n sufficiently large dom $(A + F_r) \cap$ int B_n is nonempty and such an x_n exists. This is a classical result (see, for example, [14, Cor. 4.(b)]). Then there exists $c(x_n) \in A(x_n), e(x_n) \in \partial g_r(x_n)$ such that

$$\langle c(x_n), x - x_n\rangle + r\langle e(x_n), x - x_n\rangle \ \ge 0 \quad \forall x \in B_n.$$

Furthermore, since g_r is convex we get

$$\langle c(x_n), x - x_n\rangle + rg_r(x) \ge rg_r(x_n) \quad \forall x \in B_n, \tag{3.3}$$

and from this inequality, since A is monotone it follows that

$$\langle c, x\rangle + rg_r(x) \ge \langle c, x_n\rangle + rg_r(x_n) \quad \forall x \in B_n \cap \text{ dom } A, \quad \forall c \in A(x). \tag{3.4}$$

a) Let $\overline{x}_n$ be the projection of x_n on $L_B^{\perp}$; then from (2.26) and (3.4) it follows that

$$\langle c, x\rangle + rg_r(x) \ge \langle c, \overline{x}_n\rangle + rg_r(\overline{x}_n) \quad \forall x \in B_n \cap \ C, c \in A(x). \tag{3.5}$$

Let us prove now that the sequence $\{\overline{x}_n\}$ is bounded. In the contrary case, we can suppose, without loss of generality, that

$$\lim_{n\to\infty} \|\overline{x}_n\| = +\infty, \ \lim \frac{\overline{x}_n}{\|\overline{x}_n\|} = d \text{ with } d \in L_B^{\perp}, d \ne 0.$$

Let $x \in C$ if $\eta > 0$ and $x \in$ ri C if $\eta = 0$. In both cases $g_r(x) < +\infty$, and then dividing both members of (3.5) by $\|\overline{x}_n\|$, it follows from formula (2.1), if we pass to the limit, that

$$\langle c, d\rangle + r(g_r)_\infty(d) \le 0 \quad \forall c \in A(x).$$

Set

$$h_i(x) = r\theta\left(\frac{f_i(x)}{r}\right) \text{ if } x \in \text{ dom } f_i, = +\infty \text{ otherwise.}$$

Then by [4, Prop. 2.1], h_i is a closed proper convex function and we have

$$(h_i)_\infty(d) = \theta_\infty((f_i)_\infty(d)) \text{ if } d \in \text{ dom } (f_i)_\infty, = +\infty \text{ otherwise}$$

and since $\theta_\infty(1) = +\infty, \theta_\infty(-1) = 0$, it follows from the definition of g_r that

$$(3.6) \qquad (f_i)_\infty(d) \leq 0 \quad \forall i = 1, 2, ..., m, \langle c, d\rangle \leq 0 \quad \forall c \in A(x).$$

Furthermore, if $\eta > 0$, since (3.6) holds for all $x \in C$, it follows that $f_\infty^{A,C}(d) \leq 0$, a contradiction to (2.23). If $\eta = 0$, for each $y \in C$ consider $x_\lambda = \lambda x_0 + (1-\lambda)y$ with $\lambda \in]0, 1]$ and $x_0 \in$ ri C. It follows that $x_\lambda \in$ ri C, and by (3.6) we have $\langle A(x_\lambda), d\rangle \leq 0$. Since A is single valued on C and int dom $A \supset C$, A is locally bounded and closed on C. As a consequence A is continuous on C and, passing to the limit when $\lambda \to 0^+$, we obtain

$$\langle A(y), d\rangle \leq 0 \quad \forall y \in C$$

and again $f_\infty^{A,C}(d) \leq 0$, a contradiction to (2.23).

b) Now let $\overline{x}$ be a limit point of the sequence $\{\overline{x}_n\}$. Such a point exists and, without loss of generality, we can suppose that $\overline{x} = \lim_{n\to+\infty}\overline{x}_n$.

b$_1$) Consider the case where $\eta = 0$. Then, A is single valued and continuous on C. Since in this case $\overline{x} \in C$, A is continuous at $\overline{x}$. Take in formula (3.5) $x \in$ ri C. Then passing to the limit in this formula, we obtain

$$\langle A(x), x - \overline{x}\rangle + rg_r(x) - rg_r(\overline{x}) \geq 0.$$

Let $x_t = \overline{x} + t(x - \overline{x})$ with $t \in]0, 1]$. Then from the above formula it follows that

$$\langle A(x_t), x - \overline{x}\rangle + \left[\frac{rg_r(x_t) - rg_r(\overline{x})}{t}\right] \geq 0.$$

Passing to the limit when $t \to 0^+$ in this inequality, we obtain

$$\langle A(\overline{x}), x - \overline{x}\rangle + rg_r'(\overline{x}; x - \overline{x}) \geq 0,$$

and since g_r is convex it follows that

$$(3.7) \qquad \langle A(\overline{x}), x\rangle + rg_r(x) \geq \langle A(\overline{x}), \overline{x}\rangle + rg_r(\overline{x}).$$

Let $y \in C$ and set $y_\lambda = (1-\lambda)y + \lambda x_0$ with $\lambda \in]0, 1]$ and $x_0 \in$ ri C. Then $y_\lambda \in$ ri C, and from (3.7) we have

$$\langle A(\overline{x}), y_\lambda\rangle + rg_r(y_\lambda) \geq \langle A(\overline{x}), \overline{x}\rangle + rg_r(\overline{x}).$$

Since $g_r(y) = \lim_{\lambda\to 0} g_r(y_\lambda)$, passing to the limit it follows that (3.7) holds for each $x \in C$ and which is, by the way, valid for all $x \in \mathbb{R}^N$. As a consequence $\overline{x}$ minimizes on $\mathbb{R}^N$ the function $x \to h(x) = \langle A(\overline{x}), x\rangle + rg_r(x)$. Then $0 \in \partial h(\overline{x}) = A(\overline{x}) + r\partial g_r(\overline{x})$ and S_r is nonempty.

b$_2$) Suppose now that $\eta > 0$. Then int $R(B)$ is nonempty, and $L_B^\perp = \mathbb{R}^N$ so that $x_n = \overline{x}_n$. Since A is locally bounded and closed at $\overline{x}$, passing to the limit in (3.3) we get

$$\langle \overline{c}, x - \overline{x} \rangle + r g_r(x) - r g_r(\overline{x}) \geq 0 \quad \forall x$$

for some $\overline{c} \in A(\overline{x})$. Then $\overline{x}$ minimizes the function $x \to \langle \overline{c}, x \rangle + r g_r(x)$ and, as above, it follows that S_r is nonempty. $\square$

Theorem 3.2. *Suppose the assumptions of Theorem* 3.1 *hold. For each solution* $x_r \in S_r$, *let* $x_r^1 = \sqcap_E(x_r)$ *be the projection of* x_r *on* $E = L_B^\perp$ *(the linear space generated by* $R(B)$*). Then the sequence* $\{x_r^1\}$ *stays bounded when* $r \to 0+$ *with all its limit points in* S *and*

$$\lim_{r \to 0+} d(x_r | S) = 0.$$

Furthermore, if $\eta > 0$, *the same result holds with* x_r *instead of* x_r^1.

Proof. Let $x_r \in S_r, x_r^1 = \Pi_E(x_r)$ with $E = L_B^\perp$. Since A is monotone and g_r convex, it follows from (3.1) that there exists $c(x_r) \in A(x_r)$ such that

$$\langle c(x_r), x - x_r \rangle + r g_r(x) \geq r g_r(x_r) \quad \forall x \tag{3.8}$$

and that

$$\langle c, x \rangle + r g_r(x) \geq \langle c, x_r \rangle + r g_r(x_r) \quad \forall c \in A(x), x \in \text{dom } A \tag{3.9}$$

a) From formulas (3.9) and (2.26) it follows that

$$\langle c, x \rangle + r g_r(x) \geq \langle c, x_r^1 \rangle + r g_r(x_r^1) \quad \forall x \in C, \quad \forall c \in A(x). \tag{3.10}$$

Let us prove that each selection $\{x_r^1\}$ with $r \to 0^+$ is bounded. In the contrary case, there would exist a sequence $r_k \to 0^+$ and an associated sequence $\{\overline{x}_k = \frac{x_{r_k}^1}{\|x_{r_k}^1\|}\}$ such that

$$\|x_{r_k}^1\| \to +\infty, \overline{x}_k \to d \text{ with } d \neq 0 \text{ and } d \in L_B^\perp.$$

Let $x \in C$ when $\eta > 0$ and $x \in$ ri C when $\eta = 0$. Let $a(x) = \max \ \{f_i(x) \mid i = 1, ..., m\}$.
Since θ is nondecreasing, it follows from (3.10) that

$$\langle c, x \rangle + m \ r_k \ \theta\left(\frac{a(x)}{r_k}\right) \geq \langle c, x_{r_k}^1 \rangle + r_k \sum_{i=1}^m \theta\left(\frac{f_i(x_{r_k}^1)}{r_k}\right). \tag{3.11}$$

Let $\varepsilon_i < (f_i)_\infty(d)$ for $i = 1, 2, ..., m$. Then from formula (2.1) there exists k_0 such that $f_i(x_{r_k}^1) \geq \varepsilon_i \ \|x_{r_k}^1\|$ for all $k \geq k_0$. Again since θ is nondecreasing, we deduce from (3.11)

$$\frac{\langle c, x \rangle}{\|x_{r_k}^1\|} + \frac{m \ r_k}{\|x_{r_k}^1\|} \ \theta\left(\frac{a(x)}{r_k}\right) \geq \langle c, \overline{x}_k \rangle + \sum_{i=1}^m \theta\left(\varepsilon_i \frac{\|x_{r_k}^1\|}{r_k}\right) \ \frac{r_k}{\|x_{r_k}^1\|}.$$

Since $a(x) \leq 0, \theta_\infty(a(x)) = 0$. Then passing to the limit in the above inequality, we obtain

$$\langle c, d \rangle + \sum_{i=1}^m \theta_\infty(\varepsilon_i) \leq 0. \tag{3.12}$$

Since $\theta_\infty(1) = +\infty$, it follows that $\varepsilon_i \leq 0$ so that $(f_i)_\infty(d) \leq 0$, for each i, and $d \in C_\infty$. Furthermore, $\theta_\infty(\varepsilon_i) = -\varepsilon_i \theta_\infty(-1) = 0$, and from (3.12) it follows that

$$\langle c, d\rangle \leq 0 \quad \forall c \in A(x) \text{ with } x \in C \text{ if } \eta > 0 \text{ and } x \in \text{ ri } C \text{ else.}$$

Then as in part a) of the proof of Theorem 3.1, it follows that in both cases $f_\infty^{A,C}(d) \leq 0$ which is a contradiction to (2.23).

b) Now let $\overline{x} = \lim\ x_{r_k}^1$ be a limit point of the sequence $\{x_r^1\}$ when $r \to 0^+$. Such a point exists. Take in formula (3.10) $x \in$ ri C if $\eta = 0$, and $x \in C$ if $\eta > 0$. Take also $\delta_i < f_i(\overline{x})$ for each i. Since f_i is closed, then $\delta_i < f_i(x_{r_k}^1)$ for k sufficiently large and from (3.10) we obtain

$$(3.13) \qquad \langle c, x\rangle + m\ r_k\ \theta\left(\frac{a(x)}{r_k}\right) \geq \langle c, x_{r_k}^1\rangle + r_k \sum_{i=1}^{m} \theta\left(\frac{\delta_i}{r_k}\right) \quad \forall c \in A(x)$$

with $a(x) = \max\ \{f_i(x) \mid i = 1, ..., m\}$.

b$_1$) Suppose that $\eta = 0$. Then obviously $\overline{x} \in C$. Furthermore in this case A is single valued, $\delta_i < 0, a(x) < 0$. Since $\theta_\infty(-1) = 0$, passing to the limit in 3.13 it follows that

$$\langle A(x), x - \overline{x}\rangle \geq 0 \quad \forall x \in riC.$$

Then taking the same arguments as in part b$_1$) of the proof of Theorem 3.1, it follows that

$$\langle A(\overline{x}), x - \overline{x}\rangle \geq 0\ \forall x \in C$$

and $\overline{x}$ is a solution of (VI). As a consequence, $d(x_r^1 \mid S) \to 0$ if $r \to 0$ and from (2.21) we have $\lim_{r\to 0} d(x_r \mid S) = 0$.

b$_2$) Suppose now that $\eta > 0$. Since $a(x) \leq 0, \theta_\infty(-1) = 0, \theta_\infty(1) = \infty$; if we pass to the limit in (3.13) it follows that $\delta_i \leq 0$ for each i so that $f_i(\overline{x}) \leq 0$ and $\overline{x} \in C$. Furthermore, in this case $x_r^1 = x_r$ and from (3.8) we get

$$(3.14) \qquad \langle c(x_{r_k}), x - x_{r_k}\rangle + m\ r_k\ \theta\ \left(\frac{a(x)}{r_k}\right) \geq r_k \sum_{i=1}^{m} \theta\left(\frac{\delta_i}{r_k}\right).$$

Since A is locally bounded and closed at $\overline{x}$, taking the same arguments as above, we obtain by passing to the limit in (3.14)

$$(\overline{c}, x - \overline{x}) \geq 0 \quad \forall x \in C$$

for some $\overline{c} \in A(\overline{x})$. As a consequence $\overline{x} \in S$ and $\lim_{r\to 0} d(x_r \mid S) = 0$. □

Remark 3.1. When θ and α satisfy (H_0) and (H_1), similar theorems such as Theorems 3.1 and 3.2 can be obtained using the same kinds of arguments. The only difference is that S_r is nonempty only for r sufficiently small. The proofs are left to the reader.

Remark 3.2. When $C = \mathbb{R}_+^N$ (complementarity problems), for $\theta = \theta_6$ the method coincides with the classical logarithm interior point method studied by Güler [12]. The existence theorem and the proof of primal convergence given in [12] suppose a stronger condition (condition 1.2 in [12]) than ours. Moreover, the use of formula (2.11) simplifies the proofs considerably, since we don't need the Brezis–Haraux's theorem as in [12], for which it is difficult to apply.

4. Dual convergence. The setting in this section is the same as in section 3.

4.1. The dual variational inequality. For each $\lambda \in \mathbb{R}^m_+$ we define $M(\lambda)$ as the set of vectors $x^* \in C$ for which there exist $c_0 \in A(x^*)$, $c_i \in \partial f_i(x^*)$, $i \in I(x^*)$ such that

$$c_0 + \sum_{i \in I(x^*)} \lambda_i c_i = 0, \quad \lambda_i = 0 \text{ for } i \notin I(x^*), \tag{4.1}$$

where $I(x^*) = \{i : f_i(x^*) = 0\}$.

We denote by M the set of all vectors $\lambda \in \mathbb{R}^m_+$ such that $M(\lambda)$ is nonempty and the dual problem to (VI) is defined as follows:

$$\text{Find a point } \lambda \in M. \tag{VID}$$

The next proposition in particular will help us to justify the denomination of "dual problem."

PROPOSITION 4.1. *Suppose that the functions f_i are closed proper convex functions defined on $\mathbb{R}^N$, that Slater's condition holds, and that $x^* \in C$. Then $x^* \in S$ if and only if there exists $\lambda \in \mathbb{R}^m_+$ such that $x^* \in M(\lambda)$. Furthermore, if S is a nonempty compact set, if each f_i is continuous on S and if A is maximal monotone with* int dom $A \supset S$, *then M is nonempty and compact.*

Proof. i) Let $x^* \in C$ with $\lambda \in \mathbb{R}^m_+$ such that $x^* \in M(\lambda)$. Take c_0 and c_i as in (4.1); then for each $x \in C$ it follows from (4.1) that

$$\langle c_0, x - x^* \rangle \geq \langle c_0, x - x^* \rangle + \sum_{i \in I(x^*)} \lambda_i \left[f_i(x) - f_i(x^*) \right] \geq \left\langle c_0 + \sum_{i \in I(x^*)} \lambda_i c_i, x - x^* \right\rangle = 0$$

and as a consequence $x^* \in S$.

ii) Conversely if $x^* \in S$, then there exists $c_0^* \in A(x^*)$ such that

$$\langle c_0^*, x - x^* \rangle \geq 0 \quad \forall x \in C.$$

Set $J(x) = \langle c_0^*, x - x^* \rangle$. From the above inequality it follows that x^* minimizes J on C, and since Slater's condition is satisfied, this implies the existence of multipliers $\lambda_i \geq 0$ for $i \in I(x^*)$ such that

$$c_0^* + \sum_{i \in I(x^*)} \lambda_i c_i^* = 0, \quad \lambda_i = 0 \text{ for } i \notin I(x^*),$$

where $c_i^* \in \partial f_i(x^*)$. In other words $x^* \in M(\lambda)$.

iii) Suppose now that the other assumptions are satisfied. As a consequence of part i) M is nonempty. Let us prove that M is bounded. In the contrary case there would exist a sequence $\{(\lambda^n, x_n, c_0^n, c_i^n)\}$ with $\lambda^n \in \mathbb{R}^m_+, x_n \in S, c_0^n \in A(x_n), c_i^n \in \partial f_i(x_n)$ for $i = 1, \ldots, m$ such that

$$\|\lambda^n\| \to +\infty, \frac{\lambda^n}{\|\lambda^n\|} \to \overline{\lambda} \neq 0, \quad x_n \to \overline{x}, \quad \lambda_i^n = 0 \text{ for } i \notin I(x_n)$$

and

$$c_0^n + \sum_{i \in I(x_n)} \lambda_i^n c_i^n = 0.$$

From the above equation it follows that

$$\frac{\langle c_0^n, x_0 - x_n \rangle}{\|\lambda^n\|} + \sum_{i \in I(x_n)} \frac{\lambda_i^n}{\|\lambda^n\|} f_i(x_0) \geq 0,$$

where x_0 satisfies Slater's condition.

Since A is locally bounded at $\overline{x} \in S$ and the functions f_i are continuous at $\overline{x}$, passing to the limit in the above inequality we obtain

$$\sum_{i \in I(\overline{x})} \overline{\lambda}_i f_i(x_0) \geq 0, \quad \overline{\lambda}_i = 0 \quad \text{for } i \notin I(\overline{x}), \overline{\lambda} \neq 0,$$

which is impossible since Slater's condition holds at x_0. □

If the variational inequality corresponds to a convex minimization problem ($A = \partial f_0$), then Proposition 4.1 tells us that M coincides with the set of Kuhn–Tucker vectors associated with each optimal solution. As a consequence, it follows from [18, Cor. 28.4.1] that M coincides with the optimal set of the usual dual problem. This justifies the denomination of dual problem for (VID).

4.2. The dual trajectory and its accumulation points.

THEOREM 4.2. *Consider the case where θ and α satisfy (H_0) and (H_2), and suppose that*

i) *θ is C^1 on $]-\infty, \eta[$.*

ii) *the functions f_i are closed proper convex functions defined on $\mathbb{R}^N$, and Slater's condition holds.*

iii) *A is maximal monotone, $C \subset$ int dom A.*

iv) *when $\eta = 0$, we assume that A is single valued on C, and when $\eta > 0$ we suppose that* dom $A = \mathbb{R}^n$.

Set $B = A + N_C$ and suppose in addition that $0 \in$ int $R(B)$. Then, all the assumptions of Theorems 3.1 and 3.2 are satisfied and to each solution $x_r \in S_r$ we can associate a "dual" vector λ^r with coordinates given by

$$\lambda_i^r = \theta' \left(\frac{f_i(x_r)}{r} \right) \quad \forall i = 1, 2, \ldots, m. \tag{4.2}$$

If we suppose also that all the functions f_i are continuous on a neighborhood V of the solution set S, then the sequence $\{x_r, \lambda^r\}$ stays bounded when $r \to 0^+$ and all its limit points $(\overline{x}, \overline{\lambda})$ are in $S \times M$ with $\overline{x} \in M(\overline{\lambda})$.

Proof. a) Obviously all the assumptions of Theorems 3.1 and 3.2 are satisfied and as a consequence, λ^r is well defined and the sequence $\{x_r\}$ is bounded with all its limit points in S. Furthermore, since θ is nondecreasing $\lambda^r \in \mathbb{R}_+^m$.

b) Set $\overline{\lambda}^r = \frac{\lambda^r}{\|\lambda^r\|}$ and suppose that the sequence $\{\lambda^r\}$ is not bounded when $r \to 0^+$. Then there exists a sequence $\{\lambda^{r_n}, x_{r_n}\}$ such that

$$r_n \to 0^+, \; x_{r_n} \to \overline{x}, \; \|\lambda^{r_n}\| \to \infty, \; \overline{\lambda}^{r_n} \to \overline{\lambda} \geq 0, \; \overline{\lambda} \neq 0.$$

From Theorem 3.2, $\overline{x} \in S$ and since the functions f_i are continuous at $\overline{x}$, it follows that $\frac{f_i(x_{r_n})}{r_n} \to -\infty$ for $i \notin I(\overline{x})$, and then from (4.2) we have

$$\overline{\lambda}_i = \lim_{n \to \infty} \overline{\lambda}_i^{r_n} = \lim_{n \to \infty} \lambda_i^{r_n} = 0 \quad \text{for } i \notin I(\overline{x}). \tag{4.3}$$

Now since the functions f_i are continuous on a neighborhood V of S, for r sufficiently small we have

$$r\partial g_r(x_r) = \sum_{i=1}^{m} \theta'\Big(\frac{f_i(x_r)}{r}\Big)\ \partial f_i(x_r)$$

and then from the definition of x_r and λ^r there exist $c_0^r \in A(x_r)$ and $c_i^r \in \partial f_i(x^r)$ such that

$$0 = c_0^r + \sum_{i=1}^{m} \lambda_i^r c_i^r. \tag{4.4}$$

As a consequence we get

$$0 = \frac{c_0^{r_n}}{\|\lambda^{r_n}\|} + \sum_{i=1}^{m} \overline{\lambda}_i^{r_n} c_i^{r_n}.$$

Then since the maps A and ∂f_i are locally bounded and closed at $\overline{x}$ (relative to C if $\eta = 0$), passing to the limit in the above equation, we obtain

$$0 = \sum_{i=1}^{m} \overline{\lambda}_i c_i$$

for some $\overline{c}_i \in \partial f_i(\overline{x}), i = 1, ..., m.$

Since $\overline{\lambda}_i = 0$ for $i \notin I(\overline{x}), \overline{\lambda} \neq 0$, and $\overline{\lambda}_i \geq 0$ for each i, obviously this yields a contradiction to Slater's condition.

c) Now let $(\overline{x}, \lambda)$ be a limit point of the sequence $\{x^r, \lambda^r\}$ when $r \to 0^+$. Since this sequence is bounded, such a point exists. By Theorem 3.2 $\overline{x} \in S$, and as in formula (4.3) we have $\lambda_i = 0$ for each $i \notin I(\overline{x})$. Since (4.4) is always valid, passing to the limit in this equation, we get

$$0 = c_0 + \sum_{i=1}^{m} c_i \lambda_i,$$

where $c_0 \in A(\overline{x})$, $c_i \in \partial f_i(\overline{x})$ (for the same reasons as in part b).

Since $\lambda \geq 0$ and $\lambda_i = 0$ for $i \notin I(\overline{x})$, this is equivalent to saying that $\lambda \in M$ and $\overline{x} \in M(\lambda)$. □

4.3. Convergence to a single point. In this section we suppose that A is maximal monotone with dom $A = \mathbb{R}^N$ and that A is strongly monotone; i.e., there exists $\alpha > 0$ such that

$$\langle c - d, x - y\rangle \geq \alpha\ \|x - y\|^2 \quad \forall c \in A(x) \quad d \in A(y) \quad \forall x, y \in \mathbb{R}^N. \tag{4.5}$$

As a consequence, $A + N_C$ is maximal monotone and S is reduced to a single point.

We suppose also that Slater's condition holds and that the functions f_i are finite on $\mathbb{R}^N$ and satisfy the following condition:

$$\limsup_{\|y\| \to \infty} \frac{|f_i(x + y) - f_i(x)|}{\|y\|^2} < +\infty. \tag{4.6}$$

This condition is not too restrictive since it is satisfied in particular by Lipschitz convex functions on $\mathbb{R}^N$, or more generally [16, Thm. 10.3] by uniformly continuous convex functions and also by quadratic functions and piecewise linear-quadratic functions. Since A is strongly monotone, for each $p \in \mathbb{R}^m_+$ the map $x \to A(x) + \sum_{i=1}^m p_i \partial f_i(x)$ is maximal monotone and strongly monotone, and it follows that the generalized equation

$$(VI)(p) \qquad 0 \in A(x) + \sum_{i=1}^{m} p_i \partial f_i(x)$$

admits one and only one solution $x(p)$.

We write

$$F(x) = (f_1(x), ..., f_m(x)), \ G(p) = -F(x(p)),$$
$$\widetilde{G}(p) = G(p) + N_{\mathbb{R}^m_+}(p) \text{ if } p \in \mathbb{R}^m_+, \ \widetilde{G}(p) = \phi \text{ else.}$$

Theorem 4.3. *Suppose that A is a maximal monotone map, strongly monotone with* dom $A = \mathbb{R}^N$. *Suppose also that Slater's condition holds and that the functions f_i are convex, finite on $\mathbb{R}^N$, and satisfy condition* (4.6). *Then $G(.)$ is monotone on $\mathbb{R}^m_+$. More precisely, we have*

$$(4.7) \qquad \langle G(p) - G(p'), \ p - p' \rangle \geq \alpha \ \|x(p) - x(p')\|^2 \quad \forall p, p' \in \mathbb{R}^m_+.$$

Furthermore, the function $p \to x(p)$ is continuous on $\mathbb{R}^m_+$ and $\widetilde{G}$ is maximal monotone and coercive. Finally, λ belongs to the dual set M of solutions (which is now a nonempty convex compact set) if and only if λ satisfies the variational inequality

$$(4.8) \qquad \lambda \in \mathbb{R}^m_+, \quad (G(\lambda), p - \lambda) \geq 0 \quad \forall p \in \mathbb{R}^m_+$$

which is equivalent to

$$(4.9) \qquad 0 \in \widetilde{G}(\lambda).$$

Proof. a) Let $p, q \in \mathbb{R}^m_+$; since the functions f_i are convex, from the definition of $x(p)$ and $x(q)$ we have

$$\big\langle c(p), x(q) - x(p) \big\rangle + \big\langle p, F\big(x(q)\big) - F\big(x(p)\big) \big\rangle \geq 0,$$
$$\big\langle c(q), x(p) - x(q) \big\rangle + \big\langle q, F\big(x(p)\big) - F\big(x(q)\big) \big\rangle \geq 0$$

with $c(p) \in A(x(p))$ and $c(q) \in A(x(q))$.

Then if we add both these inequalities we obtain

$$\langle c(p) - c(q), x(p) - x(q) \rangle + \langle p - q, F\big(x(p)\big) - F\big(x(q)\big) \rangle \leq 0$$

and (4.7) follows immediately by using (4.5) in the above inequality.

b) Let us prove now that the function $x(.)$ is continuous on $\mathbb{R}^m_+$. We first prove that it is locally bounded. In the contrary case there would exist $p \in \mathbb{R}^m_+$ and a sequence $p_n \in \mathbb{R}^m_+$ converging to p such that

$$\|x(p_n)\| \to +\infty, \quad \frac{x(p_n)}{\|x(p_n)\|} \to d \neq 0.$$

From (4.7) we obtain

$$\|p - p_n\| \quad \frac{\|F(x(p_n)) - F(x(p))\|}{\|x(p) - x(p_n)\|^2} \quad \geq \alpha > 0$$

and then since (4.6) holds, letting $n \to \infty$ we get $\alpha = 0$ which is impossible.

Now let $p_n \in \mathbb{R}^m_+$ converging to p, and let x^* be a limit point of the sequence $\{x(p_n)\}$. Since this sequence is bounded, such a point exists and, without loss of generality, we can suppose that $x^* = \lim_{n\to\infty} x(p_n)$. From the definition of $x(p_n)$ there exists $c_n \in A(x(p_n))$ such that

$$\langle c_n, x - x(p_n)\rangle + \langle p_n, F(x) - F(x(p_n))\rangle \geq 0 \quad \forall x.$$

Then, since A is locally bounded and closed and F is continuous, passing to the limit, we get for some $c \in A(x^*)$

$$\langle c, x - x^*\rangle + \langle p, F(x) - F(x^*)\rangle \geq 0 \quad \forall x \in \mathbb{R}^N$$

from which it follows that $x^* = x(p)$. As a consequence $x(.)$ is continuous on $\mathbb{R}^m_+$, and by the way, $G(.)$ is also.

Then from [19, Example 12.43] it follows that $\widetilde{G}$ is maximal monotone. Obviously (4.8) and (4.9) are equivalent, and λ satisfies (4.8) if and only if $\lambda \in M$. Then, from Proposition 4.1 and since $\widetilde{G}^{-1}(0)$ is closed and convex, it follows that M is a nonempty convex compact set. □

Let us now return to our algorithm and remark first that from the definitions of $x_r, \lambda^r, x(\lambda^r)$, we have

$$x_r = x(\lambda^r). \tag{4.10}$$

Let θ^* be the Fenchel conjugate of θ and write $t(\lambda) = \sum_{i=1}^m \theta^*(\lambda_i)$. Since $\partial\theta^* = (\theta')^{-1}$, it follows from (4.2) that $f_i(x_r) = r\partial\theta^*(\lambda^r_i)$. In other words we have

$$0 \in G(\lambda^r) + r\partial t(\lambda^r). \tag{4.11}$$

Since $0 \in N_{\mathbb{R}^m_+}(\lambda)$ for each $\lambda \in \mathbb{R}^m_+$, it follows that λ^r satisfies the generalized equation

$$0 \in \widetilde{G}(\lambda^r) + r\partial t(\lambda^r).$$

Hence λ^r can be interpreted as a result of a "viscosity" or "Tikhonov" regularization method, and the next proposition proves that the whole sequence $\{\lambda^r\}$ converges to a single point under additional assumptions on θ.

PROPOSITION 4.4. *Consider the case where θ and α satisfy (H_0) and (H_2). Suppose that the conditions of Theorem 4.3 are satisfied, and suppose that θ is C^1 on $]-\infty, \eta[$ and that either θ is bounded from below or there exists $\lambda \in M$ with $\lambda_i > 0$ for each i. Then there exists exactly one and only one $\overline{\lambda}$ which minimizes t on M and the whole sequence $\{\lambda_r\}$ converges to $\overline{\lambda}$ when $r \to 0^+$.*

Proof. With the above assumptions we remark that Theorem 4.2 is valid. Observing that dom $\theta^* = [0, +\infty[$ if θ is bounded below and dom $\theta^* =]0, +\infty[$ otherwise, it follows that dom $t \cap M$ is nonempty. Furthermore, since θ is differentiable on $]-\infty, \eta[$, it follows from [16, Thm. 26.3] that t is strictly convex on its domain. Now from (4.11) we have

$$\langle G(\lambda), \lambda - \lambda^r\rangle + r\ [t(\lambda) - t(\lambda^r)] \geq 0 \quad \forall \lambda \in M \cap \text{ dom } t.$$

Since $\langle G(\lambda), \lambda - \lambda^r\rangle \leq 0$ for each $\lambda \in M$ it follows that

$$t(\lambda) \geq t(\lambda^r) \quad \forall \lambda \in M.$$

Let $\tilde{\lambda}$ be a limit point of the sequence $\{\lambda^r\}$. From Theorem 4.2 such a point exists, and $\tilde{\lambda} \in M$. Since t is lower semicontinuous, passing to the limit it follows that $\tilde{\lambda}$ minimizes t on M. But since M is convex compact and t is strictly convex on its domain with $M \cap$ dom $t \neq \phi$, there exists only one point $\overline{\lambda}$ which minimizes t on M and $\tilde{\lambda} = \overline{\lambda}$. □

Remark. The assumption "θ is bounded from below" holds, for instance, in the case of $\theta_1, \theta_4, \theta_5, \theta_7, \theta_8$.

REFERENCES

[1] H. Attouch, Z. Chbani, and A. Moudafi, *Recession operators and solvability of variational problems in reflexive Banach spaces*, in Calculus of Variations, Homogenization and Continuum Mechanics, Ser. Adv. Math. Appl. Sci. 18, World Scientific, River Edge, NJ, 1994, pp. 51–67.

[2] A. Auslender, *Convergence of stationary sequences for variational inequalities with maximal monotone operators*, Appl. Math. Optim., 28 (1993), pp. 161–172.

[3] A. Auslender, *Optimisation*, Méthodes numériques, Masson, Paris, 1976.

[4] A. Auslender, R. Cominetti, and M. Haddou, *Asymptotic analysis for penalty and barrier methods in convex and linear programming*, Math. Oper. Res., 22 (1997), pp. 43–62.

[5] A. Bensoussan, J.L. Lions, and R. Temam, *Sur les méthodes de décomposition, de décentralisation, de coordinations et applications*, dans Sur les méthodes numériques en sciences physiques et économiques, J.L. Lions and G.I. Marchouk, eds., Dunod-Bordas, Paris, 1974, pp. 133–257.

[6] H. Brezis and A. Haraux, *Images d'une somme d'operateurs monotones et applications*, Israel J. Math, 23 (1976), pp. 165–186.

[7] H. Brezis and L. Nirenberg, *Characterizations of ranges of some nonlinear operators and applications to boundary value problems*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 5 (1978), pp. 225–236.

[8] F. Browder, *On the range of the sum of nonlinear operators and the Landesman-Lazer principle*, Bull, Un. Mat. Ital. B(5), (1979), pp. 364–376.

[9] C. Chen and O.L. Mangasarian, *A Class of Smoothing Functions for Nonlinear and Mixed Complementarity Problems*, Math Programming Technical Report, Center of Parallel Optimization, Computer Science Dept., University of Wisconsin, Madison, WI, 1994.

[10] J.P. Crouzeix, *Pseudomonotone variational inequality problems: Existence of solutions*, Math. Programming Ser. A, 78 (1997), pp. 305–314.

[11] M. Fukushima, *Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems*, Math Programming Ser. A, 53 (1992), pp. 99–110.

[12] O. Güler, *Existence of interior points and interior paths in nonlinear monotone complementarity problems*, Math. Oper. Res., 18 (1993), pp. 128–146.

[13] P.L. Lions, *Two remarks on the convergence of convex functions and monotone operators*, Nonlinear Anal., 2 (1978), pp. 553–562.

[14] L. McLinden, *Stable monotone variational inequalities,* Math. Programming Ser. B, 48 (1990), pp. 303–338.

[15] P. Marcotte and J.P. Dussault, *A sequential linear programming algorithm for solving monotone variational inequalities*, SIAM J. Control Optim., 27 (1989), pp. 1260–1278.

[16] G. Minty, *On the maximal domain of a "monotone" function*, Michigan Math. J., 8 (1961), pp. 135–137.

[17] J.S. Pang and D. Chan, *Iterative methods for variational and complementarity problems*, Math. Programming, 24 (1982), pp. 284–313.

[18] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[19] R.T. Rockafellar and R. Wets, *Variational Analysis*, Fundamental Principles of Mathematical Sciences 317, Springer-Verlag, Berlin, 1998.

SIAM J. Control Optim.
Vol. 37, No. 3, pp. 673–694

BOUNDARY STABILIZATION OF THE WAVE EQUATION IN ALMOST STAR-SHAPED DOMAINS[*]

PATRICK MARTINEZ[†]

Abstract. We improve several earlier results on the boundary stabilization of the wave equation in a bounded domain. First we weaken the usual geometrical conditions. Second, we improve the estimate of the decay rate in the case of polygonal domains. Then we show how to modify this method to study Maxwell's equations. The proofs are based on the construction of new multipliers, better adapted to the domain.

Key words. boundary stabilization, multipliers

AMS subject classifications. 30E25, 35B40, 49K15, 93D15

PII. S0363012997323722

1. Introduction.

The problem of exponential decay of the energy of the solutions of the wave equation in a bounded domain by the use of a boundary feedback has been studied by several authors. Bardos, Lebeau, and Rauch proved that the energy decays exponentially:

$$\forall t \geq 0, \quad E(t) \leq CE(0)e^{-\omega t} \tag{1.1}$$

if and only if a "geometric control condition" is satisfied [2]. However, their method closely relies on the hyperbolic structure of the problem, requires smooth geometric conditions, and does not give any information on the constants C and ω. On the other hand, the elementary multiplier methods developed, for example, by Lagnese [11], Lions [13], or Komornik [8] prove (1.1) with explicit constants, so as to measure the observability cost, but only under special geometrical conditions. Lasiecka and Triggiani [12] combined these two methods to obtain intermediate results.

The aim of this paper is to weaken the usual assumption of star-shapedness of the domain using adapted multipliers. In section 2, we introduce the notion of "almost star-shaped domains," and we give some examples. In section 3, we give a result of fast decay of the energy in such domains. In section 4, we use these results to improve the decay rate estimate of the energy in special plane star-shaped domains. At last, we show how to adapt this method to study Maxwell's equations in some non-star-shaped domains. This method is presented in a very well known situation but can also be applied to problems that microlocal analysis techniques are not yet able to solve, for example, to study the nonlinear stabilization of Kirchhoff plate models on some non-star-shaped domains. We can apply this method in another situation: Zuazua [18] studied the stabilization of a semilinear wave equation, with Dirichlet boundary conditions, damped by a linear velocity dissipation that is localized on a neighborhood of a part of the boundary; the notion of an "almost star-shaped domain" allows us to find analogous results for the similar problem with Neumann boundary condition (see [14]). This also allows us to improve the results found on a ball by Alabau and Komornik [1] in the study of the uniform decay of the solutions of general

[*] Received by the editors June 27, 1997; accepted for publication (in revised form) May 12, 1998; published electronically February 19, 1999.

http://www.siam.org/journals/sicon/37-3/32372.html

[†] Département de Mathématiques, E.N.S. Cachan, Antenne de Bretagne, Campus de Ker Lann, 35 170 Bruz, France (martinez@bretagne.ens-cachan.fr).

elastodynamic systems (see also Horn [6] for the study of isotropic elastodynamic systems).

2. Almost star-shaped domains: Definition and examples. In the following, Ω will be a bounded open set of $\mathbb{R}^N$. Let $\{\Gamma_0, \Gamma_1\}$ be a partition of the boundary Γ of Ω. Let us denote by ν the outward unit normal vector to Γ and by $(\mid)$ the canonical Euclidean structure of $\mathbb{R}^N$.

DEFINITION. *$(\Omega, \Gamma_0, \Gamma_1)$ is an almost star-shaped domain if there exists $\phi \in \mathcal{C}^2(\overline{\Omega})$ such that*

$$\text{Sup}\,\{\Delta\phi(x) - 2\lambda_1(x), x \in \Omega\} < \text{Inf}\,\{\Delta\phi(x), x \in \Omega\}, \tag{2.1}$$

$$\partial_\nu \phi \leq \text{ on } \Gamma_0, \tag{2.2}$$

$$\partial_\nu \phi \geq \text{ on } \Gamma_1, \tag{2.3}$$

where $\lambda_1(x)$ is the smallest eigenvalue of the real symmetric squared matrix $D^2\phi(x)$.

In particular, $(\Omega, \Gamma_0, \Gamma_1)$ is almost star-shaped if there exists $\phi \in \mathcal{C}^2(\overline{\Omega})$ such that

$$\Delta\phi = 1 \text{ in } \Omega, \tag{2.4}$$

$$\lambda_1(\phi) := \text{Inf}\,\{\lambda_1(x), x \in \Omega\} > 0, \tag{2.5}$$

$$\partial_\nu \phi \leq 0 \text{ on } \Gamma_0, \tag{2.6}$$

$$\partial_\nu \phi \geq 0 \text{ on } \Gamma_1. \tag{2.7}$$

Remark. Lagnese [11] and Kapitonov [7] introduced similar notions.

The simplest example is the case where Ω is a star-shaped domain with respect to x^0; i.e., there exists a point x^0 such that $(x - x^0 | \nu(x)) \geq 0$ for all $x \in \Gamma$. Then the function

$$\phi_0(x) = \frac{1}{2N}|x - x^0|^2 \tag{2.8}$$

verifies (2.4)–(2.7) with $\Gamma_0 = \emptyset$ and $\Gamma_1 = \Gamma$, so $(\Omega, \emptyset, \Gamma)$ is almost star-shaped.

2.1. Perturbation of a star-shaped domain. In this section, we show that any sufficiently small perturbation of class $\mathcal{C}^2$ of a star-shaped domain Ω of class $\mathcal{C}^2$ gives an almost star-shaped domain. Let Ω be a bounded domain of class $\mathcal{C}^2$, star-shaped with respect to x^1. We can construct $f \in \mathcal{C}^2(\mathbb{R}^N)$ such that

$$\Gamma = \{x \in \mathbb{R}^N, f(x) = 0\} \quad \text{and} \quad |f(x)| \longrightarrow +\infty \text{ as } |x| \longrightarrow +\infty,$$
$$\forall x \in \Gamma,\ \nabla f(x) \neq 0,$$
$$\forall x \in \Gamma,\ \Big(x - x^1 | \nabla f(x)\Big) \geq 0,$$

using the signed distance to Γ, for example.

Set $\varepsilon > 0$ and $g \in \mathcal{C}^2(\mathbb{R}^N)$. Let Ω_ε be the bounded domain whose boundary Γ_ε is the set of points x_ε verifying

$$f(x_\varepsilon) + \varepsilon g(x_\varepsilon) = 0.$$

We apply the implicit functions theorem to deduce that if ε is small enough, x_ε is well defined and the function $\varepsilon \mapsto x_\varepsilon =: x(\varepsilon)$ is of class $\mathcal{C}^2$ on the neighborhood of each point of the boundary Γ.

Let us show that if ε is small enough, $(\Omega_\varepsilon, \emptyset, \Gamma_\varepsilon)$ is almost star-shaped: set $c > 0$ and define

$$\phi(x) = \frac{1}{2N}|x - x^1|^2 + \varepsilon c f(x).$$

Then

$$\begin{aligned}\Delta\phi(x) &= 1 + \varepsilon c \Delta f(x),\\ \Delta\phi(x) - 2\lambda_1(x) &= 1 - \frac{2}{N} + \varepsilon c(\Delta f(x) - 2\lambda_1(f)(x)),\end{aligned}$$

where $\lambda_1(f)(x)$ is the smallest eigenvalue of $D^2 f(x)$. So (2.1) is satisfied if εc is small enough. Now look at the normal derivative of ϕ: if $x_\varepsilon \in \Gamma_\varepsilon$,

$$\Big(\nabla\phi(x_\varepsilon)|\nabla f(x_\varepsilon) + \varepsilon\nabla g(x_\varepsilon)\Big) = \Big(\frac{1}{N}(x_\varepsilon - x^1) + \varepsilon c\nabla f(x_\varepsilon)|\nabla f(x_\varepsilon) + \varepsilon\nabla g(x_\varepsilon)\Big)$$
$$= \frac{1}{N}\Big(x_\varepsilon - x^1|\nabla f(x_\varepsilon)\Big) + \varepsilon c\|\nabla f(x_\varepsilon)\|^2 + \frac{\varepsilon}{N}\Big(x_\varepsilon - x^1|\nabla g(x_\varepsilon)\Big) + \varepsilon^2 c\Big(\nabla f(x_\varepsilon)|\nabla g(x_\varepsilon)\Big).$$

Let us recover Γ by a finite number of neighborhoods V_{x_i} of points $x_i \in \Gamma$ such that $\varepsilon \mapsto x_\varepsilon = x(\varepsilon)$ is of class $\mathcal{C}^2$ on V_{x_i}. Then it is easy to see that there exists $M > 0$ such that

$$\Big|\frac{1}{N}\Big(x_\varepsilon - x^1|\nabla f(x_\varepsilon)\Big) + \frac{\varepsilon}{N}\Big(x_\varepsilon - x^1|\nabla g(x_\varepsilon)\Big) - \frac{1}{N}\Big(x(0) - x^1|\nabla f(x(0))\Big)\Big| \le M\varepsilon.$$

Since $x(0) \in \Gamma$, $(x(0) - x^1|\nabla f(x(0))) \ge 0$. Thus it is sufficient to take c large enough so that

$$M \le \frac{c}{4}\|\nabla f(x(0))|^2 \le \frac{c}{2}\|\nabla f(x_\varepsilon)|^2$$

if ε is small enough. Then (2.3) follows if ε is small enough.

Now assume that $(\Omega, \Gamma_0, \Gamma_1)$ is almost star-shaped with some ϕ verifying (2.4)–(2.7). Set $x^0 \in \mathbb{R}^N$ and ϕ_0 defined by (2.8). Define P by

$$P := \phi - \phi_0. \tag{2.9}$$

Then P verifies

$$\Delta P = 0 \text{ in } \Omega, \tag{2.10}$$

$$\lambda_1(P) + \frac{1}{N} > 0, \tag{2.11}$$

$$\partial_\nu P + \frac{1}{N}\Big(x - x^0|\nu(x)\Big) \ge 0 \text{ on } \Gamma_1, \tag{2.12}$$

$$\partial_\nu P + \frac{1}{N}\Big(x - x^0|\nu(x)\Big) \le 0 \text{ on } \Gamma_0. \tag{2.13}$$

Reciprocally, if there exists $P \in \mathcal{C}^2\ (\overline{\Omega})$ and $x^0 \in \mathbb{R}^N$ such that (2.10)–(2.13) are verified, then ϕ defined by (2.9) verifies (2.4)–(2.7). So the problems (2.4)–(2.7) and (2.10)–(2.13) are equivalent through (2.9). The second problem will be easier to look at, particularly in the case of plane domains, using complex analysis.

2.2. Study of the plane domains using complex analysis. Let Ω be an open bounded connected set of $\mathbb{R}^2$. Let f be a holomorphic function of the complex variable $z = x + iy$ on Ω. We will identify the real vector (x, y) and the complex number $z = x + iy$. Assume that f belongs to $\mathcal{C}^2(\overline{\Omega})$. Let us denote by P its real part and by Q its imaginary part. Differentiating with respect to x and using the Cauchy–Riemann equations, we obtain that

$$f'(z) = \frac{\partial P}{\partial x}(x,y) + i\frac{\partial Q}{\partial x}(x,y) = \frac{\partial P}{\partial x}(x,y) - i\frac{\partial P}{\partial y}(x,y),$$
$$f''(z) = \frac{\partial^2 P}{\partial x^2}(x,y) - i\frac{\partial^2 P}{\partial x \partial y}(x,y).$$

Hence the vector $\nabla P(x, y)$ corresponds to the complex number $\overline{f'(z)}$.

Next we look at the Hessian matrix of the function ϕ defined by (2.9):

$$D^2\phi = \frac{1}{2}Id + D^2 P.$$

Thus

$$\lambda_1(\phi) = \frac{1}{2} + \lambda_1(P). \tag{2.14}$$

We have to compute $\lambda_1(P)$. Since P is a harmonic function, the two eigenvalues of the real symmetric matrix $D^2P(x, y)$, for all $(x, y) \in \Omega$, are real and opposite numbers. Denote them by μ and $-\mu$. We have

$$-\mu^2 = \det\, D^2P(x,y) = \frac{\partial^2 P}{\partial x^2}\frac{\partial^2 P}{\partial y^2} - \Big(\frac{\partial^2 P}{\partial x \partial y}\Big)^2$$
$$= -\Big(\frac{\partial^2 P}{\partial x^2}\Big)^2 - \Big(\frac{\partial^2 P}{\partial x \partial y}\Big)^2 = -|f''(z)|^2.$$

Hence

$$\lambda_1(\phi) = \frac{1}{2} - \|f''\|_\infty, \tag{2.15}$$

where $\|f''\|_\infty = \sup\{|f''(z)|, z \in \Omega\}$.

Now assume $(\Omega, \Gamma_0, \Gamma_1)$ is almost star-shaped and that there exists a harmonic function P that satisfies (2.10)–(2.13). If Ω is simply connected, then there exists a holomorphic function called f whose real part is P. Then f satisfies

$$\|f''\|_\infty < \frac{1}{2}, \tag{2.16}$$

$$\Big(\overline{f'(z)}|\nu(z)\Big) + \frac{1}{2}\Big(z - z_0|\nu(z)\Big) \le 0 \text{ on } \Gamma_0, \tag{2.17}$$

$$\Big(\overline{f'(z)}|\nu(z)\Big) + \frac{1}{2}\Big(z - z_0|\nu(z)\Big) \ge 0 \text{ on } \Gamma_1. \tag{2.18}$$

f'' must be small enough (condition (2.16)), but f' must be important enough, to correct the lack of star-shapedness of Ω if it is necessary (condition (2.18)). These two conditions are not always compatible; in the following section we show how to use them to obtain an explicit example of a domain that is almost star-shaped, although not star-shaped.

2.3. Application: Example of a truly almost star-shaped domain. Let Ω be the following polygonal domain $ABCDEFGH$ (see Fig. 2.1).

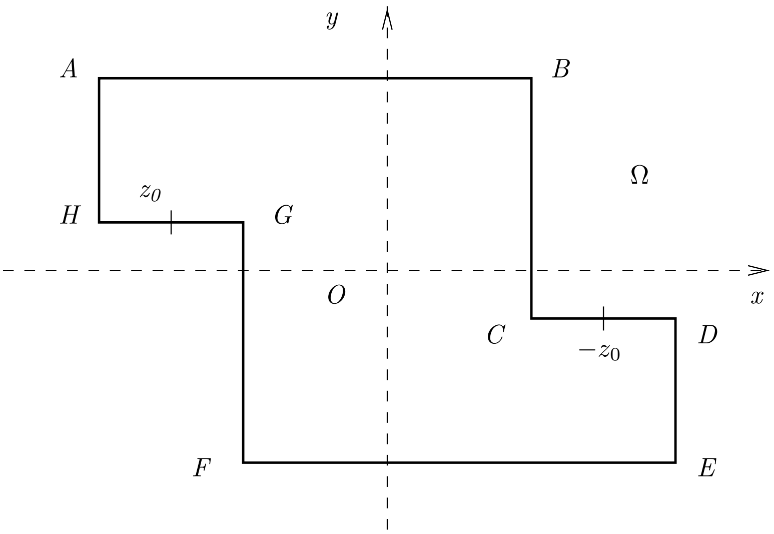


FIG. 2.1. *Example of an almost star-shaped domain.*

Denote

- $\rho :=$ min $\{|z|, z \in \Gamma\}$,
- $R :=$ max $\{|z|, z \in \Gamma\}$,
- z_0 the medium of the segment $[GH]$,
- L the length of the segment $[GH]$.

In order to simplify the calculations, assume that $[CD]$ and $[GH]$ are symmetric with respect to O.

PROPOSITION 2.1. *Let Ω be the polygonal domain ABCDEFGH. If z_0, L, ρ, R verify*

$$0 < \text{Im } z_0 \leq \frac{\rho}{R}\Big(|z_0| - \frac{L}{2}\Big), \tag{2.19}$$

then $(\Omega, \emptyset, \Gamma)$ is almost star-shaped without being star-shaped.

Remark. This provides an explicit example of an almost star-shaped domain that is not star-shaped. Then it is sufficient to smooth the corners to obtain an explicit example of an almost star-shaped domain of class $\mathcal{C}^2$.

Proof of Proposition 2.1. The hypothesis Im $z_0 > 0$ ensures that Ω is not star-shaped. In order to correct that lack of star-shapedness, we perturb $\phi_0(z) = \frac{1}{4}|z|^2$ by a harmonic function, chosen to be small enough so that (2.16) can be true, but whose normal derivative will correct $\partial_\nu \phi_0$ where it is necessary.

Set $\eta > 0$; we construct a holomorphic function f such that

$$\overline{f'(z_0)} = -\eta i \text{ and } \overline{f'(-z_0)} = \eta i.$$

We can choose

$$f(z) = \frac{1}{2}\eta i \frac{z^2}{z_0}. \tag{2.20}$$

Let us define ϕ by

$$\phi(x,y) = \frac{1}{4}|z|^2 + \operatorname{Re} f(z).$$

We have already shown that

$$\begin{aligned} \nabla\phi(x,y) &= \frac{1}{2}z + \overline{f'(z)}, \\ \lambda_1(\phi) &= \frac{1}{2} - \|f''\|_\infty. \end{aligned}$$

First we study $\partial_\nu\phi$ on $[CD]$ and on $[GH]$; using the symmetries of the domain and of the function, it is sufficient to study on $[GH]$, using the following parametrization:

$$z = t\Big(z_0 + \frac{L}{2}\Big) + (1-t)\Big(z_0 - \frac{L}{2}\Big) = z_0 + \Big(t - \frac{1}{2}\Big)L,\ t \in [0,1].$$

So

$$\begin{aligned} \partial_\nu\phi(z) &= -\frac{1}{2}\operatorname{Im} z_0 + \Big(\overline{\frac{\eta i}{z_0}\Big(z_0 + \Big(t-\frac{1}{2}\Big)L\Big)} | -i\Big) \\ &= -\frac{1}{2}\operatorname{Im} z_0 - \operatorname{Im}\Big(\overline{\frac{\eta i}{z_0}\Big(z_0 + \Big(t-\frac{1}{2}\Big)L\Big)}\Big) \\ &\geq \eta - \frac{1}{2}\operatorname{Im} z_0 - \frac{\eta L}{2|z_0|}. \end{aligned}$$

Next we study $\partial_\nu\phi$ on the other sides:

$$\partial_\nu\phi(z) = \frac{1}{2}\Big(z|\nu(z)\Big) + \Big(\overline{f'(z)}|\nu(z)\Big) \geq \frac{1}{2}\rho - \eta\frac{R}{|z_0|}.$$

That gives us the following conditions:

$$\|f''\|_\infty = \frac{\eta}{|z_0|} < \frac{1}{2}, \tag{2.21}$$

$$\eta - \frac{1}{2}\operatorname{Im} z_0 - \frac{\eta L}{2|z_0|} \geq 0, \tag{2.22}$$

$$\rho - 2\eta\frac{R}{|z_0|} \geq 0. \tag{2.23}$$

Equation (2.21) is satisfied when (2.23) is true, and the hypothesis (2.19) allows us to find η verifying (2.22)–(2.23). Thus $(\Omega, \emptyset, \Gamma)$ is almost star-shaped without being star-shaped. □

3. Uniform boundary stabilization of the wave equation in a smooth almost star-shaped domain. Let Ω be a bounded open set of $\mathbb{R}^N$ of class $\mathcal{C}^2$. Let us denote by ν the outward unit normal vector to Γ. Assume that the partition $\{\Gamma_0, \Gamma_1\}$ of the boundary satisfies

$$\overline{\Gamma_0} \cap \overline{\Gamma_1} = \emptyset. \tag{3.1}$$

Let q be a bounded nonnegative function on Ω, and a, ℓ be two nonnegative functions of class $\mathcal{C}^1$ on Γ_1 such that

$$q \not\equiv 0 \ \text{ or } \ \Gamma_0 \not\equiv 0 \ \text{ or } \ a \not\equiv 0. \tag{3.2}$$

As usual, let us denote $H^1_{\Gamma_0}(\Omega) := \{u \in H^1(\Omega), u = 0 \text{ on } \Gamma_0\}$. We consider the following evolutionary problem that has been studied by several authors, such as Komornik and Zuazua [10], Komornik [8], and Tcheugoué Tébou [17], on star-shaped domains:

$$u'' - \Delta u + qu = 0 \text{ in } \Omega, \tag{3.3}$$

$$u = 0 \text{ on } \Gamma_0, \tag{3.4}$$

$$\partial_\nu u + au + \ell u' = 0 \text{ on } \Gamma_1, \tag{3.5}$$

$$u(0) = u^0, u'(0) = u^1. \tag{3.6}$$

Applying a carefully chosen feedback and using a method introduced in Komornik and Zuazua [10], we shall obtain uniform stabilization with rather precise decay rate estimates. Then we will show how to use this result to improve the decay rate on some domains, carefully choosing the function ϕ that satisfies (2.4)–(2.7).

3.1. Existence and regularity theorem.

THEOREM 3.1. *Assume* (3.1)–(3.2).

1. *Given* $(u^0, u^1) \in H^1_{\Gamma_0}(\Omega) \times L^2(\Omega)$*, the problem* (3.3)–(3.6) *has a unique (so-called weak) solution*

$$u \in \mathcal{C}\Big(\mathbb{R}_+, H^1_{\Gamma_0}(\Omega)\Big) \cap \mathcal{C}^1\Big(\mathbb{R}_+, L^2(\Omega)\Big). \tag{3.7}$$

The energy $E : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ *of the solution* u *defined by*

$$E(t) = \frac{1}{2}\int_\Omega \Big(u'^2 + |\nabla u|^2 + qu^2\Big)\, dx + \frac{1}{2}\int_{\Gamma_1} au^2\, d\sigma \tag{3.8}$$

is a nonincreasing function.

2. *If* u^0 *and* u^1 *satisfy the stronger conditions*

$$\begin{cases} (u^0, u^1) \in H^2(\Omega) \cap H^1_{\Gamma_0}(\Omega) \times H^1_{\Gamma_0}(\Omega), \\ \partial_\nu u^0 + au^0 + \ell u^1 = 0 \ \text{ on } \Gamma_1, \end{cases} \tag{3.9}$$

then the solution u *has the stronger regularity property:*

$$u \in L^\infty\Big(\mathbb{R}_+, H^2(\Omega)\Big), u' \in L^\infty\Big(\mathbb{R}_+, H^1_{\Gamma_0}(\Omega)\Big), u'' \in L^\infty\Big(\mathbb{R}_+, L^2(\Omega)\Big). \tag{3.10}$$

In this case, u *will be called strong solution of the problem* (3.3)–(3.6).

This theorem is well known (see, for example, Komornik [9]).

3.2. Uniform stabilization under a carefully chosen feedback.

THEOREM 3.2. *Let* Ω *be a bounded open set of* $\mathbb{R}^N$ *of class* $\mathcal{C}^2$*; assume* $(\Omega, \Gamma_0, \Gamma_1)$ *is almost star-shaped. Assume that there exists* $\phi \in \mathcal{C}^2(\overline{\Omega})$ *satisfying* (2.4)–(2.7). *Define*

$$C := \max\left(\frac{2}{3}, 1 - \lambda_1(\phi)\right) \text{ and } Q_1 := \|\sqrt{q}\nabla\phi\|_\infty, \tag{3.11}$$

and assume that

$$C + Q_1 < 1. \tag{3.12}$$

Define the functions a and ℓ by

$$a = \frac{C}{2\|\nabla\phi\|_\infty^2}\partial_\nu\phi \text{ and } \ell = \frac{1}{\|\nabla\phi\|_\infty}\partial_\nu\phi. \tag{3.13}$$

Given $(u^0, u^1) \in H^1_{\Gamma_0}(\Omega) \times L^2(\Omega)$, the energy of the weak solution u satisfies the following estimate:

$$\forall t \geq 0, \quad E(t) \leq E(0)e^{1-\omega t} \text{ with } \omega := \frac{1 - C - Q_1}{2\|\nabla\phi\|_\infty}. \tag{3.14}$$

Remarks. 1. If ϕ satisfies only (2.1)–(2.3), the estimate (3.14) holds true. The only difference comes from the value of C.

2. One can easily generalize Theorem 3.2 to more general functions a, ℓ, and q, for example, assuming that a, ℓ, and $\partial_\nu\phi$ are positive on Γ_1 or using a nonlinear feedback as in Komornik [9].

3. Assume now that Ω is a bounded polygonal domain without crack such that each side belongs to Γ_0 or to Γ_1. Moreover, assume that the angles of the corners where there is a change of boundary condition have a measure strictly less than π. Then it is easy to adapt the method that Grisvard [4] or Moussaoui [15] used to study the regularity of the solutions of the problem of the wave equation with mixed boundary conditions (Dirichlet condition on Γ_0 and Neumann condition on Γ_1) in such domains. We find that the strong solutions of (3.3)–(3.6) have the following regularity:

$$u \in L^\infty(\mathbb{R}_+, H^s(\Omega))$$

for some $s > \frac{3}{2}$. This regularity allows us to justify the computations that lead to the key identity given in Lemma 3.4. Thus the estimate (3.14) holds true in such domains.

Proof of Theorem 3.2. The proof of this theorem is based on the multiplier method. We will prove only the estimate (3.14) for smooth initial data, i.e., (u^0, u^1) satisfying (3.9). The general case then follows by an easy density argument. The validity of all the computations is guaranteed by the regularity of u given by (3.10). We use the following convention:

$$\int_S^T \int_\Omega \quad \text{means} \quad \int_S^T \int_\Omega dx\ dt \quad \text{and} \quad \int_S^T \int_\Gamma \quad \text{means} \quad \int_S^T \int_\Gamma d\sigma\ dt.$$

Lemma 3.3. *Given (u^0, u^1) verifying* (3.9), *the strong solution of* (3.3)–(3.6) *satisfies*

$$\forall 0 \leq S < T < +\infty, \quad E(S) - E(T) = \int_S^T \int_{\Gamma_1} \ell u'^2. \tag{3.15}$$

Proof of Lemma 3.3. We multiply (2.5) by u' and we integrate by parts on

$\Omega \times [S,T]$:

$$\begin{aligned} 0 &= \int_S^T \int_\Omega u'(u'' - \Delta u + qu) \\ &= \frac{1}{2}\left[\int_\Omega u'^2 + |\nabla u|^2 + qu^2\right]_S^T - \int_S^T \int_\Gamma u'\partial_\nu u \\ &= \frac{1}{2}\left[\int_\Omega u'^2 + |\nabla u|^2 + qu^2 + \int_{\Gamma_1} au^2\right]_S^T + \int_S^T \int_{\Gamma_1} \ell u'^2. \qquad \square \end{aligned}$$

Set $c \in \mathbb{R}$ and

$$M(u) = 2\nabla\phi \cdot \nabla u + cu.$$

The key to the proof of Theorem 3.2 is the following basic identity.

LEMMA 3.4. *Given* (u^0, u^1) *verifying* (2.11) *and* $0 \le S < T < +\infty$, *we have*

$$\begin{aligned} &\int_S^T \int_{\Gamma_0} (\partial_\nu u)^2 \partial_\nu \phi + \int_S^T \int_{\Gamma_1} -(au + \ell u')M(u) + \partial_\nu \phi\ (u'^2 - |\nabla u|^2) \qquad (3.16) \\ &= \left[\int_\Omega u'M(u)\right]_S^T + \int_S^T \int_\Omega (1-c)(u'^2 - |\nabla u|^2) + quM(u) + 2\Big(D^2\phi \cdot \nabla u|\nabla u\Big). \end{aligned}$$

Proof of Lemma 3.4. We multiply (3.3) by $M(u)$ and we integrate by parts on $\Omega \times [S,T]$.

$$0 = \int_S^T \int_\Omega (u'' - \Delta u + qu)M(u)$$

$$\begin{aligned} = \left[\int_\Omega u'M(u)\right]_S^T - \int_S^T \int_\Omega u'(2\nabla\phi \cdot \nabla u' + cu') - \int_S^T \int_\Gamma \partial_\nu u\ M(u) \\ + \int_S^T \int_\Omega \nabla u \cdot \nabla(2\nabla\phi \cdot \nabla u + cu) + \int_S^T \int_\Omega quM(u) \end{aligned}$$

$$\begin{aligned} = \left[\int_\Omega u'M(u)\right]_S^T - \int_S^T \int_{\Gamma_1} \partial_\nu u\ M(u) - \int_S^T \int_{\Gamma_0} 2(\partial_\nu u)^2 \partial_\nu \phi + \int_S^T \int_\Omega quM(u) \\ - \int_S^T \int_\Omega \nabla\phi \cdot \nabla(u'^2) + cu'^2 + \int_S^T \int_\Omega c|\nabla u|^2 + 2\partial_i u\ \partial_i(\partial_j \phi\ \partial_j u) \end{aligned}$$

$$\begin{aligned} = \left[\int_\Omega u'M(u)\right]_S^T - \int_S^T \int_{\Gamma_1} \partial_\nu u M(u) - \int_S^T \int_{\Gamma_0} 2(\partial_\nu u)^2 \partial_\nu \phi + \int_S^T \int_\Omega quM(u) \\ - \int_S^T \int_\Gamma u'^2 \partial_\nu \phi + \int_S^T \int_\Omega u'^2 - c(u'^2 - |\nabla u|^2) + 2\Big(D^2\phi \cdot \nabla u|\nabla u\Big) + \nabla\phi \cdot \nabla\Big(|\nabla u|^2\Big). \end{aligned}$$

Thus, putting the boundary integrals in the left-hand side,

$$\int_S^T \int_{\Gamma_0} (\partial_\nu u)^2 \partial_\nu \phi + \int_S^T \int_{\Gamma_1} \partial_\nu u \ M(u) + \partial_\nu \phi \ (u'^2 - |\nabla u|^2)$$
$$= \left[\int_\Omega u' M(u) \right]_S^T + \int_S^T (1-c)(u'^2 - |\nabla u|^2) + 2(D^2\phi \cdot \nabla u | \nabla u) + quM(u). \qquad \square$$

Remark. As $u = 0$ on Γ_0, we used the fact that

$$\nabla u = (\partial_\nu u)\nu \text{ on } \Gamma_0.$$

First we estimate the terms of the left-hand side of (3.16). In the following, we will take $c > 0$, and we define a and ℓ by

$$a = \frac{c}{2\|\nabla\phi\|_\infty^2} \partial_\nu \phi \text{ and } \ell = \frac{1}{\|\nabla\phi\|_\infty} \partial_\nu \phi. \tag{3.17}$$

Lemma 3.5. *Assume a and ℓ are defined by* (3.17)*. Given (u^0, u^1) satisfying* (3.9)*, we have*

$$\int_S^T \int_{\Gamma_0} (\partial_\nu u)^2 \partial_\nu \phi + \int_S^T \int_{\Gamma_1} -(au + \ell u')M(u) + \partial_\nu \phi \ (u'^2 - |\nabla u|^2)$$
$$\leq 2\|\nabla\phi\|_\infty \Big(E(S) - E(T)\Big) - \frac{c}{2} \int_S^T \int_{\Gamma_1} au^2. \tag{3.18}$$

Proof of Lemma 3.5. According to (3.17), denote $a = \alpha \partial_\nu \phi$ and $\ell = \lambda \partial_\nu \phi$. Then

$$-(au + \ell u')(2\nabla\phi \cdot \nabla u + cu) + \partial_\nu \phi (u'^2 - |\nabla u|^2)$$
$$\leq \partial_\nu \phi \Big(\|\nabla\phi\|_\infty^2 (\alpha u + \lambda u')^2 + |\nabla u|^2 - c\alpha u^2 - c\lambda u u' + u'^2 - |\nabla u|^2 \Big)$$
$$\leq \partial_\nu \phi \Big((1 + \lambda^2 \|\nabla\phi\|_\infty^2) u'^2 + (\alpha^2 \|\nabla\phi\|_\infty^2 - c\alpha) u^2 + (2\alpha\lambda \|\nabla\phi\|_\infty^2 - c\lambda) uu' \Big).$$

Equation (3.18) follows from the carefully chosen values of α and λ given by (3.17), from the identity (3.15), and from (2.6). □

Lemma 3.6.

$$\left| \int_\Omega u' M(u) \right| \leq 2\|\nabla\phi\|_\infty E(t) \quad \textit{if } c \in [0,2]. \tag{3.19}$$

Proof of Lemma 3.6.

$$\int_\Omega (M(u))^2 = \int_\Omega (2\nabla\phi \cdot \nabla u + cu)^2 = \int_\Omega (2\nabla\phi \cdot \nabla u)^2 + c^2u^2 + 2c\nabla\phi \cdot \nabla(u^2)$$
$$= \int_\Omega (2\nabla\phi \cdot \nabla u)^2 + c^2u^2 + 2c\int_\Gamma \partial_\nu \phi \ u^2 - \int_\Omega 2cu^2$$
$$= \int_\Omega (2\nabla\phi \cdot \nabla u)^2 + c(c-2)u^2 + 2c\int_{\Gamma_1} \partial_\nu \phi \ u^2.$$

Assume $c \in [0,2]$. Then

$$\int_\Omega M(u)^2 \leq \int_\Omega (2\nabla\phi \cdot \nabla u)^2 + 2c \int_{\Gamma_1} \partial_\nu \phi \; u^2.$$

Set $d^2 = 2\|\nabla\phi\|_\infty$. Applying the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}\left|\int_\Omega u' M(u)\right| &\leq \left(\int_\Omega u'^2\right)^{1/2} \left(\int_\Omega M(u)^2\right)^{1/2}\\ &\leq \frac{d^2}{2}\int_\Omega u'^2 + \frac{1}{2d^2}\int_\Omega (2\nabla\phi\cdot\nabla u)^2 + \frac{c}{d^2}\int_{\Gamma_1} \partial_\nu\phi \; u^2\\ &\leq 2\|\nabla\phi\|_\infty E(t). \qquad \square\end{aligned}$$

Using the estimates (3.18) and (3.19), (3.16) becomes

$$\int_S^T \int_\Omega (1-c)(u'^2 - |\nabla u|^2) + 2(D^2\phi \cdot \nabla u|\nabla u) + quM(u) \\ + \frac{c}{2}\int_S^T \int_{\Gamma_1} au^2 \leq 4\|\nabla\phi\|_\infty E(S). \tag{3.20}$$

Thanks to (2.5), we can say that

$$\int_S^T \int_\Omega (1-c)u'^2 + \Big(2\lambda_1(\phi) - (1-c)\Big)|\nabla u|^2 + cqu^2 + 2qu\nabla\phi\cdot\nabla u \\ + \frac{c}{2}\int_S^T \int_{\Gamma_1} au^2 \leq 4\|\nabla\phi\|_\infty E(S). \tag{3.21}$$

Last, using the Cauchy–Schwarz inequality in the same way, we have

$$\left|\int_\Omega 2qu\nabla\phi\cdot\nabla u\right| \leq 2\|\sqrt{q}\nabla\phi\|_\infty E(t) = 2Q_1 E(t). \tag{3.22}$$

Thus

$$2\Big(\inf\Big\{\frac{c}{2}, (1-c), 2\lambda_1(\phi) - (1-c)\Big\} - Q_1\Big)\int_S^T E \leq 4\|\nabla\phi\|_\infty E(S). \tag{3.23}$$

It is easy to see that the constant c gives the better result, obtained when the coefficient of $\int_S^T E$ in (3.23) is the biggest, i.e., when c has the value C given by (3.11). Then (3.23) becomes

$$2(1 - C - Q_1)\int_S^T E \leq 4\|\nabla\phi\|_\infty E(S). \tag{3.24}$$

If $1-C-Q_1 > 0$, letting T go to $+\infty$, we see that E is a nonnegative and nonincreasing function that satisfies

$$\forall S \geq 0, \quad \int_S^{+\infty} E \leq \frac{1}{\omega} E(S).$$

Using a Gronwall-type inequality as in Haraux [5], we conclude that

$$\forall t \geq 0, \quad E(t) \leq E(0)e^{1-\omega t}.$$

We recall the proof of this inequality briefly: set $g(t) = \int_t^{+\infty} E(\tau)\, d\tau$. g satisfies the differential inequality

$$\forall t \geq 0,\ g'(t) + \omega g(t) \leq 0.$$

Thus

$$\forall t \geq 0,\ g(t) \leq g(0)e^{-\omega t} \leq \frac{1}{\omega} E(0)e^{-\omega t}.$$

Then since E is nonnegative and nonincreasing, for all $\varepsilon > 0$ we have

$$E(t) \leq \frac{1}{\varepsilon} \int_{t-\varepsilon}^{t} E(\tau)\, d\tau \leq \frac{1}{\varepsilon} g(t-\varepsilon) \leq \frac{1}{\omega\varepsilon} E(0)e^{\omega\varepsilon}e^{-\omega t},$$

and the best estimate is obtained for $\omega\varepsilon = 1$. □

Remark. When ϕ satisfies only the weaker conditions (2.1)–(2.3), one has to replace the coefficient $(1-c)$ by $(\Delta\phi - c)$ in (3.20). The assumption (2.1) allows us to choose a constant c such that (3.24) holds true.

4. Application: Fast decay of the energy in regular polygons. Let Ω be a convex polygon and $x^0 \in \Omega$. The function $\phi(x) = \phi_0(x) := \frac{1}{4}|x - x^0|^2$ satisfies

$$\Delta\phi = \frac{1}{2} + \frac{1}{2} = 1 \text{ in } \Omega, \tag{4.1}$$

$$\lambda_1(\phi) \left(= \frac{1}{2} \right) > 0, \tag{4.2}$$

$$\partial_\nu \phi = \Big(x - x^0 | \nu(x) \Big) \geq 0 \text{ on } \Gamma. \tag{4.3}$$

Define a and ℓ by (2.15). Given $(u^0, u^1) \in H^1(\Omega) \times L^2(\Omega)$, the solution of the evolutionary system

$$u'' - \Delta u = 0 \text{ in } \Omega, \tag{4.4}$$

$$\partial_\nu u + au + \ell u' = 0 \text{ on } \Gamma, \tag{4.5}$$

$$u(0) = u^0,\ u'(0) = u^1, \tag{4.6}$$

satisfies the energy estimate:

$$E(t) \leq E(0)e^{1-\omega t}, \text{ with } \omega = \frac{\inf\{\frac{1}{3}, \lambda_1(\phi)\}}{2\|\nabla\phi\|_\infty} = \frac{1}{3R(x^0)}, \tag{4.7}$$

where $R(x^0) := \sup_{x \in \Omega} |x - x^0|$. Our purpose is to improve this decay rate, first found by Tcheugoué Tébou [17], perturbing ϕ_0 by harmonic functions adapted to the domain.

4.1. Main result and first properties. Let Ω be the n-sided regular polygon centered in O whose vertices are the points $\mathrm{Re}^{2ik\pi/n}, 0 \leq k \leq n-1$.

THEOREM 4.1. *Let Ω be the n-sided regular polygon centered in O. Thanks to a careful choice of the function ϕ (and consequently the choice of a and ℓ), the solutions of the problem* (4.4)–(4.6) *satisfy the estimate*

$$\forall t \geq 0, \quad E(t) \leq E(0)e^{1-t/\rho_n}, \tag{4.8}$$

with

$$\rho_n = \left(3 - \frac{1}{n-1}\right) R \textit{ if } 3 \leq n \leq 8, \tag{4.9}$$

$$\rho_n = (3 - 6\varepsilon_n)R \textit{ with } \varepsilon_n = \frac{\sin^2 \frac{\pi}{2n}}{1 + (\cos \frac{\pi}{n})^{n-1}} \simeq \frac{\pi^2}{8n^2} \textit{ if } n \geq 9. \tag{4.10}$$

Moreover, in the case of the equilateral triangle ($n = 3$), *improved computations give the better estimate*

$$\rho_3 \leq 2,36R. \tag{4.11}$$

The function ϕ will be explicitly given in the proof.

Remarks. 1. A general principle of Russell [16] shows that uniform stabilization implies the exact controllability of the associated system in time $T > T_0 = \rho_n$. Applying this principle for the equilateral triangle, we obtain $T_0 \geq \rho_3 \geq 2R > R\sqrt{3}$. Thus the method is not powerful enough to find results that are already known (exact controllability in time $T > 2R$). So we cannot expect to find optimal results like Bardos, Lebeau, and Rauch [2] obtained when the domain is analytical.

2. It is an interesting question whether there exists an "optimal" function ϕ providing an optimal decay rate.

3. A technical difficulty (that Lemma 4.3 can help to imagine) prevents us from applying this method in dimension 3 to improve the decay rate for the tetrahedron, for example.

Proof of Theorem 4.1. Equation (4.8) is a consequence of Theorem 3.2 with a good choice of ϕ. We have to find ϕ such that $\omega(\phi)$ is the biggest possible. Because of (4.1), ϕ can be written as the perturbation of ϕ_0 by a harmonic function. It is clear that ∇P has to be entering at the vertices of Ω in order to reduce $\|\nabla\phi\|_\infty$. For example, if Ω is the equilateral triangle, ∇P has to be as indicated in Fig. 4.1.

First we study some properties we need to optimize the decay rate.

LEMMA 4.2. *Set $\phi \in \mathcal{C}^2(\overline{\Omega})$ such that $\lambda_1(\phi) > 0$. Then $\|\nabla\phi\|$ attains its upper bound only on the boundary of Ω.*

Proof of Lemma 4.2. An easy continuity argument shows that $\|\nabla\phi\|$ is bounded and attains its upper bound on $\overline{\Omega}$. Assume this happens in $z_1 = (x_1, y_1) \in \Omega$. As ϕ is $\mathcal{C}^2$,

$$\frac{\partial}{\partial x}\Big(\Big(\frac{\partial\phi}{\partial x}\Big)^2(x_1, y_1) + \Big(\frac{\partial\phi}{\partial y}\Big)^2(x_1, y_1)\Big) = 0,$$
$$\frac{\partial}{\partial y}\Big(\Big(\frac{\partial\phi}{\partial x}\Big)^2(x_1, y_1) + \Big(\frac{\partial\phi}{\partial y}\Big)^2(x_1, y_1)\Big) = 0.$$

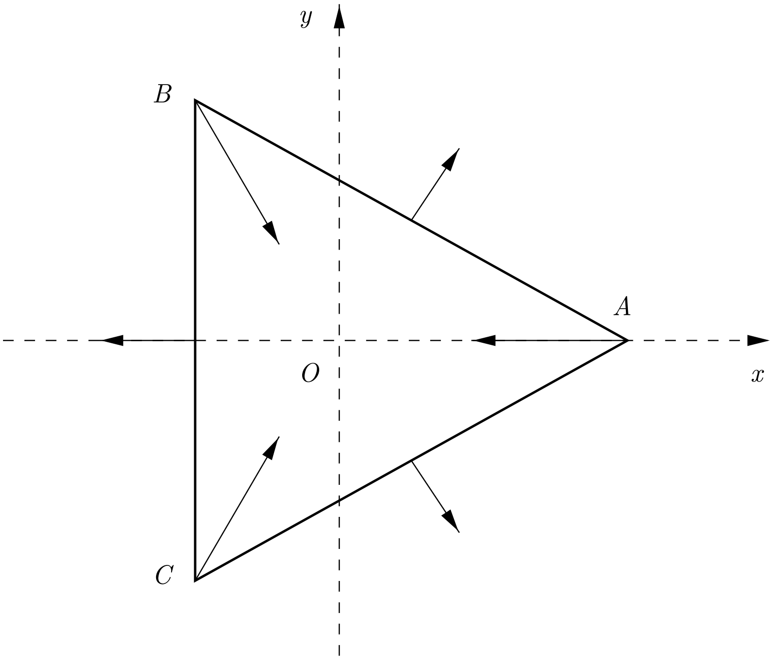


FIG. 4.1. *Behavior of* ∇P *on* Γ.

Thus we deduce that

$$D^2\phi \cdot \nabla\phi(x_1, y_1) = 0,$$

and this is impossible because $D^2\phi$ is invertible everywhere, and so $\nabla\phi(x_1, y_1)$ cannot be equal to $\overrightarrow{0}$ because of its definition. □

LEMMA 4.3. *Given* ϕ *verifying* (4.1)–(4.3) *and*

$$\nabla\phi(O) = 0, \tag{4.12}$$

the decay rate given by ϕ *can be optimal only if*

$$\lambda_1(\phi) \geq \frac{1}{3}. \tag{4.13}$$

Remark. We will see next that the additional hypothesis (4.12) is only a consequence of the symmetries of the polygon.

Proof of Lemma 4.3. Assume ϕ verifies (4.1)–(4.3), (4.12), and $\lambda_1(\phi) < \frac{1}{3}$. As usual, denote $\phi(z) - \frac{1}{4}|z|^2 = P(z) =$ Re $f(z)$. Define

$$\phi_\varepsilon = \phi_0 + \varepsilon P, \tag{4.14}$$

and choose $\varepsilon > 0$ such as

$$\lambda_1(\phi_\varepsilon) = \frac{1}{2} - \|\varepsilon f''\|_\infty = \frac{1}{3}.$$

Then ϕ_ε verifies (4.1), (4.2), (4.12), and (4.13). Assume ϕ_ε verifies (4.3). Then we shall prove that

$$\omega(\phi_\varepsilon) = \frac{1}{6\|\nabla\phi_\varepsilon\|_\infty} \geq \omega(\phi) = \frac{\lambda_1(\phi)}{2\|\nabla\phi\|_\infty}. \tag{4.15}$$

Effectively,

$$\begin{aligned}&\phi = \phi_0 + P = \phi_0 + \varepsilon P + (1-\varepsilon)P = \phi_\varepsilon + (1-\varepsilon)P,\\ &\nabla\phi(z) = \nabla\phi_\varepsilon(z) + (1-\varepsilon)\overline{f'(z)}.\end{aligned}$$

As $\overline{f'(0)} = \nabla P(O) = 0$, $|f'(z)| \leq R\|f''\|_\infty$. So

$$\|\nabla\phi\|_\infty \geq \|\nabla\phi_\varepsilon\|_\infty - (1-\varepsilon)\|f'\|_\infty,$$
$$\lambda_1(\phi) = \frac{1}{2} - \|f''\|_\infty = \frac{1}{2} - \varepsilon\|f''\|_\infty - (1-\varepsilon)\|f''\|_\infty = \frac{1}{3} - (1-\varepsilon)\|f''\|_\infty.$$

We deduce from this study that

$$\frac{2\|\nabla\phi\|_\infty}{\lambda_1(\phi)} \geq 2\frac{\|\nabla\phi_\varepsilon\|_\infty - (1-\varepsilon)R\|f''\|_\infty}{\frac{1}{3} - (1-\varepsilon)\|f''\|_\infty}, \tag{4.16}$$

but on the other hand

$$M := \frac{\|\nabla\phi_\varepsilon\|_\infty}{R} \geq \frac{1}{2} - \frac{\varepsilon}{R}\|\nabla P\|_\infty \geq \frac{1}{2} - \varepsilon\|f''\|_\infty = \frac{1}{3},$$

so (4.16) gives

$$\frac{2\|\nabla\phi\|_\infty}{\lambda_1(\phi)} \geq 2R\frac{M - (1-\varepsilon)\|f''\|_\infty}{\frac{1}{3} - (1-\varepsilon)\|f''\|_\infty} \geq 2R\frac{M}{\frac{1}{3}} = 6\|\nabla\phi_\varepsilon\|_\infty. \qquad \square$$

Next we study the problem of the normal derivative of ϕ_ε.

LEMMA 4.4. *Set $\phi \in \mathcal{C}^2(\overline{\Omega})$ verifying* (4.1), (4.12), *and* (4.13)*; then the condition* (4.3) *is automatically satisfied: $\partial_\nu\phi \geq 0$ on the boundary.*

Proof of Lemma 4.4. Set $z_k = \mathrm{Re}^{2ik\pi/n}, 0 \leq k \leq n-1$, and denote ν_k the outward unit normal vector to $]z_k, z_{k+1}[$.

$$\nu_k = \frac{z_{k+1} - z_k}{|z_{k+1} - z_k|}e^{-i\pi/2}. \tag{4.17}$$

All along the segment $]z_k, z_{k+1}[$, the normal derivative of ϕ is

$$\begin{aligned}\partial_\nu\phi(z) = \left(\frac{z}{2} + \overline{f'(z)}|\nu_k\right) &= \left(\frac{z}{2}|\nu_k\right) + \left(\overline{f'(z)}|\nu_k\right)\\ &= \left(\frac{z_k}{2}|\nu_k\right) + \left(\overline{f'(z)}|\nu_k\right) \geq \frac{R}{2}\cos\frac{\pi}{n} - R\|f''\|_\infty.\end{aligned}$$

However, $\cos\frac{\pi}{n} \geq \cos\frac{\pi}{3} \geq \frac{1}{2}$ and $\|f''\|_\infty = \frac{1}{2} - \lambda_1(\phi) \leq \frac{1}{6}$, so (4.3) follows. $\square$

This drives us to study the following optimization problem:

$$\begin{cases}\|f''\|_\infty \leq \frac{1}{6},\\ 6\|\nabla\phi\|_\infty \text{ the smallest possible.}\end{cases} \tag{4.18}$$

First assume that Ω is the equilateral triangle.

LEMMA 4.5. *If $n = 3$, we will seek only the functions f that have a development of the form*

$$f(z) = \sum_{k=1}^{+\infty} a_{3k} z^{3k}. \tag{4.19}$$

This lemma has a simple geometrical application: it is sufficient to study the harmonic functions P such that the gradient in each vertex S of the polygon is colinear to the vector $\overrightarrow{\mathrm{OS}}$.

Proof of Lemma 4.5. Set f a holomorphic function on Ω defined by a power series whose convergence radius is bigger than R:

$$f(z) = \sum_{k=0}^{+\infty} a_k z^k.$$

Then define for $i = 0$ to 2

$$g_i(z) = \sum_{k=1}^{+\infty} a_{3k+i} z^{3k+i}. \tag{4.20}$$

We show in the following that g_0 is better than f for our optimization problem: if ϕ (respectively, ψ_0) represents the function associated with f (respectively, with g_0), then

$$\begin{cases} \|g_0''\|_\infty \leq \|f''\|_\infty, \\ \|\nabla \psi_0\|_\infty \leq \|\nabla \phi\|_\infty. \end{cases} \tag{4.21}$$

Let z_0 be a maximum point of g_0'': $\|g_0''\|_\infty = |g_0''(z_0)|$. Let us show that the following system cannot be possible:

$$\begin{cases} |f''(z_0)| < |g_0''(z_0)|, \\ |f''(z_0 e^{2i\pi/3})| < |g_0''(z_0 e^{2i\pi/3})|, \\ |f''(z_0 e^{-2i\pi/3})| < |g_0''(z_0 e^{-2i\pi/3})|. \end{cases} \tag{4.22}$$

Denote $\beta_i := g_i''(z_0)$. Since $f = g_0 + g_1 + g_2$, the system (4.22) can be written

$$\begin{cases} |\beta_0 + \beta_1 + \beta_2| < |\beta_0|, \\ |\beta_0 e^{-4i\pi/3} + \beta_1 e^{-2i\pi/3} + \beta_2| < |\beta_0 e^{-4i\pi/3}|, \\ |\beta_0 e^{4i\pi/3} + \beta_1 e^{2i\pi/3} + \beta_2| < |\beta_0 e^{4i\pi/3}|. \end{cases} \tag{4.23}$$

If $\beta_0 = 0$, the result is clear. Otherwise, define

$$\gamma_1 := \frac{\beta_1}{\beta_0} \quad \text{and} \quad \gamma_2 := \frac{\beta_2}{\beta_0}.$$

Dividing each equation by $|\beta_0|$, we get from (4.23)

$$\begin{cases} |1 + \gamma_1 + \gamma_2| < 1, \\ |1 + \gamma_1 e^{2i\pi/3} + \gamma_2 e^{-2i\pi/3}| < 1, \\ |1 + \gamma_1 e^{-2i\pi/3} + \gamma_2 e^{2i\pi/3}| < 1. \end{cases} \tag{4.24}$$

But this is impossible:

$$|1+\gamma_1+\gamma_2|^2+|1+\gamma_1 e^{2i\pi/3}+\gamma_2 e^{-2i\pi/3}|^2+|1+\gamma_1 e^{-2i\pi/3}+\gamma_2 e^{2i\pi/3}|^2 \\ =3(1+|\gamma_1|^2+|\gamma_2|^2)\geq 3.$$

Thus

$$\|g_0''\|_\infty \leq \|f''\|_\infty.$$

We can proceed in the same way to prove that

$$\|\nabla\psi_0\|_\infty \leq \|\nabla\phi\|_\infty. \qquad \square$$

In the same way, when Ω is the n-sided regular polygon centered in O, we will seek f in the form

$$f(z)=\sum_{k=0}^{+\infty} a_{kn} z^{kn}. \tag{4.25}$$

4.2. Explicit computations: First study. Let Ω be the n-sided regular polygon. Thanks to the preliminary results, we study the results given by the function

$$f(z)=-\varepsilon\frac{1}{nR^{n-2}}z^n \text{ with } \varepsilon>0. \tag{4.26}$$

That gives the relations

$$\nabla\phi(z)=\frac{z}{2}-\varepsilon\frac{\overline{z}^{n-1}}{R^{n-2}}, \tag{4.27}$$

$$\lambda_1(\phi)=\frac{1}{2}-\varepsilon(n-1). \tag{4.28}$$

Because of Lemma 4.3, we will take $\varepsilon\in[0,\frac{1}{6(n-1)}]$. Moreover, the relation

$$|\nabla\phi(z)|=|\nabla\phi(ze^{2i\pi/n})|$$

allows us to study $\|\nabla\phi\|_\infty$ on only one side side of Ω. We shall treat the case "n odd," $n=2n'+1$; in order to simplify the computations, we will study the problem on the vertical side $I_{n'}:=[\mathrm{Re}^{2in'\pi/n},\mathrm{Re}^{2i(n'+1)\pi/n}]$. When n is even, it is convenient to do a rotation on Ω so that it has a vertical side. Set $z\in I_{n'}$:

$$z=R\cos\frac{2\pi n'}{n}+iy,\ y\in\left[-R\sin\frac{2\pi n'}{n},R\sin\frac{2\pi n'}{n}\right].$$

Denote

$$\rho:=R\cos\frac{2\pi n'}{n}=-R\cos\frac{\pi}{n}, \tag{4.29}$$

$$Y:=y^2 \text{ and } Y_{\max}:=\left(R\sin\frac{2\pi n'}{n}\right)^2=R^2\sin^2\frac{\pi}{n}. \tag{4.30}$$

So we have on $I_{n'}$

$$\|\nabla\phi(z)\|^2=\frac{1}{4}(\rho^2+Y)-\frac{\varepsilon\rho}{R^{n-2}}\sum_{q=0}^{n'}C_n^{2q}(-1)^q\rho^{n-2q-1}Y^q+\left(\frac{\varepsilon}{R^{n-2}}\right)^2(\rho^2+Y)^{n-1}.$$

Denote by $F_\varepsilon(Y)$ this function. In order to calculate $\|\nabla\phi\|_\infty$, we study the variations of F_ε. Let us show F_ε is a convex function on $[0, Y_{\max}]$:

$$F''_\varepsilon(Y) = -\frac{\varepsilon\rho}{R^{n-2}}\sum_{q=0}^{n'} q(q-1)C_n^{2q}(-1)^q\rho^{n-2q-1}Y^{q-2} \\ + (n-1)(n-2)\Big(\frac{\varepsilon}{R^{n-2}}\Big)^2(\rho^2+Y)^{n-3}.$$

- If $n' \leq 2$, F''_ε is clearly nonnegative.
- If $n' \geq 3$, putting the terms two by two, we see that F''_ε is the sum of the terms

$$\begin{aligned} J_{2k}(Y) &= 2k\rho^{n-4k-3}Y^{2k-2}\Big((2k-1)C_n^{4k}\rho^2 - (2k+1)C_n^{4k+2}Y\Big) \\ &= \frac{n!2k\rho^{n-1-4k-2}Y^{2k-2}}{(4k)!(n-4k-2)!}\left(\frac{2k-1}{(n-4k)(n-4k-1)}\rho^2 - \frac{2k+1}{(4k+2)(4k+1)}Y\right). \end{aligned}$$

So it is sufficient to show that each term $J_{2k}(Y)$ is nonnegative. $J_{2k}(Y)$ has the same sign than $V_{2k}(Y) := \Big(\frac{2k-1}{(n-4k)(n-4k-1)}\rho^2 - \frac{2k+1}{(4k+2)(4k+1)}Y\Big)$. By monotonicity,

$$\begin{aligned} V_{2k}(Y) \geq V_2(Y_{\max}) &= \frac{1}{(n-4)(n-5)}\rho^2 - \frac{1}{10}Y_{\max} \\ &\geq \left(\frac{10}{(n-4)(n-5)} - \tan^2\frac{\pi}{n}\right)\frac{\rho^2}{10}, \end{aligned}$$

but for all $n \geq 7$,

$$\tan^2\frac{\pi}{n} \leq \frac{10}{(n-4)(n-5)}.$$

Thus F_ε is convex, and then

$$\|F_\varepsilon\|_\infty = \sup\{F_\varepsilon(0), F_\varepsilon(Y_{\max})\}.$$

That ends the study of F_ε for ε fixed. Next we study the behavior of $\|F_\varepsilon\|_\infty$ with respect to ε:

$$F_\varepsilon(0) = \left(\frac{1}{2}\rho - \varepsilon\frac{\rho^{n-1}}{R^{n-2}}\right)^2 \text{ increases when } \varepsilon \text{ increases, because } \rho < 0;$$

$$F_\varepsilon(Y_{\max}) = |\nabla\phi(Re^{\frac{2i\pi n'}{n}})|^2 = |\nabla\phi(R)|^2 = \left(\frac{1}{2} - \varepsilon\right)^2 R^2 \text{ decreases when } \varepsilon \text{ increases.}$$

In order to minimize $\|\nabla\phi\|_\infty$, we will take, if it is possible, ε such that $F_\varepsilon(0) = F_\varepsilon(Y_{\max})$. That gives the value given in Theorem 4.1: $\varepsilon = \varepsilon_n$. However, we have still to verify that this value belongs to the interval $[0, \frac{1}{6(n-1)}]$ in which we search for ε. Computing the numerical values, we see that

- if $n \leq 8$, $\varepsilon_n > \frac{1}{6(n-1)}$; hence the best value of ε is $\frac{1}{6(n-1)}$, and

$$\omega(\phi) = 6\|\nabla\phi\|_\infty = 6|\nabla\phi(R)| = \left(3 - \frac{1}{n-1}\right)R;$$

- if $n \geq 9$, $\varepsilon_n \leq \frac{1}{6(n-1)}$; hence the best value of ε is ε_n, and

$$\omega(\phi) = (3 - 6\varepsilon_n)R. \qquad \square$$

4.3. Explicit computations: Second study. When Ω is the equilateral triangle, we can go further and study the results given by

$$f(z) = az^3 + bz^6 \quad \text{and} \quad \phi(z) = \frac{1}{4}|z|^2 + \varepsilon \text{ Re } f(z). \tag{4.31}$$

Assume $\nabla P(R) = (-R, 0)$. A careful study of the variations of $G(Y) := |f''(z)|^2$ on the vertical side of Ω shows that

- if $-5\frac{b}{a}R^3 \in [0, \frac{1}{4}]$, G is increasing on $[0, Y_{\max}]$; so the smallest value of f'' is obtained when

$$aR = -\frac{10}{27} \text{ and } bR^4 = \frac{1}{54}; \text{ then } \|f''\|_\infty = f''(R) = \frac{5}{3};$$

- if $-5\frac{b}{a}R^3 \in [\frac{1}{4}, \frac{1}{2}]$, G attains its upper bound only once and in $]0, Y_{\max}[$; we see with numerical values that

$$\text{if } aR \simeq -0,4 \text{ and } bR^3 \simeq 0,033, \text{ then } \|f''\|_\infty \simeq 1,56.$$

Then it is easy to see that $\|\nabla\phi\|$ is an increasing function of Y on $[0, Y_{\max}]$ if $\varepsilon\|f''\|_\infty \leq \frac{1}{6}$. So

$$6\|\nabla\phi\|_\infty = 6|\nabla\phi(R)| = 6\left(\frac{1}{2} - \varepsilon\right) R \leq 2,36R,$$

which proves the estimate (4.11). □

5. Uniform stabilization of Maxwell's equations in almost star-shaped domains.

The notion of "almost star-shaped domains" allows us to extend results of Komornik [9] on the uniform stabilization of Maxwell's equations. Kapitonov [7] also treated this problem in another class of "substarlike domains."

5.1. Statement of the problem and main results. We consider the following evolutionary system:

$$E' - \text{rot } H = 0 = H' + \text{rot } E \text{ in } \Omega \times \mathbb{R}_+, \tag{5.1}$$

$$\text{div } E = 0 = \text{div } H \text{ in } \Omega \times \mathbb{R}_+, \tag{5.2}$$

$$\nu \times (E \times \nu + H) = 0 \text{ on } \Gamma \times \mathbb{R}_+, \tag{5.3}$$

$$E(0) = E^0, H(0) = H^0 \text{ in } \Omega. \tag{5.4}$$

As usual denote the energy space

$$\mathcal{H} = \{(E, H) \in L^2(\Omega)^6, \text{div } E = \text{div } H = 0 \text{ dans } \Omega\}.$$

According to Barucq and Hanouzet [3], we have the following existence theorem.

THEOREM 5.1. *Let Ω be a bounded open domain of $\mathbb{R}^3$ of class $\mathcal{C}^2$.*

1. *Given $(E^0, H^0) \in \mathcal{H}$, there exists a unique weak solution*

$$(E, H) \in \mathcal{C}(\mathbb{R}_+, \mathcal{H}). \tag{5.5}$$

2. *Given $(E^0, H^0) \in \{(E, H) \in H^1(\Omega)^6, \nu \times (E \times \nu + H) = 0$ on $\Gamma\}$, which is a dense subspace of $\mathcal{H}$, there exists a unique strong solution*

$$(E, H) \in \mathcal{C}^1(\mathbb{R}_+, \mathcal{H}) \cap \mathcal{C}(\mathbb{R}_+, H^1(\Omega)^6). \tag{5.6}$$

Define the energy of solutions by

$$\mathcal{E}(t) = \frac{1}{2}\int_\Omega |E(t)|^2 + |H(t)|^2 \; dx = \|(E,H)\|^2_{\mathcal{H}}(t). \tag{5.7}$$

Assume there exists $\phi \in \mathcal{C}^2(\overline{\Omega})$ such that

$$\begin{cases} \Delta\phi = 1 \text{ in } \Omega, \\ \lambda_1(\phi) > \frac{1}{4}, \\ \partial_\nu \phi > 0 \text{ on } \Gamma. \end{cases} \tag{5.8}$$

The method exposed before allows us to construct almost star-shaped domains that are not star-shaped (and are not strictly starlike) that verify (5.8). On such domains we have the following stabilization result.

Theorem 5.2. *Given $(E^0, H^0) \in \mathcal{H}$, the energy of the associated weak solution satisfies the estimate*

$$\forall t \geq 0, \quad \mathcal{E}(t) \leq \mathcal{E}(0)e^{1-\frac{4\lambda_1(\phi)-1}{R(\phi)+\|\nabla\phi\|_\infty}t} \quad \textit{with } R(\phi) = \sup_\Gamma \frac{\|\nabla\phi\|^2_\infty}{\partial_\nu\phi}. \tag{5.9}$$

5.2. Key elements of the proof of Theorem 5.2. We prove (5.9) only for smooth initial data. Thanks to the boundary condition, the energy is nonincreasing and we have the following.

Lemma 5.3.

$$\forall\ 0 < S < T < +\infty: \quad \mathcal{E}(S) - \mathcal{E}(T) = \int_S^T \int_\Gamma |E_\tau|^2 = \int_S^T \int_\Gamma |H_\tau|^2, \tag{5.10}$$

where E_τ and H_τ designate the tangential components of E and H.

The key to the proof of Theorem 5.2 is the following identity.

Lemma 5.4.

$$\int_S^T \int_\Gamma (|E|^2 + |H|^2)\partial_\nu\phi - 2\int_S^T \int_\Gamma (H\cdot\nabla\phi)(H\cdot\nu) + (E\cdot\nabla\phi)(E\cdot\nu) \tag{5.11}$$
$$= -2\left[\int_\Omega (E\times H).\nabla\phi\right]_S^T + \int_S^T \int_\Omega |E|^2 + |H|^2 - 2\Big(D^2\phi\cdot E|E\Big) - 2\Big(D^2\phi\cdot H|H\Big).$$

Proof. Using (5.1) and (5.2), it is easy to verify that

$$\begin{aligned} 2&\left[\int_\Omega (E_1H_2 - E_2H_1)\phi_{,3}\right]_S^T \\ &= -\int_S^T \int_\Gamma (H_1^2 + H_2^2 + H_3^2 + E_1^2 + E_2^2 + E_3^2)\phi_{,3}\nu_3 \\ &\quad + 2\int_S^T \int_\Gamma \phi_{,3}H_3(H_1\nu_1 + H_2\nu_2) + \phi_{,3}E_3(E_1\nu_1 + E_2\nu_2) \\ &\quad + 2\int_S^T \int_\Gamma (H_3^2 + E_3^2)\phi_{,3}\nu_3 + \int_S^T \int_\Omega (H_1^2 + H_2^2 + H_3^2 + E_1^2 + E_2^2 + E_3^2)\phi_{,3,3} \\ &\quad - 2\int_S^T \int_\Omega H_3H_2\phi_{,3,2} + H_3H_1\phi_{,3,1} + E_3E_1\phi_{,3,1} + E_3E_2\phi_{,3,2} + H_3^2\phi_{,3,3} + E_3^2\phi_{,3,3}. \end{aligned}$$

Two similar identities are obtained by cyclical permutation of the indexes 1, 2, and 3. Adding the three identities, we obtain (5.11). Then it is easy to conclude, thanks to (5.8) and to the boundary condition (5.3), that we have

$$\int_\Omega |E|^2+|H|^2-2\Big(D^2\phi\cdot E|E\Big)-2\Big(D^2\phi\cdot H|H\Big) \\ \geq (4\lambda_1(\phi)-1)\int_\Omega |E|^2+|H|^2, \tag{5.12}$$

$$(|E|^2+|H|^2)\partial_\nu\phi-2(E.\nabla\phi)(E.\nu)-2(H.\nabla\phi)(H.\nu) \\ \leq \frac{\|\nabla\phi\|^2}{\partial_\nu\phi}(|E_\tau|^2+|H_\tau|^2). \tag{5.13}$$

Then using the estimates (5.10), (5.12), and (5.13), the identity (5.11) gives

$$\begin{aligned} 2(4\lambda_1(\phi)-1)\int_S^T\mathcal{E} &\leq \int_S^T\int_\Gamma \frac{\|\nabla\phi\|^2}{\partial_\nu\phi}(|E_\tau|^2+|H_\tau|^2)+2\|\nabla\phi\|_\infty(\mathcal{E}(S)+\mathcal{E}(T)) \\ &\leq 2R(\phi)(\mathcal{E}(S)-\mathcal{E}(T))+2\|\nabla\phi\|_\infty(\mathcal{E}(S)+\mathcal{E}(T)). \end{aligned}$$

Since by definition $R(\phi)\geq\|\nabla\phi\|_\infty$, we obtain that

$$(4\lambda_1(\phi)-1)\int_S^T\mathcal{E}\leq\Big(R(\phi)+\|\nabla\phi\|_\infty\Big)\mathcal{E}(S). \tag{5.14}$$

We conclude using the Gronwall-type inequality we used in section 3. □

Acknowledgments. I wish to thank B. Rao for fruitful discussions that led to the construction of Figure 2.1 and V. Komornik for his remarks.

REFERENCES

[1] F. Alabau and V. Komornik, *Boundary observability, controllability, and stabilization of linear elastodynamic systems*, SIAM J. Control Optim., 37 (1998), pp. 521–542.

[2] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[3] H. Barucq and B. Hanouzet, *Etude asymptotique du système de Maxwell avec la condition aux limites absorbantes de Silver-Müller* II, C. R. Acad. Sci. Paris. Sér. I Math., 316 (1993), pp. 547–552.

[4] P. Grisvard, *Contrôlabilité exacte des solutions de l'équation des ondes en présence de singularités*, J. Math. Pures Appl., 68 (1989), pp. 215–259.

[5] A. Haraux, *Semi-groupes linéaires et équations d'évolution linéaires périodiques*, Publication du Laboratoire d'Analyse Numérique no. 78011, Université Pierre et Marie Curie, Paris, 1978.

[6] M. A. Horn, *Implications of sharp trace regularity results on boundary stabilization of the system of linear elasticity*, J. Math. Anal. Appl., 223 (1998), pp. 126–150.

[7] B. V. Kapitonov, *Stabilization and exact boundary controllability for Maxwell's equations*, SIAM J. Control Optim., 32, (1994), pp. 408–420.

[8] V. Komornik, *Rapid boundary stabilization of the wave equation*, SIAM J. Control Optim., 29 (1991), pp. 197–208.

[9] V. Komornik, *Exact Controllability and Stabilization. The Multiplier Method*, John Wiley, Chichester and Masson, Paris, 1994.

[10] V. Komornik and E. Zuazua, *A direct method for the boundary stabilization of the wave equation*, J. Math. Pures Appl., 69 (1990), pp. 33–54.

[11] J. Lagnese, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163–182.

[12] I. Lasiecka and R. Triggiani, *Uniform stabilization of the wave equation with Dirichlet or Neumann feedback control without geometrical conditions*, Appl. Math. Optim., 25 (1992), pp. 189–224.

[13] J. L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués*, Vol. 1, Masson, Paris, 1998.

[14] P. Martinez, *Stabilisation de systèmes distribués semilinéaires: Domaines presque étoilés et inégalités intégrales généralisées,* Thesis, University Louis Pasteur, Strasburg, France, 1998.

[15] M. Moussaoui, *Singularités des solutions du problème mélé, contrôlabilité exacte et stabilisation frontière*, ESAIM Proc., 2, Soc. Math. Appl. Indust., Paris, 1997, pp. 195–201 (electronic).

[16] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations. Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.

[17] L. R. Tcheugoué Tébou, *On the stabilization of the wave and linear elasticity equations in dimension* 2, Panamer. Math. J., 1 (1996), pp. 41–55.

[18] E. Zuazua, *Exponential decay for the semi-linear wave-equation with locally distributed damping*, Comm. Partial Differential Equations, 15 (1990), pp. 205–235.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 3, pp. 695–709

SUPREMUM OPERATORS AND COMPUTATION OF SUPREMAL ELEMENTS IN SYSTEM THEORY*

S. HASHTRUDI ZAD[†], R. H. KWONG[†], AND W. M. WONHAM[†]

Abstract. Constrained supremum and supremum operators are introduced to obtain a general procedure for computing supremal elements of upper semilattices. Examples of such elements include supremal (A, B)-invariant subspaces in linear system theory and supremal controllable sublanguages in discrete-event system theory. For some examples, we show that the algorithms available in the literature are special cases of our procedure. Our iterative algorithms may also provide more insight into applications; in the case of supremal controllable subpredicate, the algorithm enables us to derive a lookahead policy for supervisory control of discrete-event systems.

Key words. discrete-event systems, linear systems, lattice theory, supervisory control, partition, supremal elements, supremum operators

AMS subject classifications. 93B, 68Q20

PII. S0363012997317189

1. Introduction.

In system theory, we sometimes encounter lattice structures [2], [5]. Examples are the lattice of equivalence relations in the theory of sequential machines [9], the lattice of subspaces in geometric control [29], and the lattice of sublanguages in algebraic discrete-event system theory [24]. In these cases, we are usually interested in computing the supremal (or dually, infimal) element of an upper (resp., lower) semilattice: for instance, supremal (A, B)-invariant subspaces [29] or supremal controllable sublanguages [28]. For some of these supremal elements, algorithms are available in the literature. For example, in [14] certain supremal and infimal elements are obtained as extremal solutions to systems of inequalities over the corresponding lattices. In this paper we present a general framework based on *constrained supremum* and *supremum operators* for computing supremal elements, which unifies many of the above-mentioned algorithms, including those of [14]. We will see that previous results are special cases of our general procedure, which we call the Δ-*method.* In some cases, these procedures provide us with more insight into the applications of system theory. For example, in the computation of supremal controllable subpredicate, we are led to a lookahead policy for supervisory control.

The supremum operator (referred to as the maximum operator in [9]) is used in [9] in the computation of the supremal elements of finite lattices. There, the underlying binary relation is assumed to be a pair algebra, and therefore is closed with respect to component-wise join and meet operations. In system theory, however, we are usually interested in the supremal elements of infinite upper semilattices. In these cases, the underlying relation of the supremum operator is not a pair algebra. In this paper, we will generalize the approach of [9] and obtain a procedure suitable for computing supremal elements in system theory. We will also introduce a constrained supremum operator, based on which we will present a second iterative algorithm for computing

*Received by the editors February 24, 1997; accepted for publication (in revised form) October 30, 1997; published electronically February 19, 1999. This research was supported in part by AlliedSignal Aerospace Canada.

http://www.siam.org/journals/sicon/37-3/31718.html

[†]Systems Control Group, Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, M5S 3G4, Canada (hashtrud@control.toronto.edu, stk@control.toronto.edu, wonham@control.toronto.edu).

supremal elements. Although our focus will be on computing supremal elements, dual results for infimal elements will be presented without proof to complete the discussion.

In section 2, we define constrained supremum and supremum operators and derive procedures for computing supremal elements. Next, in section 3, we apply our results to a number of examples in system theory. Dual results for computing infimal elements are given in section 4. Finally, we summarize our work in section 5.

In this paper, the bottom and top elements of lattices are denoted by $\perp$ and $\top$. In a poset L, $x \in L$ is a fixpoint of an operator $\varphi : L \to L$ iff $\varphi(x) = x$, and $\varphi(\cdot)$ is monotone iff $x \leq x'$ implies $\varphi(x) \leq \varphi(x')$ for all x, $x' \in L$. Also $|Q|$ denotes the cardinality of the set Q.

2. Supremum operators.

Throughout this section, $\leq$ denotes a partial order on the corresponding lattice (or poset). For any two elements x and y of a lattice (poset), $x \geq y$ means $y \leq x$. We shall also denote the meet and join operations by $\wedge$ and $\vee$ and use $\vee$ (resp., $\wedge$) and sup (resp., inf) interchangeably.

Consider a complete lattice L. Let $S \subseteq L$ be a complete upper semilattice under the join operation $\vee$ of L. For every $z \in L$, define

$$z^{\uparrow} := \sup\{x \mid x \in S, \quad x \leq z\}.$$

Of course $z^{\uparrow} \in S$. In this section, we will present algorithms for computing $z^{\uparrow}$. Toward this end, the *supremum operators* will be introduced.

We begin by defining *contractive operators.*

DEFINITION 1. *Let L be a poset. An operator $\varphi : L \to L$ is contractive iff $\varphi(x) \leq x$ for all $x \in L$.*

Theorem 1 presents the main algorithm for computing $z^{\uparrow}$.

THEOREM 1. *Let L be a complete lattice, $S \subseteq L$ be an upper semilattice under the join operation $\vee$ of L, and $\varphi : L \to L$ be an operator with the following properties:*

(i) $S = \{x \in L \mid \varphi(x) = x\}$,

(ii) $\varphi(\cdot)$ *is monotone,*

(iii) $\varphi(\cdot)$ *is contractive.*

Let $z \in L$, and suppose there exists an integer $k^* \geq 0$ such that the recursion

$$z_0 = z,$$
$$z_{k+1} = \varphi(z_k)$$

terminates in k^* steps. Then

$$z^{\uparrow} = \sup\{x \mid x \in S, \quad x \leq z\} = z_k \qquad \textit{for all } k \geq k^*.$$

Proof. Since $\varphi(\cdot)$ is contractive, $z_{k+1} = \varphi(z_k) \leq z_k$ for all $k \geq 0$. Therefore $\{z_k\}$ is a nonincreasing sequence. Since $z_{k+1} = z_k$ for all $k \geq k^*$ by assumption, $\varphi(z_{k^*}) = z_{k^*}$; hence $z_{k^*} \in S$. Moreover $z_{k^*} \leq z_0 = z$; therefore $z_{k^*} \in \{x \mid x \in S, \ x \leq z\}$. For any $x \in S$ such that $x \leq z$, if $x \leq z_k$ then $x = \varphi(x) \leq \varphi(z_k) = z_{k+1}$. Therefore, by induction, $x \leq z_k$ for all $k \geq 0$. As a result, z_{k^*} is the supremal element sought: $z^{\uparrow} = z_{k^*}$. □

Remark 1. In the theorem, the sequence $\{z_k\}$ is required to be finite. A poset in which, for any chain $z_0 \geq z_1 \geq \cdots \geq z_k \geq \cdots$ of elements there exists an integer $n \geq 0$ such that $z_k = z_n$ for all $k \geq n$ is said to satisfy the descending chain condition (DCC) [5]. A lattice of finite length, i.e., one in which the length of the longest chain is finite, satisfies the DCC. Graded lattices and finite lattices are of finite length and therefore satisfy the DCC [2]. Clearly, in Theorem 1, satisfaction of the DCC (by L) guarantees finite termination of the recursion. However, if L does not satisfy the DCC, then finite termination has to be established before the theorem can be used. One way to do this is to show that the chain $\{z_k\}$ is contained in a subset of L which satisfies the DCC. Formally we express this simple result as the following lemma.

LEMMA 1. *Suppose the chain* $z_0 \geq z_1 \geq \cdots \geq z_k \geq \cdots$ *of elements of a given poset is contained in a subset of the poset which satisfies the descending chain condition. Then there exists an integer* $n \geq 0$ *such that* $z_k = z_n$ *for all* $k \geq n$. □

In the next section, we shall see an application of this lemma in the computation of supremal controllable sublanguages.

Remark 2. Theorem 1 still holds if L is a complete upper semilattice.

In Theorem 1, $z^{\uparrow}$ is computed as the greatest fixpoint of the contractive monotone operator $\varphi(\cdot)$. By assumption, the set of fixpoints of φ coincides with the upper semisublattice S. We note that the fixpoints of a contractive monotone operator always form a complete upper semilattice (under the join operation of the underlying lattice). To see this, recall that by the Knaster–Tarski fixpoint theorem [26], the set of fixpoints of a monotone operator is nonempty. Now let $\{x_\alpha \mid \alpha \in A\}$ be a subset of the set of fixpoints of a contractive monotone operator φ. Then $\vee_\alpha x_\alpha = \vee_\alpha \varphi(x_\alpha) \leq \varphi(\vee_\alpha x_\alpha)$ because φ is monotone. Also since φ is contractive, $\varphi(\vee_\alpha x_\alpha) \leq \vee_\alpha x_\alpha$. Therefore $\vee_\alpha x_\alpha = \varphi(\vee_\alpha x_\alpha)$; i.e., $\vee_\alpha x_\alpha$ is a fixpoint of φ.

Theorem 1 presents a list of properties for the operator $\varphi(\cdot)$ sufficient to guarantee the convergence of the algorithm, but it does not suggest any particular structure for the operator explicitly. The *constrained supremum operator*, to be introduced shortly, is contractive and, as will be illustrated in the next section, with a suitable choice of its underlying binary relation can be made to satisfy the assumptions of Theorem 1.

DEFINITION 2. *Let* L *be a complete lattice and* $\Delta \subseteq L \times L$. *The constrained supremum operator* $\hat{\Psi} : L \to L$ *is defined as*

$$\hat{\Psi}(y) := \sup\{x \mid (x, y) \in \Delta, \quad x \leq y\}.$$

The symbol $\hat{\Psi}(\cdot)$ does not show the dependency of the constrained supremum operator on its underlying relation Δ. In this paper, no ambiguity will result because we will not work with more than one binary relation at a time. We will denote the supremum operator (Definition 4) and the infimum operators (section 4) following the same convention.

If for some $y \in L$ the set of x's in Definition 2 is empty, then by "empty set logic" $\hat{\Psi}(y) = \bot$.

LEMMA 2. *The constrained supremum operator is contractive.*

Proof. The proof follows immediately from Definition 2. □

We will need the definition of S-reflexive relations.

DEFINITION 3. *Let* L *be a poset and* $S \subseteq L$. *A relation* $\Delta \subseteq L \times L$ *is reflexive with respect to* S *or simply* S*-reflexive iff*

$$L_\Delta := \{x \in L \mid (x, x) \in \Delta\} = S.$$

A reflexive relation on L is obviously L-reflexive too.

Theorem 2 follows from Theorem 1 and presents an iterative procedure for computing $z^{\uparrow}$ in terms of the constrained supremum operator.

THEOREM 2. *Assume that* L *is a complete lattice and* $S \subseteq L$ *is an upper semilattice under the join operation* $\vee$ *of* L. *Let* $\Delta \subseteq L \times L$ *be* S*-reflexive, and assume that*

(i) $L_\Delta = \{x \in L \mid \hat{\Psi}(x) = x\}$,
(ii) $\hat{\Psi}(\cdot)$ *is monotone.*

Let $z \in L$, and suppose there exists an integer $k^* \geq 0$ such that the iteration

$$\begin{aligned} z_0 &= z, \\ z_{k+1} &= \hat{\Psi}(z_k) \end{aligned}$$

terminates in k^* steps. Then

$$z^\uparrow = \sup\{x \mid x \in S, \quad x \leq z\} = z_k \quad \textit{for all } k \geq k^*.$$

Proof. The proof follows from Theorem 1 and Lemma 2. □

Remarks 1 and 2 apply to Theorem 2 as well. Regarding assumption (i) in Theorem 2, we note that $L_\Delta \subseteq \{x \in L \mid \hat{\Psi}(x) = x\}$ is always true but the reverse inclusion $\supseteq$ depends on Δ. Assumptions (i) and (ii) are expressed in terms of $\hat{\Psi}(\cdot)$. As we shall see in the next section, in order to use Theorem 2, we start by defining a suitable Δ, and then we compute $\hat{\Psi}$. Sometimes verifying (i) and (ii) directly in terms of $\hat{\Psi}$ is difficult. The following propositions provide sufficient conditions for the satisfaction of (i) and (ii) in terms of Δ. These conditions are usually easier to check.

Proposition 1. *Let L be a complete lattice, $\Delta \subseteq L \times L$ and $\hat{\Psi}(\cdot)$ the corresponding constrained supremum operator. If, for every set of pairs $\{(x_\alpha, y) \in \Delta \mid \alpha \in A\}$, with A being some index set, we have $(\vee_\alpha x_\alpha, y) \in \Delta$, then*

$$L_\Delta = \{x \in L \mid \hat{\Psi}(x) = x\}.$$

Proof. If $x \in L_\Delta$, then $(x, x) \in \Delta$, therefore $\hat{\Psi}(x) = x$. Conversely, if $x = \hat{\Psi}(x)$, then $(x, x) \in \Delta$ (i.e., $x \in L_\Delta$) because $(\hat{\Psi}(x), x) \in \Delta$ by assumption. □

Proposition 2. *Let L be a complete lattice and $\Delta \subseteq L \times L$. If $(x, y) \in \Delta$ and $y \leq y'$ imply $(x, y') \in \Delta$, then the constrained supremum operator $\hat{\Psi}$ must be monotone.*

Proof. If $y \leq y'$, then it follows from the assumption that

$$\begin{aligned} \{x \mid (x, y) \in \Delta,\ x \leq y\} &\subseteq \{x \mid (x, y') \in \Delta,\ x \leq y\} \\ &\subseteq \{x \mid (x, y') \in \Delta,\ x \leq y'\}; \end{aligned}$$

therefore $\hat{\Psi}(y) \leq \hat{\Psi}(y')$. □

Now we discuss another iterative algorithm which is probably more familiar.

Definition 4. *Let L be a complete lattice and $\Delta \subseteq L \times L$. The supremum operator $\Psi : L \to L$ is defined according to*

$$\Psi(y) := \sup\{x \mid (x, y) \in \Delta\}.$$

If, for some $y \in L$, there exists no $x \in L$ such that $(x, y) \in \Delta$, then $\Psi(y) = \bot$ by "empty set logic."

Theorem 3. *Assume that L is a complete lattice and $S \subseteq L$ is an upper semilattice under the join operation $\vee$ of L. Let $\Delta \subseteq L \times L$ be S-reflexive, and assume that*

(i) $L_\Delta = \{x \in L \mid x \leq \Psi(x)\}$,
(ii) $\Psi(\cdot)$ *is monotone.*

Let $z \in L$ and suppose there exists an integer $k^* \geq 0$ such that the iteration

$$z_0 = z,$$
$$z_{k+1} = \Psi(z_k) \wedge z$$

terminates in k^* steps. Then

$$z^{\uparrow} = \sup\{x \mid x \in S,\ x \leq z\} = z_k \quad \textit{for all } k \geq k^*.$$

Proof. Let $\varphi(x) := \Psi(x) \wedge x$ for $x \in L$. Obviously $\varphi(\cdot)$ is contractive. It follows from the assumptions of the theorem that $\varphi(\cdot)$ is also monotone and L_Δ is its set of fixpoints. Using induction, we can show that with $\varphi(x) = \Psi(x) \wedge x$, the sequence in Theorem 1 is identical to the sequence generated in Theorem 3 and hence terminates in k^* steps. By Theorem 1, $z^{\uparrow} = z_k$ for all $k \geq k^*$. □

Remark 1 applies to Theorem 3 as well.

$L_\Delta \subseteq \{x \in L \mid x \leq \Psi(x)\}$ always holds, but the reverse inclusion $\supseteq$ depends on Δ. The following propositions will help in verifying assumptions (i) and (ii) of Theorem 3.

PROPOSITION 3. *Let L be a complete lattice, $\Delta \subseteq L \times L$, and $\Psi(\cdot)$ be the corresponding supremum operator. Suppose that*

(i) *for every set of pairs $\{(x_\alpha, y) \in \Delta \mid \alpha \in A\}$ with A being some index set, $(\vee_\alpha x_\alpha, y) \in \Delta$; and*

(ii) *$(x, y) \in \Delta$ and $x' \leq x$ imply $(x', y) \in \Delta$.*

Then

$$L_\Delta = \{x \in L \mid x \leq \Psi(x)\}.$$

Proof. If $x \in L_\Delta$, then $(x, x) \in \Delta$; therefore $x \leq \Psi(x)$. Conversely, let $x \leq \Psi(x)$. It follows from (i) that $(\Psi(x), x) \in \Delta$. Hence by (ii), $(x, x) \in \Delta$; i.e., $x \in L_\Delta$. □

PROPOSITION 4. *Let L be a complete lattice and $\Delta \subseteq L \times L$. If $(x, y) \in \Delta$ and $y \leq y'$ imply $(x, y') \in \Delta$, then the supremum operator $\Psi(\cdot)$ must be monotone.*

Proof. If $y \leq y'$, then $\{x \mid (x, y) \in \Delta\} \subseteq \{x \mid (x, y') \in \Delta\}$. Therefore $\Psi(y) \leq \Psi(y')$. □

Remark 3. In Theorems 2 and 3, one "sup" operation (over L_Δ) is replaced with another one (over x's satisfying $(x, z_k) \in \Delta$). As we shall see in the following section, the latter operation is usually easier to do. The usefulness of Theorems 2 and 3 is due to this fact.

Remark 4. Theorems 1, 2, and 3 compute the supremal element as the greatest fixpoint of monotone operators. This is a standard technique (see, e.g., [22], [27]) which relies on the Knaster–Tarski fixpoint theorem [26], [15]. In this paper we are presenting *a method for obtaining* these monotone operators, i.e., defining a suitable S-reflexive relation Δ and computing $\hat{\Psi}(\cdot)$ or $\Psi(\cdot)$. We shall refer to this method as the Δ-method. Examples given in the following section will show that the method is fairly general. Note that in this paper no systematic way for defining a suitable relation Δ is provided; however, the next section will demonstrate how such a relation can be chosen in a number of cases.

3. Applications.

In this section, we will apply the Δ-method to some examples in system theory. In each example, the first step is to identify the lattice L, then define an appropriate S-reflexive relation Δ. Next we have to compute the (constrained) supremum operator. If L does not satisfy the DCC, we also have to see if the iteration terminates in a finite number of steps.

Example 1. *Extremal solutions of inequalities.* In [14], the existence and computation of extremal solutions of inequalities over lattices are studied. Based on this, algorithms for the computation of some of the supremal and infimal elements in supervisory control of discrete-event systems are obtained. Specifically the following result is established.

THEOREM 4 (see [14]). *Consider the system of inequalities* $\{f_i(x) \leq g_i(x)\}_{1\leq i\leq n}$ *over a complete lattice* L, *with* f_i*'s being disjunctive and* g_i*'s monotone. Let* $f_i^{\perp}$ *denote the dual of* f_i, *and define* $h_1 : L \rightarrow L$ *as*

$$h_1(x) := \bigwedge_{i=1}^{n} f_i^{\perp}(g_i(x)).$$

If the iteration

$$z_0 = \top,$$
$$z_{k+1} = h_1(z_k)$$

terminates in a finite number of steps—say k^**—then*

$$z_{k^*} = \sup\{x \mid \textit{for all } i \leq n : f_i(x) \leq g_i(x)\}. \qquad \square$$

In a complete lattice L, the dual of a function $f : L \rightarrow L$ has been defined to be a function $f^{\perp} : L \rightarrow L$, with $f^{\perp}(y) = \sup\{x \in L \mid f(x) \leq y\}$ for every $y \in L$ [14]. Moreover $f : L \rightarrow L$ is called disjunctive iff for every subset $\{x_\alpha \in L \mid \alpha \in A\}$, with A being an index set, $f(\vee_\alpha x_\alpha) = \vee_\alpha f(x_\alpha)$ [7], [14]. Note that f disjunctive implies f monotone.

The remaining results in [14] are derived from the above theorem. We shall show that Theorem 4 and hence the resulting algorithms of [14] are special cases of Theorem 3.

Let $S = \{x \mid \text{for all } i \leq n : f_i(x) \leq g_i(x)\}$. Define $\Delta \subseteq L \times L$ to be

$$\Delta = \{(x, y) \mid x, y \in L \text{ and for all } i \leq n : f_i(x) \leq g_i(y)\}.$$

For every $x \in L$, $(x, x) \in \Delta$ iff x is a solution of $\{f_i(x) \leq g_i(x)\}_{1\leq i\leq n}$; therefore Δ is S-reflexive and

$$\sup\{x \mid \text{for all } i \leq n : f_i(x) \leq g_i(x)\} = \sup\{x \mid x \in L_\Delta\}.$$

For a set $\{(x_\alpha, y) \in \Delta \mid \alpha \in A\}$

$$\begin{aligned} f_i\left(\bigvee_\alpha x_\alpha\right) &= \bigvee_\alpha f_i(x_\alpha) \quad (f_i \text{ is disjunctive}) \\ &\leq g_i(y). \end{aligned}$$

Therefore $(\vee_\alpha x_\alpha, y) \in \Delta$. If $(x, y) \in \Delta$ and $x' \leq x$, then $f_i(x') \leq f_i(x) \leq g_i(y)$ for all $i \leq n$; hence $(x', y) \in \Delta$. Also $(x, y) \in \Delta$ and $y \leq y'$ imply $f_i(x) \leq g_i(y) \leq g_i(y')$ for all $i \leq n$, which in turn implies $(x, y') \in \Delta$.

By Definition 4

$$\begin{aligned} \Psi(y) &= \sup\{x \mid \text{for all } i \leq n : f_i(x) \leq g_i(y)\} \\ &= \sup\{x \mid x \leq h_1(y)\} \quad \text{(by Lemma 5 of [14])} \\ &= h_1(y). \end{aligned}$$

Now it is easy to see that Theorem 4 follows from Theorem 3 and Propositions 3 and 4.

The framework presented in this paper is more general than the one given in [14]. To see this, consider the lattice $L = \{\bot, x_1, x_2, \top\}$ with $\bot < x_1 < x_2 < \top$. Define functions f and g according to $f(\bot) = \bot$, $f(x_1) = \top$, $f(x_2) = \top$, $f(\top) = x_2$, $g(\bot) = \bot$, $g(x_1) = x_1$, $g(x_2) = x_2$, $g(\top) = x_2$. Consider the problem of finding $\sup\{x \in L \mid f(x) \leq g(x)\}$. Theorem 4 does not apply to this case, because f is not disjunctive. However, it is easy to see that the iteration in Theorem 2 with $\Delta = \{(x, y) \in L \times L \mid f(x) \leq g(y)\}$ will give us the supremal solution of the inequality. Note that there are problems that can be solved using our approach but cannot be (easily) cast into a problem of solving a system of inequalities. The "relational coarsest partition" problem [11], [23] is one example, to be discussed later in this paper.

Example 2. *Supremal* (A, B)*-invariant subspace.* Consider the finite-dimensional linear time-invariant system

$$\dot{x}(t) = Ax(t) + Bu(t),$$

where $x(t) \in \mathcal{X}$. A subspace $\mathcal{V} \subseteq \mathcal{X}$ is called (A, B)-invariant iff $A\mathcal{V} \subseteq \mathcal{V} + \mathrm{Im}B$ [29], where $\mathrm{Im}B$ is the subspace spanned by the columns of B. For every $\mathcal{K} \subseteq \mathcal{X}$, there exists a supremal (A, B)-invariant subspace $\mathcal{K}^{\uparrow}$ contained in $\mathcal{K}(\mathcal{K}^{\uparrow} \subseteq \mathcal{K})$ [29]. A recursion for computing $\mathcal{K}^{\uparrow}$ can be derived as follows.

The set of subspaces $\mathcal{V} \subseteq \mathcal{X}$ partially ordered by subspace inclusion $\subseteq$ and under operations sum $+$ and intersection $\cap$ of subspaces forms a complete lattice; call it L. In this case $\bot = 0$ and $\top = \mathcal{X}$. Let S be the set of (A, B)-invariant subspaces. Define $\Delta \subseteq L \times L$ as follows:

$$\Delta = \{(\mathcal{V}_1, \mathcal{V}_2) \mid \mathcal{V}_1, \mathcal{V}_2 \subseteq \mathcal{X} \text{ and } A\mathcal{V}_1 \subseteq \mathcal{V}_2 + \mathrm{Im}B\}.$$

For every $\mathcal{V} \subseteq \mathcal{X}$, $(\mathcal{V}, \mathcal{V}) \in \Delta$ iff $\mathcal{V}$ is (A, B)-invariant; hence Δ is S-reflexive and

$$\mathcal{K}^{\uparrow} = \sup\{\mathcal{V} \mid \mathcal{V} \in L_{\Delta}, \quad \mathcal{V} \subseteq \mathcal{K}\}.$$

Moreover Δ satisfies the assumptions of Propositions 3 and 4. Define $A^{-1}\mathcal{V} := \{x \in \mathcal{X} \mid Ax \in \mathcal{V}\}$ and observe that $(A^{-1}(\mathcal{V}+\mathrm{Im}B), \mathcal{V}) \in \Delta$ for any $\mathcal{V} \subseteq \mathcal{X}$ since $A(A^{-1}(\mathcal{V}+\mathrm{Im}B)) \subseteq \mathcal{V}+\mathrm{Im}B$. Also if for some $\mathcal{V}_1 \subseteq \mathcal{X}$ we have $(\mathcal{V}_1, \mathcal{V}) \in \Delta$, then $A\mathcal{V}_1 \subseteq \mathcal{V}+\mathrm{Im}B$ and hence $\mathcal{V}_1 \subseteq A^{-1}(\mathcal{V} + \mathrm{Im}B)$. Therefore $\Psi(\mathcal{V}) = A^{-1}(\mathcal{V} + \mathrm{Im}B)$. The lattice L is not finite, but it is of finite length (in fact, graded) with $d(\mathcal{V})$, denoting dimension of $\mathcal{V} \subseteq \mathcal{X}$, as the height function (in a poset of finite length, the height of an element is the least upper bound of the lengths of the chains that start at $\bot$ and end at the element) [2]. Therefore by Theorem 3 and Propositions 3 and 4

$$\begin{aligned}\mathcal{K}_0 &= \mathcal{K},\\ \mathcal{K}_{k+1} &= \mathcal{K} \cap A^{-1}(\mathcal{K}_k + \mathrm{Im}B),\\ \mathcal{K}^{\uparrow} &= \mathcal{K}_k, \qquad k \geq d(\mathcal{K}).\end{aligned}$$

This is a well-known result [29].

Similarly, Theorem 2 gives the following iteration:

$$\begin{aligned}\mathcal{K}_0 &= \mathcal{K},\\ \mathcal{K}_{k+1} &= \hat{\Psi}(\mathcal{K}_k) = \mathcal{K}_k \cap A^{-1}(\mathcal{K}_k + \mathrm{Im}B),\\ \mathcal{K}^{\uparrow} &= \mathcal{K}_k, \qquad k \geq d(\mathcal{K}).\end{aligned}$$

Example 3. *Supremal controllable sublanguage.* Let Σ be a nonempty finite alphabet of elements and Σ^* be the set of all finite strings of elements of Σ, including the empty string. A subset $K \subseteq \Sigma^*$ is called a language. Let N be a fixed language and Σ_u be a fixed subset of Σ. A language $K \subseteq \Sigma^*$ is controllable (with respect to N and Σ_u) iff $\overline{K}\Sigma_u \cap N \subseteq \overline{K}$, where $\overline{K}$ is the prefix closure of K [24]. Every language E has a supremal controllable sublanguage, denoted here by $E^{\uparrow}$ [24].

Let $L = \text{Pwr}(\Sigma^*)$ be the power set of Σ^*. Then L, partially ordered by set inclusion $\subseteq$ and under the operations union $\cup$ and intersection $\cap$ of sets, forms a complete lattice, with $\bot = \emptyset$ and $\top = \Sigma^*$. Let S be the set of controllable sublanguages. Define $\Delta \subseteq L \times L$ to be

$$\Delta = \{(K_1, K_2) \mid K_1, K_2 \in L \text{ and } \overline{K}_1\Sigma_u \cap N \subseteq \overline{K}_2\}.$$

For every $K \in L$, $(K, K) \in \Delta$ iff K is controllable; as a result Δ is S-reflexive and

$$E^{\uparrow} = \sup\{K \mid K \in L_\Delta, \quad K \subseteq E\}.$$

We see that for the set $\{(K_1^\alpha, K_2) \in \Delta \mid \alpha \in A\}$

$$\overline{\bigcup_\alpha K_1^\alpha}\Sigma_u \cap N = \bigcup_\alpha (\overline{K_1^\alpha}\Sigma_u \cap N) \subseteq \overline{K}_2;$$

hence $(\cup_\alpha K_1^\alpha, K_2) \in \Delta$. Also if $(K_1, K_2) \in \Delta$ and $K_1' \subseteq K_1$, then

$$\overline{K_1'}\Sigma_u \cap N \subseteq \overline{K}_1\Sigma_u \cap N \subseteq \overline{K}_2;$$

i.e., $(K_1', K_2) \in \Delta$. Therefore by Proposition 3, $L_\Delta = \{x \in L \mid x \leq \Psi(x)\}$. Similarly, using Proposition 4, we can show that $\Psi(\cdot)$ is monotone. According to Definition 4, for $K \in L$,

$$\Psi(K) = \sup\{D \mid \overline{D}\Sigma_u \cap N \subseteq \overline{K}\}.$$

The iteration in Theorem 3 becomes

$$\begin{aligned} E_0 &= E, \\ E_{k+1} &= E \cap \sup\{D \mid \overline{D}\Sigma_u \cap N \subseteq \overline{E}_k\}, \end{aligned}$$

which, when E and N are regular languages, is known to converge in a finite number of steps [28]—say, k^*. Hence by Theorem 3: $E^{\uparrow} = E_k$ for all $k \geq k^*$. Here the lattice L does not satisfy the DCC. However, when E and N are regular over the fixed, finite alphabet Σ, the languages E_k belong to the finite set $\{K \mid \|K\| \leq \|E\| \cdot \|N\| + 1\}$ [28]. Here $\|K\|$ denotes the minimal cardinality of the state set of an automaton that "recognizes" K [10]. Therefore, finite termination of the recursion follows from Lemma 1.

Example 4. *Supremal normal sublanguage.* With Σ and Σ^* as in Example 3, let K be a fixed language. Suppose that Σ_o is a fixed subset of Σ, and let $P : \Sigma^* \to \Sigma_o^*$ denote the natural projection. A language $N \subseteq K$ is (K, P)-normal iff $K \cap P^{-1}(PN) = N$ [20], [21]. Note that $N \subseteq K \cap P^{-1}(PN)$ always holds. Every language $E \subseteq K$ has a unique supremal normal sublanguage, denoted here by E^* [21].

Let L be the set of sublanguages of K; i.e., $L := \{N \mid N \subseteq K \subseteq \Sigma^*\}$. Then L is partially ordered by set inclusion $\subseteq$ and, under the operations union $\cup$ and intersection $\cap$ of sets, forms a complete lattice, with $\bot = \emptyset$ and $\top = K$. Let S be the set of (K, P)-normal languages. Now define

$$\Delta = \{(N_1, N_2) \mid N_1, N_2 \in L \text{ and } K \cap P^{-1}(PN_1) \subseteq N_2\}.$$

For every $N \in L$, $(N, N) \in \Delta$ iff N is (K, P)-normal; hence Δ is S-reflexive and

$$E^* = \sup\{N \mid N \in L_\Delta, \quad N \subseteq E\}.$$

Using Propositions 1 and 2 it is straightforward to verify that $\hat{\Psi}(\cdot)$ satisfies assumptions (i) and (ii) of Theorem 2. In this case the constrained supremum operator is

$$\begin{aligned}\hat{\Psi}(N) &= \sup\{N' \subseteq K \mid K \cap P^{-1}(PN') \subseteq N, \quad N' \subseteq N \subseteq K\}\\ &= \{x \in N \mid K \cap P^{-1}(Px) \in N \subseteq K\}.\end{aligned}$$

Using Figure 1 with

$$X = (K \cap P^{-1}(PN)) - N$$

we can see that $N \cap P^{-1}(PX)$ contains those strings $x \in N$ that violate $K \cap P^{-1}(Px) \in N$. Therefore

$$\begin{aligned}\hat{\Psi}(N) &= N - P^{-1}(PX)\\ &= N - P^{-1}P(K - N).\end{aligned}$$

The last equality holds because $P(K - (X \cup N)) \cap P(N) = \emptyset$. The recursion in Theorem 2 will be

$$\begin{aligned}E_0 &= E,\\ E_{k+1} &= E_k - P^{-1}P(K - E_k).\end{aligned}$$

The above iteration converges in one step since $E^* = E - P^{-1}P(K - E)$ [12]. We note that in this example the lattice L does not satisfy the DCC. If K and E are closed and L contains only the closed sublanguages of K, then the iteration becomes

$$\begin{aligned}E_0 &= E,\\ E_{k+1} &= E_k - P^{-1}P(K - E_k)\Sigma^*\end{aligned}$$

to keep E_k's closed. The above recursion also converges in only one step because $E^* = E - P^{-1}P(K - E)\Sigma^*$ [3].

Example 5. *Supremal controllable subpredicate.* Predicates and predicate transformers [6] are used in state-based control of discrete-event systems [25]. For other examples refer to [18] and [13] and their references. Using the notation in [18], let Q be the set of states of a discrete-event system G. A predicate P is controllable (with respect to G) iff $P \preceq M_\sigma(P)$ for all $\sigma \in \Sigma_u$ and $P \preceq R(G, P)$ [17], where $\preceq$ means "implies," Σ_u is the set of uncontrollable events, $M_\sigma(P)$ the "weakest liberal precondition" of P, and $R(G, P)$ the "reachability predicate." For every predicate P, there exists a supremal controllable subpredicate $P^\uparrow(\preceq P)$ [18].

Let L be the family of predicates on Q. Then L is partially ordered by $\preceq$ and under operations disjunction $\vee$ and conjunction $\wedge$ of predicates forms a complete

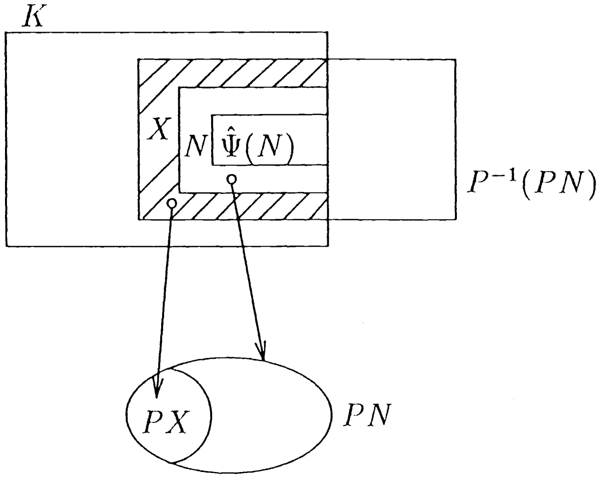


FIG. 1. *Computing* $\hat{\Psi}(N)$.

lattice. Here $\perp =$FALSE and $\top =$TRUE. Let S be the set of controllable predicates. If $\Delta \subseteq L \times L$ is defined as

$$\Delta = \{(P_1, P_2) \mid P_1, P_2 \in L \text{ and } (P_1 \preceq M_\sigma(P_2) \text{ for all } \sigma \in \Sigma_u) \text{ and } P_1 \preceq R(G, P_2)\},$$

then for every predicate P, $(P, P) \in \Delta$ iff P is controllable; therefore Δ is S-reflexive and

$$P^\uparrow = \sup\{P' \mid P' \in L_\Delta, \quad P' \preceq P\}.$$

Moreover, using the fact that $M_\sigma(\cdot)$ and $R(G, \cdot)$ are monotone, along with Propositions 1 and 2, it can be shown that Δ satisfies assumptions (i) and (ii) of Theorem 2. Theorem 2 provides the following recursion, which converges in a finite number of steps when the number of states is finite, $|Q| < \infty$:

$$\begin{aligned} P_0 &= P, \\ P_{k+1} &= \sup\{P' \mid P' \preceq M_\sigma(P_k) \text{ for all } \sigma \in \Sigma_u, \quad P' \preceq R(G, P_k)\} \\ &= R(G, P_k) \bigwedge \left(\bigwedge_{\sigma \in \Sigma_u} M_\sigma(P_k) \right), \\ P^\uparrow &= P_k, \qquad k \geq |P|. \end{aligned}$$

In the above recursion, $|P| := |\{q \in Q \mid q \in P\}|$. Computing $R(G, P_k)$ is computationally complex. The above iteration can be replaced with the simpler recursion:

$$\begin{aligned} P_0 &= P, \\ P_{k+1} &= \sup\{P' \mid P' \preceq M_\sigma(P_k) \text{ for all } \sigma \in \Sigma_u, \quad P' \preceq P_k\} \\ &= P_k \bigwedge \left(\bigwedge_{\sigma \in \Sigma_u} M_\sigma(P_k) \right), \\ P^* &= P_k \quad \text{for all } k \geq k^*, \\ P^\uparrow &= R(G, P^*). \end{aligned}$$

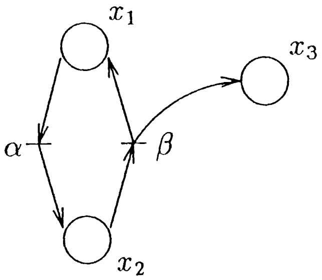


FIG. 2. *Petri net of the small factory.*

Here $k^* \leq |P|$. To verify the above iteration, we have to note only that [25] $P^* = [P]$ (the weakest subpredicate of P that is invariant under the flow induced by uncontrollable events) and use Proposition 8 of [18].

For the purpose of supervisory control, only [P] has to be computed and $P^{\uparrow}$ is not required [19]. We claim that knowing $[P]$ is not always necessary; knowing k^* is enough. To see this, recall that the above recursion for computing $[P]$ ends after k^* steps. This means that if a state $q \in P$, then either $q \in [P]$ or $q \notin [P]$ and there exists a string s of uncontrollable events of length less than or equal to k^* such that $\delta(q, s) \notin P$ (δ is the transition function of the discrete-event system). Therefore, to check whether $q \in [P]$ at any state q, we need only examine strings of uncontrollable events of length less than or equal to k^*. This is a lookahead policy. (See [4] and its references for other approaches to and applications of lookahead policies.) If k^* is not large, this forward search can be done on-line. Note that computation of k^* is done *off-line*. When the number of states, $|Q|$, is not finite, e.g., in vector discrete-event systems (VDES), k^* may still be finite. For VDES, k^* (if finite) may be calculated *off-line* using integer linear programming *without* the burden of computing $[P]$. The details are left for future research.

To illustrate the above point, consider a small factory consisting of 2 machines and a buffer with capacity 3. The Petri net of the factory is shown in Figure 2. Here x_1 is the number of idle machines, x_2 is the number of working machines, and x_3 is the number of workpieces produced. Event α ("take a workpiece") is controllable, while β ("one workpiece done") is uncontrollable. Since $x_1 + x_2 = 2$, the state of the system is described by two variables; we take $x := [x_2, x_3]$. The state graph of the system is shown in Figure 3. Horizontal transitions are α and diagonal ones β. The objective of supervision is to prevent buffer overflow; i.e., $x \in P$ iff $0 \leq x_3 \leq 3$ and $0 \leq x_2 \leq 2$. We assume initially $x_2 = x_3 = 0$. The sequence P_k can be obtained by inspection and is shown in Figure 3. In this figure, predicate P_k holds on those states that are below its corresponding dashed line. We see that $k^* = 2$, so to check whether $x \in [P]$, one has to examine strings β and $\beta\beta$. In general with n machines and a buffer capacity b, $k^* = \min(n, b)$.

Example 6. *The relational coarsest partition problem (RCP).* Consider a finite-state labeled transition system [1] $\mathcal{S} = (Q, \Sigma, R, q_0)$, where Q is the set of states and Σ the set of elementary actions (events). For simplicity we assume $\Sigma = \{\alpha\}$, $|\Sigma| = 1$. R is a binary relation on Q such that $(q, q') \in R$ iff $q' \in \alpha(q)$ with $\alpha(q)$ denoting the set of states reachable from q via a single α transition. Let $\pi = \{B_1, \ldots, B_p\}$ be a partition of Q, with B_i denoting the blocks of π. Then π is compatible with R [8] iff

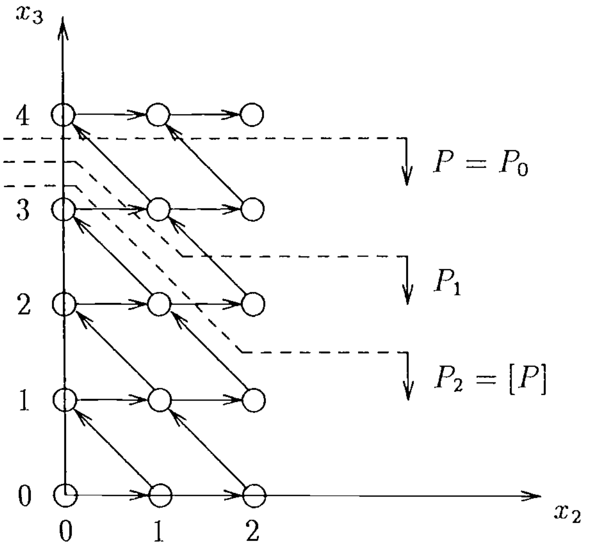


FIG. 3. *State graph.*

whenever q, q' are in the same block B_i, for any block B_j,

$$\alpha(q) \cap B_j \neq \emptyset \Leftrightarrow \alpha(q') \cap B_j \neq \emptyset.$$

Equivalently π is compatible with R iff for each pair of blocks B_i and B_j, either $B_i \subseteq R^{-1}(B_j)$ or $B_i \cap R^{-1}(B_j) = \emptyset$ [8]. Given a partition ρ, the RCP is to find the coarsest partition $\rho^\uparrow$ which is finer than ρ and compatible with R [11], [23]. For applications of the RCP in observer theory and the study of equivalence of labeled transition systems, the reader may refer to [1] and [16] and their references.

Let L be the set of partitions of (or equivalence relations on) Q. For two partitions π_1 and π_2, let $\pi_1 \leq \pi_2$ iff for every pair of blocks $B_1 \in \pi_1$ and $B_2 \in \pi_2$, either $B_1 \subseteq B_2$ or $B_1 \cap B_2 = \emptyset$. Then L partially ordered by $\leq$ forms a complete lattice [2]. In this lattice $\perp = \{\{q\} \mid q \in Q\}$ and $\top = \{Q\}$. Let S be the set of partitions compatible with R. Now we define $\Delta \subseteq L \times L$ to be

$$\begin{aligned}\Delta = \{(\pi_1, \pi_2) \mid \pi_1, \pi_2 \in L \text{ and}\\ ((\text{for all } B_1 \in \pi_1, \text{ for all } B_2 \in \pi_2)\ B_1 \subseteq R^{-1}(B_2) \quad \text{or} \quad B_1 \cap R^{-1}(B_2) = \emptyset)\}.\end{aligned}$$

Note that $(\pi_1, \pi_2) \in \Delta$ iff whenever q and q' are in the same block $B_1 \in \pi_1$, then for any block $B_2 \in \pi_2$,

$$\alpha(q) \cap B_2 \neq \emptyset \quad \Leftrightarrow \quad \alpha(q') \cap B_2 \neq \emptyset.$$

Δ is the generalization of the set of *partition pairs* of [9] to nondeterministic systems. When $\mathcal{S}$ is deterministic, Δ is a pair algebra [9]. In our case, $\mathcal{S}$ can be nondeterministic; therefore Δ is not a pair algebra in general. For example, consider $Q = \{a, b, c, d\}$ and $\alpha(a) = \{a, c\}$, $\alpha(b) = \{b, d\}$, $\alpha(c) = \alpha(d) = \emptyset$. Let $\pi_1 = \pi_1' = \{\{a, b\}, \{c, d\}\}$, $\pi_2 = \{\{a, b\}, \{c, d\}\}$, $\pi_2' = \{\{a, d\}, \{b, c\}\}$. Then $(\pi_1, \pi_2) \in \Delta$, $(\pi_1', \pi_2') \in \Delta$, but $(\pi_1 \wedge \pi_1', \pi_2 \wedge \pi_2') \notin \Delta$. Obviously π is compatible with R iff $(\pi, \pi) \in \Delta$; therefore Δ is S-reflexive and

$$\rho^\uparrow = \sup\{\pi \mid \pi \in L_\Delta, \quad \pi \leq \rho\}.$$

Using Propositions 1 and 2, the reader can show that Δ satisfies the assumptions (i) and (ii) of Theorem 2.

By Definition 2, for every $\rho \in L$,

$$\hat{\Psi}(\rho) = \sup\{\pi \mid (\pi, \rho) \in \Delta, \;\; \pi \leq \rho\}.$$

Let $\rho = \{B_1, \ldots, B_m\}$. Following [23], for a partition $\pi \in L$ and a subset $X \subseteq Q$, let $\mathit{split}(X, \pi)$ be the refinement of π obtained by replacing each block $B \in \pi$ such that $B \cap R^{-1}(X) \neq \emptyset$ and $B - R^{-1}(X) \neq \emptyset$, with the two blocks $B' = B \cap R^{-1}(X)$ and $B'' = B - R^{-1}(X)$. The reader can now verify that the following iteration computes $\hat{\Psi}(\rho)$:

$$\begin{aligned} \pi_0 &= \rho, \\ \pi_k &= \mathit{split}(B_k, \pi_{k-1}), \qquad 1 \leq k \leq m, \\ \hat{\Psi}(\rho) &= \pi_m. \end{aligned}$$

Therefore, using Theorem 2, the relational coarsest partition problem can be solved with the following recursion:

$$\begin{aligned} \rho_0 &= \rho, \\ \rho_{k+1} &= \hat{\Psi}(\rho_k), \\ \rho^{\uparrow} &= \rho_k \quad \text{for all } k \geq |Q| - 1. \end{aligned}$$

This procedure can be generalized to the case $|\Sigma| > 1$. The resulting algorithm turns out to be the one given in Proposition 3.10 of [8].

4. Dual results. Dual results for computing infimal elements are presented here. The proofs are similar to those in section 2 and are omitted.

Given a poset L, an operator $\varphi : L \to L$ is called *expansive* iff $x \leq \varphi(x)$ for all $x \in L$.

THEOREM 5. *Let L be a complete lattice, $S \subseteq L$ be a lower semilattice under the meet operation $\wedge$ of L, and $\varphi : L \to L$ be an operator with the following properties:*

(i) $S = \{x \in L \mid \varphi(x) = x\}$,

(ii) *$\varphi(\cdot)$ is monotone,*

(iii) *$\varphi(\cdot)$ is expansive.*

Let $z \in L$, and suppose that there exists an integer $k^* \geq 0$ such that $\varphi^{k^*}(z) = \varphi^{k^*+1}(z)$. Then

$$z^{\downarrow} := \inf\{x \mid x \in S, \; z \leq x\} = \varphi^k(z) \quad \text{for all } k \geq k^*. \qquad \square$$

For a binary relation $\Delta \subseteq L \times L$, the *constrained infimum operator* is defined as

$$\hat{\psi}(x) := \inf\{y \mid (x, y) \in \Delta, \quad x \leq y\}.$$

This operator is expansive. If Δ is S-reflexive and $\hat{\psi}(\cdot)$ satisfies properties (i) and (ii) of Theorem 5, then the iteration of the theorem can be used for computing $z^{\downarrow}$. If for every set of pairs $\{(x, y_\alpha) \in \Delta \mid \alpha \in A\}$, with A being some index set, we have $(x, \wedge_\alpha y_\alpha) \in \Delta$, then $L_\Delta = \{x \mid \hat{\psi}(x) = x\}$. Also, if $(x, y) \in \Delta$ and $x' \leq x$ imply $(x', y) \in \Delta$, then $\hat{\psi}(\cdot)$ is monotone.

Let the *infimum operator* corresponding to a relation $\Delta \subseteq L \times L$ be

$$\psi(x) := \inf\{y \mid (x, y) \in \Delta\}.$$

Assume that Δ is S-reflexive, $L_\Delta = \{x \in L \mid \psi(x) \leq x\}$, and $\psi(\cdot)$ is monotone. If the iteration

$$z_0 = z,$$
$$z_{k+1} = \psi(z_k) \vee z$$

terminates in a finite number of steps—say, k^*—then

$$z^{\downarrow} = \inf\{x \mid x \in S, \quad z \leq x\} = z_k \quad \text{for all } k \geq k^*.$$

If (i) for every set of pairs $\{(x, y_\alpha) \in \Delta \mid \alpha \in A\}$, with A being some index set, $(x, \wedge_\alpha y_\alpha) \in \Delta$, and (ii) $(x, y) \in \Delta$ and $y \leq y'$ imply $(x, y') \in \Delta$, then $L_\Delta = \{x \in L \mid \psi(x) \leq x\}$. If $(x, y) \in \Delta$ and $x' \leq x$ imply $(x', y) \in \Delta$, then $\psi(\cdot)$ is monotone.

5. Conclusion. In this paper, we introduced constrained supremum and supremum operators to obtain a general procedure, the Δ-method, for computing supremal elements of upper semilattices. These elements are used in system theory, and for some of them, specific algorithms are given in the literature. We applied our procedure to some well-known examples, and in all cases the algorithms available in the literature turned out to be instances of our procedure. These iterations are also informative from a practical point of view. For instance, in the case of supremal controllable subpredicate, the recursion resulted in a lookahead policy for supervisory control in which the required length of the lookahead window was equal to the number of the recursion steps.

Dual results for infimal elements of lower semilattices were presented for completeness.

Defining a suitable binary relation on the lattice is an important step in the procedure. This is not always easy. Also for lattices that do not satisfy the descending chain condition, establishing finite termination of the iteration is not necessarily trivial; see, e.g., [28]. Nevertheless, the procedure covers a variety of cases in control theory, especially in discrete-event system theory.

REFERENCES

[1] A. Arnold, *Finite Transition Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1994.

[2] G. Birkhoff, *Lattice Theory*, American Mathematical Society, Providence, RI, 1967.

[3] R.D. Brandt, V. Garg, R. Kumar, F. Lin, S.I. Marcus, and W.M. Wonham, *Formulas for calculating supremal controllable and normal sublanguages*, Systems Control Lett., 15 (1990), pp. 111–117.

[4] S.L. Chung, S. Lafortune, and F. Lin, *Limited lookahead policies in supervisory control of discrete event systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 1921–1935.

[5] B.A. Davey and H.A. Priestly, *Introduction to Lattices and Order*, Cambridge University Press, Cambridge, UK, 1990.

[6] E.W. Dijkstra, *A Discipline of Programming*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[7] E.W. Dijkstra and C.S. Scholten, *Predicate Calculus and Program Semantics*, Springer-Verlag, New York, 1990.

[8] J.-C. Fernandez, *An implementation of an efficient algorithm for bisimulation equivalence*, Sci. Comput. Programming, 13 (1989/90), pp. 219–236.

[9] J. Hartmanis and R.E. Stearns, *Algebraic Structure Theory of Sequential Machines*, Prentice-Hall, Englewood Cliffs, NJ, 1966.

[10] J.E. Hopcroft and J.D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, Reading, MA, 1979.

[11] P.C. Kanellakis and S.A. Smolka, *CCS expressions, finite state processes, and three problems of equivalence*, in Proc. 2nd ACM Symposium on the Principles of Distributed Computing, 1983, pp. 228–240.

[12] R. Kumar, V. Garg, and S.I. Marcus, *On controllability and normality of discrete event dynamical systems*, Systems Control Lett., 17 (1991), pp. 157–168.

[13] R. Kumar, V. Garg, and S.I. Marcus, *Predicates and predicate transformers for supervisory control of discrete event dynamical systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 232–247.

[14] R. Kumar and V.K. Garg, *Extremal solutions of inequations over lattices with applications to supervisory control*, Theoret. Comput. Sci., 148 (1995), pp. 67–92.

[15] J.-L. Lassez, V.L. Nguyen, and E.A. Sonenberg, *Fixed point theorems and semantics: A folk tale*, Inform. Process. Lett., 14 (1982), pp. 112–116.

[16] M. Lawford, W.M. Wonham, and J.S. Ostroff, *State-event observers for labeled transition systems*, in Proc. 33rd IEEE Conf. on Decision and Control, 1994, pp. 3642–3648.

[17] Y. Li and W.M. Wonham, *Controllability and observability in the state-feedback control of discrete-event systems*, in Proc. 27th IEEE Conf. on Decision and Control, 1988, pp. 203–208.

[18] Y. Li and W.M. Wonham, *Control of vector discrete-event systems* I. *The base model*, IEEE Trans. Automat. Control, 38 (1993), pp. 1214–1227.

[19] Y. Li and W.M. Wonham, *Control of vector discrete-event systems* II. *Controller synthesis*, IEEE Trans. Automat. Control, 39 (1994), pp. 512–531.

[20] F. Lin and W.M. Wonham, *On observability of discrete-event systems*, Inform. Sci., 44 (1988), pp. 173–198.

[21] F. Lin and W.M. Wonham, *Decentralized supervisory control of discrete-event systems*, Inform. Sci., 44 (1988), pp. 199–224.

[22] E.G. Manes and M.A. Arbib, *Algebraic Approaches to Program Semantics*, Springer-Verlag, New York, 1986.

[23] R. Paige and R.E. Tarjan, *Three partition refinement algorithms*, SIAM J. Comput., 16 (1987), pp. 973–989.

[24] P.J. Ramadge and W.M. Wonham, *Supervisory control of a class of discrete event processes*, SIAM J. Control Optim., 25 (1987), pp. 206–230.

[25] P.J. Ramadge and W.M. Wonham, *Modular feedback logic for discrete event systems*, SIAM J. Control Optim., 25 (1987), pp. 1202–1218.

[26] A. Tarski, *A lattice-theoretical fixpoint theorem and its applications*, Pacific J. Math., 5 (1955), pp. 285–309.

[27] W. Wechler, *Universal Algebra for Computer Scientists*, Springer-Verlag, Berlin, 1992.

[28] W.M. Wonham and P.J. Ramadge, *On the supremal controllable sublanguage of a given language*, SIAM J. Control Optim., 25 (1987), pp. 635–659.

[29] W.M. Wonham, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1985.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 3, pp. 710–730

SMOOTH NORMAL APPROXIMATIONS OF EPI-LIPSCHITZIAN SUBSETS OF $\mathbb{R}^n$*

BERNARD CORNET[†] AND MARC-OLIVIER CZARNECKI[‡]

Abstract. A sequence (M_k) of closed subsets of $\mathbb{R}^n$ converges normally to $M \subset \mathbb{R}^n$ if (sc) $M = \limsup M_k = \liminf M_k$ in the sense of Painlevé–Kuratowski and (nc) $\limsup \mathrm{G}(N_{M_k}) \subset \mathrm{G}(N_M)$, where $\mathrm{G}(N_M)$ (resp., $\mathrm{G}(N_{M_k})$) denotes the graph of N_M (resp., N_{M_k}), Clarke's normal cone to M (resp., M_k).

This paper studies the normal convergence of subsets of $\mathbb{R}^n$ and mainly shows two results. The first result states that every closed epi-Lipschitzian subset M of $\mathbb{R}^n$, with a compact boundary, can be approximated by a sequence of smooth sets (M_k), which converges normally to M and such that the sets M_k and M are lipeomorphic for every k (i.e., the homeomorphism between M and M_k and its inverse are both Lipschitzian). The second result shows that, if a sequence (M_k) of closed subsets of $\mathbb{R}^n$ converges normally to an epi-Lipschitzian set M, and if we additionally assume that the boundary of M_k remains in a fixed compact set, then, for k large enough, the sets M_k and M are lipeomorphic.

In Cornet and Czarnecki [*Cahier Eco-Maths 95-55,* 1995], direct applications of these results are given to the study (existence, stability, etc.) of the generalized equation $0 \in f(x^*) + N_M(x^*)$ when M is a compact epi-Lipschitzian subset of $\mathbb{R}^n$ and $f : M \to \mathbb{R}^n$ is a continuous map (or more generally a correspondence).

Key words. epi-Lipschitzian, normal convergence, smooth approximation, lipeomorphism, homeomorphism, Clarke's normal cone

AMS subject classifications. Primary, 49J45, 49J52; Secondary, 57R12

PII. S0363012997320122

1. Introduction.

A closed subset M of $\mathbb{R}^n$ is said to be epi-Lipschitzian if its Clarke's normal cone $N_M(x)$ is pointed (i.e., if $N_M(x) \cap -N_M(x) = \{0\}$) at every $x \in M$. This class of sets, introduced in optimization by Rockafellar [16], is of particular importance since it includes both (i) closed convex sets with a nonempty interior and (ii) sets defined by finite smooth inequality constraints satisfying a nondegeneracy assumption (independence of the binding constraints). Closed epi-Lipschitzian subsets M of $\mathbb{R}^n$ are equivalently defined as sets that can be locally written as the epigraph of a Lipschitzian function (see [16]).

A sequence (M_k) of closed subsets of $\mathbb{R}^n$ converges normally to $M \subset \mathbb{R}^n$ if (sc) $M = \limsup M_k = \liminf M_k$ in the sense of Painlevé–Kuratowski and (nc) $\limsup \mathrm{G}(N_{M_k}) \subset \mathrm{G}(N_M)$, where $\mathrm{G}(N_M) = \{(x,y) \in \mathbb{R}^n \times \mathbb{R}^n | x \in M, y \in N_M(x)\}$ (resp., $\mathrm{G}(N_{M_k})$) denotes the graph of N_M (resp., N_{M_k}), Clarke's normal cone to M (resp., M_k).

This paper studies the normal convergence of subsets of $\mathbb{R}^n$ and mainly shows two results. The first result (Theorem 2.1) states that every closed epi-Lipschitzian subset M of $\mathbb{R}^n$, with a compact boundary, can be approximated by a sequence of smooth sets (M_k), which converges normally to M and such that, for every k, the sets M_k and M are lipeomorphic (i.e., the homeomorphism between M and M_k and its

*Received by the editors April 18, 1997; accepted for publication (in revised form) May 1, 1998; published electronically February 19, 1999.

http://www.siam.org/journals/sicon/37-3/32012.html

[†] CERMSEM, Université de Paris 1 Panthéon-Sorbonne, 90 rue de Tolbiac, 75013 Paris, France (cornet@univ-paris1.fr).

[‡] CERMSEM, Université de Paris 1 and Laboratoire d'Analyse Convexe, Université de Montpellier 2, place Eugène Bataillon, 34095 Montpellier cedex 5, France (marco@math.univ-montp2.fr).

inverse are both Lipschitzian). Moreover, we prove that one can additionally assume that the approximating sequence (M_k) is internal (resp., external) in the following sense: $M_k \subset \mathrm{int} M_{k+1}$ (resp., $M_{k+1} \subset \mathrm{int} M_k$) for all $k \in \mathbb{N}$. This result extends previous ones in the literature (which do not consider the lipeomorphism properties); see Benoist [1] and (with a different formalism, without the geometrical concept of Clarke's cones) Nečas [15] (in Russian), Massari and Pepe [13], and Doktor [8].

In the above result, the lipeomorphism property is in fact a consequence of the normal convergence of the sequence (M_k). This is a consequence of our second result (Theorem 2.2), which states that if (M_k) is a sequence of closed subsets of $\mathbb{R}^n$ which converges normally to an epi-Lipschitzian set M, then the sets M_k and M are lipeomorphic for k large enough if we additionally assume that, for all k, $\mathrm{bd} M_k$ remains in some given compact set $K \subset \mathbb{R}^n$. In fact, we shall show (Theorem 2.3) that one can weaken assertion (nc) by only assuming that the convex hull of the set $\{p \in \mathbb{R}^n | (x,p) \in \limsup \mathrm{G}(N_{M_k})\}$ is pointed for every $x \in M$.

In [7], direct applications of these results are given to the study (existence, stability, etc.) of the generalized equation $0 \in f(x^*) + N_M(x^*)$ when M is a compact epi-Lipschitzian subset of $\mathbb{R}^n$ and $f : M \to \mathbb{R}^n$ is a continuous map (or more generally a correspondence).

The paper is organized as follows. The definitions and the main results are given in section 2. The proof of the approximation result (Theorem 2.1) is given in section 3, and the proof of the lipeomorphism result (Theorem 2.3) is given in section 4.

2. Definitions and statement of the results.

2.1. Preliminaries[1]**.** Let M be a closed subset of $\mathbb{R}^n$ and let $x \in M$. We define Clarke's normal cone to M at x, denoted $N_M(x)$, in two steps as follows. We first call perpendicular vector to M at x every vector in the set

$$\perp_M(x) = \{v \in \mathbb{R}^n | \exists \alpha > 0, B(x + \alpha v, \alpha \|v\|) \cap M = \emptyset\}.$$

Then Clarke's normal cone to M at x is the closure of the convex hull of the following limiting cone:

$$\hat{N}_M(x) = \{v \in \mathbb{R}^n | \exists (x_k)_{k \in \mathbb{N}} \subset M,\ \forall k \in \mathbb{N},\ \exists v_k \in \perp_M(x_k),\ (x_k) \to x, (v_k) \to v\}.$$

We now define Clarke's tangent cone to M at x, denoted $T_M(x)$, as the negative polar cone of $N_M(x)$, i.e.,

$$T_M(x) = \{u \in \mathbb{R}^n \mid \forall v \in N_M(x), (u|v) \leq 0\}.$$

We recall that a closed subset M of $\mathbb{R}^n$ is said to be epi-Lipschitzian if $N_M(x)$ is pointed (i.e., $N_M(x) \cap -N_M(x) = \{0\}$) for all $x \in M$. We say that M is C^k-smooth if it is a C^k (with $k \in \{1, \dots, \infty\}$) submanifold with a boundary of $\mathbb{R}^n$ of full dimension,

[1] We let $\mathbb{R}_+ = \{x \in \mathbb{R} | x \geq 0\}$. If $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ belong to $\mathbb{R}^n$, we denote the scalar product of $\mathbb{R}^n$ by $(x|y) = \sum_{i=1}^n x_i y_i$, the Euclidean norm by $\|x\| = \sqrt{(x|x)}$; we denote $B(x,r) = \{y \in \mathbb{R}^n | \ \|x - y\| < r\}$, $\overline{B}(x,r) = \{y \in \mathbb{R}^n | \ \|x - y\| \leq r\}$, $S(x,r) = \{y \in \mathbb{R}^n | \ \|x - y\| = r\}$, $\overline{B} = \overline{B}(0,1)$, and $S = S(0,1)$. If $X \subset \mathbb{R}^n$, $Y \subset \mathbb{R}^n$, and $x \in \mathbb{R}^n$, we let $d(x,X) = \inf_{y \in X} \|x - y\|$ (also denoted $d_X(x)$), and we denote by $X + Y = \{x + y | x \in X, y \in Y\}$ the sum of X and Y, $B(X,r) = X + B(0,r)$, $\mathrm{cl} X$ or $\overline{X}$ the closure of X, $\mathrm{int} X$ the interior of X, $\mathrm{bd} X$ the boundary of X, and $\mathrm{co} X$ the convex hull of X. If X and Y are two nonempty compact subsets of $\mathbb{R}^n$, $\delta(X,Y) = \max\{\sup_{x \in X} d(x,Y), \sup_{y \in Y} d(y,X)\}$ is the Hausdorff distance between X and Y. A correspondence Φ from $X \subset \mathbb{R}^n$ to $\mathbb{R}^m$ is a map from X to the set of all the subsets of $\mathbb{R}^m$ and the graph of Φ, denoted $\mathrm{G}(\Phi)$, is defined by $\mathrm{G}(\Phi) = \{(x,y) \in X \times \mathbb{R}^m | y \in \Phi(x)\}$.

i.e., if for all $\overline{x} \in M$, there is an open neighborhood U of $\overline{x}$ and a C^k function $f : U \to \mathbb{R}$ such that $M \cap U = \{x \in U | f(x) \leq 0\}$, and such that $\nabla f(x) \neq 0$ if $f(x) = 0$. M is said to be smooth if it is C^∞-smooth.

We recall the following definitions and properties associated with a sequence (M_k) of subsets of $\mathbb{R}^n$ (see, for example, Kuratowski [12]):

$$\begin{aligned} \liminf M_k &= \{x \in \mathbb{R}^n | \exists (x_k) \subset \mathbb{R}^n, x_k \to x, x_k \in M_k \text{ for all } k\}; \\ \limsup M_k &= \{x \in \mathbb{R}^n | \exists (x_k) \subset \mathbb{R}^n, \exists \varphi \in \mathcal{I}, x_k \to x, x_k \in M_{\varphi(k)} \text{ for all } k\},^2 \end{aligned}$$

where $\mathcal{I}$ is the set of all increasing maps $\varphi : \mathbb{N} \to \mathbb{N}$.

We recall that the inclusion $\liminf M_k \subset \limsup M_k$ always holds true; the sequence (M_k) is said to be set-convergent if $\liminf M_k = \limsup M_k$. We say that the sequence (M_k) is smooth if the set M_k is smooth for k large enough. We say that it is increasing (resp., decreasing) if $M_k \subset \text{int} M_{k+1}$ (resp., $M_{k+1} \subset \text{int} M_k$) for all $k \in \mathbb{N}$. If (M_k) is an increasing (resp., decreasing) sequence, then one notices that it set-converges to some $M \subset \mathbb{R}^n$ if and only if $M = \text{cl}(\cup_{k \in \mathbb{N}} M_k)$ (resp., $M = \cap_{k \in \mathbb{N}} M_k$). An increasing (resp., decreasing) converging sequence is also called an internal (resp., external) approximation of its set-limit.

2.2. Statement of the results. We give a stronger notion of set-convergence which involves both the set-convergence and the convergence of the graph of the normal cones in the following sense.

Definition 2.1. *We say that a sequence (M_k) of closed subsets of $\mathbb{R}^n$ is a normal approximation of a closed subset $M \subset \mathbb{R}^n$ (or converges normally to M) if the two following assertions hold:*

(sc) *(set convergence)* $M = \limsup M_k = \liminf M_k$;
(nc) *(normal convergence)* $\limsup \mathrm{G}(N_{M_k}) \subset \mathrm{G}(N_M)$.

Remark 2.1 (the convex case). Let (M_k) be a sequence of closed convex subsets of $\mathbb{R}^n$. Assume that (M_k) set-converges to some subset $M \subset \mathbb{R}^n$. Then one easily notices that the set M is convex and that (M_k) is a normal approximation of M.

The next theorem shows the existence of internal and external smooth normal approximations of a compact epi-Lipschitzian subset of $\mathbb{R}^n$, which satisfy additional properties also of interest for themselves (in fact we weaken the compactness assumption by assuming only that M is closed and that $\text{bd}M$ is compact). We recall that subsets M and N of $\mathbb{R}^n$ are lipeomorphic if there exists a map $\Phi : M \to N$ which is a lipeomorphism, i.e., is a Lipschitzian invertible map with a Lipschitzian inverse.

Theorem 2.1. *Let M be a closed epi-Lipschitzian subset of $\mathbb{R}^n$, such that* $\text{bd}M$ *is nonempty and compact. Then there exists a smooth internal normal approximation and a smooth external normal approximation of M which both additionally satisfy the following properties:*

(lip) *the sets M_k and M are lipeomorphic for all k;*
(lip^c) *the sets $\mathbb{R}^n \setminus \text{int} M_k$ and $\mathbb{R}^n \setminus \text{int} M$ are lipeomorphic for all k;*
(L) *there is $\ell > 0$ and a compact subset $K \subset \mathbb{R}^n$ such that, for all k,* $\text{bd}M_k \subset K$ *and*

$$\delta(\text{bd}M_k, \text{bd}M_{k+1}) \leq \ell \min\{\|x - y\| \, | x \in \text{bd}M_k, y \in \text{bd}M_{k+1}\}.$$

Theorem 2.1 is proved in section 3. A general discussion about assertion (L) is given in the next section.

[2] Equivalently, $\limsup M_k = \cap_{p \in \mathbb{N}} \text{cl}(\cup_{k \geq p} M_k)$. Note also that for every subsequence $(M_{\varphi(k)})$ of (M_k), one has that $\liminf M_k \subset \liminf M_{\varphi(k)} \subset \limsup M_{\varphi(k)} \subset \limsup M_k$.

At this stage, it is worth pointing out that in Theorem 2.1 the two lipeomorphism assertions (lip) and (lip$^{\mathrm{c}}$) are a consequence of the normal convergence of the sequence (M_k) as is shown in the next result.

THEOREM 2.2. *Let (M_k) be a sequence of closed subsets of $\mathbb{R}^n$ such that the boundaries* $\mathrm{bd}M_k$ *remain in a given compact subset $K \subset \mathbb{R}^n$. Assume that (M_k) set-converges to some subset $M \subset \mathbb{R}^n$ and that*

($*$) *(M_k) converges normally to M, and M is epi-Lipschitzian.*

Then, for k large enough,

(i) *the sets M, $\mathbb{R}^n \setminus \mathrm{int}M$, M_k and $\mathbb{R}^n \setminus \mathrm{int}M_k$ are epi-Lipschitzian;*

(ii) *M_k is lipeomorphic to M and $\mathbb{R}^n \setminus \mathrm{int}M_k$ is lipeomorphic to $\mathbb{R}^n \setminus \mathrm{int}M$.*

Theorem 2.2 is a consequence of the following result, which slightly weakens assertion ($*$) by noticing that the condition (nc) of normal convergence implies that

$$\{p \in \mathbb{R}^n | (x,p) \in \limsup \mathrm{G}(N_{M_k})\} = \limsup_{x' \to x, k \to \infty} N_{M_k}(x') \subset N_M(x).$$

THEOREM 2.3. *Theorem* 2.2 *remains true if one replaces assertion* ($*$) *with the following assertion:*

($**$) $\mathrm{co}\limsup_{x' \to x, k\to\infty} N_{M_k}(x')$ *is pointed for all $x \in M$.*

The proof of Theorem 2.3 is given in section 4.

Remark 2.2. Theorem 2.2 and Theorem 2.3 may not be true if $\mathrm{bd}M_k$ does not remain in a fixed compact set K. Consider $M = \mathbb{R} \times \mathbb{R}_+$ and the smooth internal normal approximation of M defined by $M_k = \mathbb{R} \times [1/k, \infty) \setminus B((k, 3/k), 1/k)$ for $k \geq 1$. Then, for every k, M_k is clearly not homeomorphic to M.

Remark 2.3. Theorem 2.2 may not be true if the set-limit M is not epi-Lipschitzian. Consider $M = \{0\}$ in $\mathbb{R}$ and the smooth external normal approximation of M defined by $M_k = [-1/k, 1/k]$. Then for every k, M_k is not homeomorphic to M. Similarly, Theorem 2.3 may not be true without the assumption that $\mathrm{co}\limsup_{x' \to x, k\to\infty} N_{M_k}(x')$ is pointed for all $x \in M$, even if one assume that M is epi-Lipschitzian. Consider $M = [1, 1]$ in $\mathbb{R}$ and the smooth internal normal approximation of M defined by $M_k = [-1, -1/k] \cup [1/k, 1]$, and notice that $\mathrm{co}\limsup_{x' \to 0, k\to\infty} N_{M_k}(x') = \mathbb{R}$.

Remark 2.4. Theorem 2.3 may not be true if we do not assume that $\limsup M_k = \liminf M_k$. Consider $M_{2k} = \overline{B}(0,1)$, $M_{2k+1} = \overline{B}(0,1) \setminus B(0, 1/2)$ (or $M_{2k+1} = \emptyset$), and notice that $\overline{B}(0,1) = \limsup M_k$.

2.3. General remarks about assertion (L).

Remark 2.5. There may exist a normal (external or internal) approximation satisfying all the conclusions of Theorem 2.1 except for assertion (L). Consider the subset $M = [0,1]$ of $\mathbb{R}$ and the sets $M_k = [-1/k, 1 + 1/2^k]$.

Remark 2.6. If $\mathrm{bd}M$ is not compact, there may not exist an internal (or an external) normal approximation of M which satisfies assertion (L). Consider the following closed epi-Lipschitzian subset of $\mathbb{R}$:

$$M = \left(\mathbb{R}_- \setminus \cup_{k=1}^{\infty}\left(-k - \frac{1}{k+1}, -k + \frac{1}{k+1}\right)\right) \cup \left(\cup_{k=1}^{\infty}\left[k - \frac{1}{k+1}, k + \frac{1}{k+1}\right]\right).$$

Then, if (M_k) is any smooth internal (or external) normal approximation of M, we let the reader check that it does not satisfy assertion (L).

Remark 2.7. Note that the inequality

$$\delta(X,Y) \geq \min\{\|x - y\| \, | x \in X, y \in Y\}$$

is always true if X and Y are two nonempty compact subsets of $\mathbb{R}^n$. Hence, one necessarily has $\ell \geq 1$ in assertion (L) of Theorem 2.1.

Remark 2.8. If we additionally assume that (M_k) is increasing or decreasing, then

$$\text{(L}'\text{)} \qquad \forall k,\ \delta(\text{bd}M_k, \text{bd}M) \leq \ell \min\{\|x-y\| \,|x \in \text{bd}M, y \in \text{bd}M_k\}.$$

Indeed, one just needs to notice that $\delta(\text{bd}M_k, \text{bd}M) \leq \sum_{i=k}^{\infty} \delta(\text{bd}M_i, \text{bd}M_{i+1})$ and that

$$\min\{\|x-y\| \,|x \in \text{bd}M, y \in \text{bd}M_k\} \geq \sum_{i=k}^{\infty} \min\{\|x-y\| \,|x \in \text{bd}M_i, y \in \text{bd}M_{i+1}\}.$$

Assertion (L$'$) is no longer true if we do not assume that the sequence (M_k) is increasing or decreasing. Consider the set $M = [0,1]$, the sets $M_{2k} = [-1/k, 1]$ and $M_{2k+1} = [0, 1 + 1/(k+1)]$ for all $k \geq 1$. Then (M_k) is a smooth approximation of M which satisfies assertion (L), but the above property (L$'$) is not true.

2.4. Other concepts of normal convergence.

2.4.1. Involving the subdifferential of the distance function. We first recall the definition of Clarke's subdifferential of a locally Lipschitzian function.[3] Let U be an open subset of $\mathbb{R}^n$ and consider $f : U \to \mathbb{R}$; if f is differentiable at $x \in U$, we denote $\nabla f(x)$ the gradient of f at x. If f is locally Lipschitzian, its subdifferential $\partial f(x)$ at $x \in U$ is defined by

$$\partial f(x) = \text{co}\{\lim_{k\to\infty} \nabla f(x_k) | x_k \to x,\ x_k \in \text{Dom}(\nabla f)\},$$

where $\text{Dom}(\nabla f)$ is the set on which f is differentiable. In the case of the distance function d_M to a closed set $M \subset \mathbb{R}^n$, one can be more precise. Indeed, from Clarke [4, Thm. 2.5.6],

$$\partial d_M(x) = \text{co}\Big((\hat{N}_M(x) \cap S) \cup \{0\}\Big), \tag{1}$$

where $\hat{N}_M(x)$ is the limiting normal cone defined previously and S is the unit sphere in $\mathbb{R}^n$.

It seems natural to compare the normal convergence with the following concept of ∂-convergence, in which one replaces the normal cone by the subdifferential of the distance function. More precisely, we say that a sequence (M_k) of closed subsets of $\mathbb{R}^n$ ∂-converges to a closed subset $M \subset \mathbb{R}^n$ if it satisfies assertion (sc) together with

(∂c) $\quad \limsup \text{G}(\partial d_{M_k}) \subset \text{G}(\partial d_M)$.

It is worth noticing that the ∂-convergence can be formulated only in terms of the distance function, by noticing that assertion (sc) can be equivalently reformulated as follows:

(sc$'$) $\quad \forall x \in \mathbb{R}^n$, $\lim_{k\to\infty} d_{M_k}(x) = d_M(x)$.

The link between normal convergence and ∂-convergence can be summarized as follows. It will appear that the concept of ∂-convergence is too strong (for a matter of normalization), even in the epi-Lipschitzian case. Indeed, in this paper we

[3] If $X \subset \mathbb{R}^n$, a map $f : X \to \mathbb{R}^m$ is locally Lipschitzian if, for all $x \in X$, there is a neighborhood U of x and a real number $K \geq 0$ such that $\|f(y) - f(z)\| \leq K\|y - z\|$ for all y and z in U.

show that every compact epi-Lipschitzian set can be approximated in the sense of normal convergence by a sequence of smooth sets. This result is no longer true with the ∂-convergence as shown below (Proposition 2.2) by taking $M = \mathbb{R}^2 \setminus \mathrm{int}\mathbb{R}^2_+$. Furthermore, in the epi-Lipschitzian case, the following proposition shows that the ∂-convergence implies the normal convergence, a result which is no longer true in general (see Remark 2.9).

Proposition 2.1. *Let (M_k) be a sequence of closed subsets of $\mathbb{R}^n$ which ∂-converges to some epi-Lipschitzian set $M \subset \mathbb{R}^n$. Then (M_k) converges normally to the set M.*

Proof of Proposition 2.1. Let $(x,p) \in \limsup \mathrm{G}(N_{M_k})$. Then there is a sequence (x_k) converging to x, a sequence (p_k) converging to p, and an increasing map $\varphi : \mathbb{N} \to \mathbb{N}$, such that $x_k \in M_{\varphi(k)}$ and $p_k \in N_{M_{\varphi(k)}}(x_k)$ for all k. Since $N_{M_{\varphi(k)}}(x_k) = \mathrm{cl}[\cup_{\lambda \geq 0} \lambda \partial d_{M_{\varphi(k)}}(x_k)]$ (see Clarke [4]) for all k, there is a sequence $(\lambda_r^k)_{r\in\mathbb{N}}$ in $\mathbb{R}_+$ and a sequence $(v_r^k)_{r\in\mathbb{N}}$ in $\partial d_{M_{\varphi(k)}}(x_k)$ such that $\lambda_r^k v_r^k$ converges to p_k when $r \to \infty$. Hence, without any loss of generality (using a diagonal argument), we may assume that $p = \lim_{k\to\infty} \lambda_k v_k$, with $\lambda_k \geq 0$ and $v_k \in \partial d_{M_{\varphi(k)}}(x_k) \subset \overline{B}(0,1)$, and that the bounded sequence (v_k) converges to some $v \in \mathbb{R}^n$. Then, from above and (1), for every integer k,

$$v_k \in \partial d_{M_{\varphi(k)}}(x_k) = \mathrm{co}\Big((\hat{N}_{M_{\varphi(k)}}(x_k) \cap S) \cup \{0\}\Big).$$

Hence, from Carathéodory's theorem, there are $n+1$ elements (v_k^i, λ_k^i) ($i \in \{1,\dots,n+1\}$) in $\mathbb{R}^n \times \mathbb{R}_+$ such that $v_k^i \in \hat{N}_{M_{\varphi(k)}}(x_k) \cap S \subset \partial d_{M_{\varphi(k)}}(x_k)$, $\sum_{i=1}^{n+1} \lambda_k^i = 1$, and $\mu_k \in [0,1]$ such that

$$v_k = \mu_k \sum_{i=1}^{n+1} \lambda_k^i v_k^i.$$

Again, without any loss of generality, we may assume that $(v_k^1,\dots,v_k^{n+1},(\lambda_k^1,\dots,\lambda_k^{n+1}),\mu_k)$ converges to some element $(v^1,\dots,v^{n+1},(\lambda^1,\dots,\lambda^{n+1}),\mu) \in S^{n+1} \times \Sigma \times [0,1]$, where Σ is the unit simplex of $\mathbb{R}^{n+1}$. From assertion (∂c), we get that $v \in \partial d_M(x)$ and that $v^i \in \partial d_M(x)$ for all $i \in \{1,\dots,n+1\}$. But for all i, from above $v^i \in S$ and $v^i \in \mathrm{co}\Big((\hat{N}_M(x) \cap S) \cup \{0\}\Big)$, noticing that

$$\mathrm{co}\Big((\hat{N}_M(x) \cap S) \cup \{0\}\Big) \cap S = \hat{N}_M(x) \cap S,$$

we deduce that $v^i \in \hat{N}_M(x) \cap S$. Then $w = \sum_{i=1}^{n+1} \lambda^i v^i \in \mathrm{co}(\hat{N}_M(x) \cap S)$, which does not contain 0 since M is epi-Lipschitzian; hence $w \neq 0$. Recalling that $p = \lim_{k\to\infty} \lambda_k v_k = \lim_{k\to\infty} \lambda_k \mu_k \sum_{i=1}^{n+1} \lambda_k^i v_k^i$, the sequence $(\lambda_k \mu_k)$ converges to $\rho = \|p\|/\|w\|$, and $p = \rho w$ with $w \in \mathrm{co}(\hat{N}_M(x) \cap S) \subset \partial d_M(x)$; hence $p \in N_M(x)$. This shows that (M_k) converges normally to M. □

Remark 2.9. Proposition 2.1 may no longer be true if M is not epi-Lipschitzian. Consider the set

$$M = \{(x,y) \in \mathbb{R}^2 | y \leq \sqrt{|x|}\}$$

and, for every integer $k \geq 1$, the set

$$M_k = \{(x,y) \in \mathbb{R}^2 | [y \geq 0 \text{ and } y \leq \sqrt{|x|} - 1/k] \text{ or } [y < 0 \text{ and } y \leq |kx| - 1/k]\}.$$

Then the sequence (M_k) ∂-converges to M (note that $\partial d_M(0) = [-1,1] \times \{0\} = \limsup_{x' \to 0, k \to \infty} \partial d_{M_k}(x')$). But $\limsup_{x' \to 0, k \to \infty} N_{M_k}(x') = \mathbb{R} \times \mathbb{R}_+$ and $N_M(0) = \mathbb{R} \times \{0\}$; hence assertion (nc) is not satisfied. □

The next proposition shows that the concept of ∂-convergence is too strong (for a matter of normalization).

Proposition 2.2. *The set* $\mathbb{R}^2 \setminus \mathrm{int}\mathbb{R}^2_+$, *i.e., the complementary of the interior of* $\mathbb{R}^2_+$ *in* $\mathbb{R}^2$, *cannot be approximated, in the sense of the* ∂*-convergence, by a sequence of smooth sets.*

Proof of Proposition 2.2. Assume that it is not true, and let (M_k) be a sequence of smooth subsets of $\mathbb{R}^2$ which ∂-converges to $M = \mathbb{R}^2 \setminus \mathrm{int}\mathbb{R}^2_+$. From Proposition 2.1, since M is clearly epi-Lipschitzian, (M_k) converges normally to M; this implies that $(\mathbb{R}^2 \setminus \mathrm{int} M_k)$ converges normally to $\mathbb{R}^2 \setminus \mathrm{int} M = \mathbb{R}^2_+$ (see Proposition 3.1) and we shall prove later (Lemma 4.1) that this implies

$$\perp_{\mathbb{R}^2_+}(0) \subset \limsup_{x \to 0, k \to \infty} N_{\mathbb{R}^2 \setminus \mathrm{int} M_k}(x).$$

Since $v = -(1/\sqrt{2}, 1/\sqrt{2}) \in \perp_{\mathbb{R}^2_+}(0) (= -\mathbb{R}^2_+)$, the above inclusion implies that there is a sequence (x_k) converging to 0, a sequence (v_k) converging to v, and an increasing map $\varphi : \mathbb{N} \to \mathbb{N}$, such that, for all k, $x_k \in \mathrm{bd}(\mathbb{R}^2 \setminus \mathrm{int} M_{\varphi(k)}) = \mathrm{bd} M_{\varphi(k)}$ and $v_k \in N_{\mathbb{R}^2 \setminus \mathrm{int} M_{\varphi(k)}}(x_k) = -N_{M_{\varphi(k)}}(x_k)$ (since the set $M_{\varphi(k)}$ is smooth). For k large enough, $v_k \neq 0$; since the set $M_{\varphi(k)}$ is smooth, $-v_k/\|v_k\|$ is the unique element in $N_{M_{\varphi(k)}}(x_k) \cap S$ and hence belongs to $\partial d_{M_{\varphi(k)}}(x_k)$. Then the sequence $(-v_k/\|v_k\|)$ converges to $-v$, which, from assertion (∂c), belongs to $\partial d_M(0)$. Hence,

$$-v = (1/\sqrt{2}, 1/\sqrt{2}) \in \partial d_M(0) = \mathrm{co}\{(1,0),(0,1),(0,0)\},$$

which is a contradiction. □

2.4.2. Involving the limiting normal cone. We now compare the normal convergence with the following concept of $\hat{N}$-convergence, in which one replaces Clarke's normal cone N_M with the limiting normal cone $\hat{N}_M$ (see Mordukhovich [14]) defined previously. More precisely, we say that a sequence (M_k) of closed subsets of $\mathbb{R}^n$ $\hat{N}$-converges to a closed subset $M \subset \mathbb{R}^n$ if it satisfies assertion (sc) together with

($\hat{\mathrm{n}}$c) $\quad \limsup \mathrm{G}(\hat{N}_{M_k}) \subset \mathrm{G}(\hat{N}_M)$.

The next proposition shows that the $\hat{N}$-convergence and the ∂-convergence are in fact equivalent.

Proposition 2.3. *Let* (M_k) *be a sequence of closed subsets of* $\mathbb{R}^n$ *and let* M *be a closed subset of* $\mathbb{R}^n$. *Then* (M_k) $\hat{N}$*-converges to* M *if and only if* (M_k) ∂*-converges to the set* M.

Proof of Proposition 2.3 ($\hat{N}$-convergence $\Rightarrow$ ∂-convergence). Let

$$(x, v) \in \limsup \mathrm{G}(\partial d_{M_k}).$$

Then there is a sequence (x_k) converging to x, a sequence (v_k) converging to v, and an increasing map $\varphi : \mathbb{N} \to \mathbb{N}$, such that $x_k \in M_{\varphi(k)}$ and $v_k \in \partial d_{M_{\varphi(k)}}(x_k)$ for all k. Since from (1)

$$\partial d_{M_{\varphi(k)}}(x_k) = \mathrm{co}\Big((\hat{N}_{M_{\varphi(k)}}(x_k) \cap S) \cup \{0\}\Big),$$

from Carathéodory's theorem, there are $n+1$ elements (v_k^i, λ_k^i) ($i \in \{1, \ldots, n+1\}$) in

$\mathbb{R}^n \times \mathbb{R}_+$ and an element $\mu_k \in [0,1]$ such that $v_k^i \in \hat{N}_{M_{\varphi(k)}}(x_k) \cap S$, $\sum_{i=1}^{n+1} \lambda_k^i = 1$, and

$$v_k = \mu_k \sum_{i=1}^{m+1} \lambda_k^i v_k^i.$$

Without any loss of generality, we may assume that $(v_k^1, \ldots, v_k^{n+1}, (\lambda_k^1, \ldots, \lambda_k^{n+1}), \mu_k)$ converges to some element $(v^1, \ldots, v^{n+1}, (\lambda^1, \ldots, \lambda^{n+1}), \mu) \in S^{n+1} \times \Sigma \times [0,1]$, where Σ is the unit simplex of $\mathbb{R}^{n+1}$. From assertion ($\hat{\text{n}}$c), we deduce that $v^i \in \hat{N}_M(x) \cap S$ for all $i \in \{1, \ldots, n+1\}$. Then $v \in \text{co}((\hat{N}_M(x) \cap S) \cup \{0\}) = \partial d_M(x)$ from (1). This shows that (M_k) ∂-converges to M.

Proof of Proposition 2.3 (∂-convergence $\Rightarrow$ $\hat{N}$-convergence). Let $(x,p) \in \limsup \mathrm{G}(\hat{N}_{M_k})$. Then there is a sequence (x_k) converging to x, a sequence (p_k) converging to p, and an increasing map $\varphi : \mathbb{N} \to \mathbb{N}$, such that $x_k \in M_{\varphi(k)}$ and $p_k \in \hat{N}_{M_{\varphi(k)}}(x_k)$ for all k. If $p = 0$, then clearly $p \in \hat{N}_M(x)$. Assume now that $p \neq 0$. Then for k large enough, $p_k \neq 0$, and $p_k/\|p_k\|$ converges to $p/\|p\|$. Since $p_k/\|p_k\| \in \hat{N}_{M_{\varphi(k)}}(x_k) \cap S \subset \partial d_{M_{\varphi(k)}}(x_k)$ (from (1)), and since (M_k) ∂-converges to M, we deduce that $p/\|p\| \in \partial d_M(x)$. Since

$$\partial d_M(x) \cap S = \text{co}\Big((\hat{N}_M(x) \cap S) \cup \{0\}\Big) \cap S = \hat{N}_M(x) \cap S,$$

we deduce that $p/\|p\| \in \hat{N}_M(x) \cap S$, hence that $p \in \hat{N}_M(x)$. This shows that the sequence (M_k) $\hat{N}$-converges to M. □

Remark 2.10. The proof of the equivalence between the ∂-convergence and the $\hat{N}$-convergence relies heavily on the following equality:

$$\partial d_M(x) = \text{co}\Big((\hat{N}_M(x) \cap S) \cup \{0\}\Big).$$

The difference between the ∂-convergence and the normal convergence might be explained by the fact that the inclusion

$$\partial d_M(x) \subset \text{co}\Big((N_M(x) \cap S) \cup \{0\}\Big)$$

may be strict, even in the epi-Lipschitzian case. Consider $M = \mathbb{R}^2 \setminus \text{int}\mathbb{R}^2_+$.

3. Proof of the approximation result. The first idea to prove Theorem 2.1 is to smooth the distance function d_M by using a classical convolution argument. Indeed, by doing so one directly gets the existence of a normal smooth approximation of M. However, the lipeomorphism properties are more difficult to obtain and our proof will consist of using a more refined argument of convolution (in fact, by using the representation theorem in [6]).

The proof of Theorem 2.1 has three steps. In the first step, we show that (M_k) is an internal (resp., external) smooth approximation of M which satisfies (lip), (lip$^{\text{c}}$), and (L) if and only if $(\mathbb{R}^n \setminus \text{int}M_k)$ is an external (resp., internal) smooth approximation of $\mathbb{R}^n \setminus \text{int}M$ which satisfies (lip), (lip$^{\text{c}}$), and (L). In view of this equivalence property, it is sufficient to only show in the following the existence of smooth internal approximations of epi-Lipschitzian sets. In the second step, we improve the representation theorem of Cornet and Czarnecki [6] when the epi-Lipschitzian set is additionally assumed to have a compact boundary. In the third step, the previous representation result allows us to get the approximating sequence. These three steps are proved successively in the following three sections.

3.1. Complementarity property of internal and external approximations.

PROPOSITION 3.1. *Let M and M_k $(k \in \mathbb{N})$ be closed epi-Lipschitzian subsets of $\mathbb{R}^n$. Then the two following assertions are equivalent:*

(i) (M_k) *is a (resp., internal, resp., external, resp., smooth, resp., satisfying* (lip), *resp.,* (lip^c), *resp.,* (L)) *normal approximation of M;*

(ii) $(\mathbb{R}^n \setminus \mathrm{int}M_k)$ *is a (resp., external, resp., internal, resp., smooth, resp., satisfying* (lip^c), *resp.,* (lip), *resp.,* (L)) *normal approximation of* $\mathbb{R}^n \setminus \mathrm{int}M$.

Remark 3.1. Proposition 3.1 is no longer true if we do not assume that M is epi-Lipschitzian. Consider $M = \{0\}$ in $\mathbb{R}$ and $M_k = [-1/k, 1/k]$ for $k \geq 1$. Then (M_k) is a smooth external normal approximation of M and $(\mathbb{R} \setminus \mathrm{int}M_k)$ is not a normal approximation of $\mathbb{R} \setminus \mathrm{int}M$. Consider also the set $M = \{(x,y) \in \mathbb{R}^2 | y \geq \sqrt{|x|}\}$ and the set $M_k = \{(x,y) \in \mathbb{R}^2 | [y \geq 0 \text{ and } y \geq \sqrt{|x|} - 1/k] \text{ or } [y < 0 \text{ and } y \geq (k^3/4)x^2 - (1/4k)]\}$ for $k \geq 1$. Then (M_k) is a smooth internal approximation of M but $(\mathbb{R}^n \setminus \mathrm{int}M_k)$ is not a normal approximation of $\mathbb{R}^n \setminus \mathrm{int}M$. Indeed, $N_{\mathbb{R}^n \setminus \mathrm{int}M}(0) = \mathbb{R} \times \{0\}$; hence the set $\{0\} \times \mathbb{R}_+$, which is contained in $\limsup G(N_{\mathbb{R}^n \setminus \mathrm{int}M_k})$, is not contained in $G(N_{\mathbb{R}^n \setminus \mathrm{int}M})$.

Before proving Proposition 3.1, we prove a claim.

CLAIM 3.1. *Let (M_k) be a sequence of closed subsets of $\mathbb{R}^n$ converging normally to a closed subset $M \subset \mathbb{R}^n$. Then*

$$\begin{array}{lll} (\mathrm{int}) & \mathrm{int}M & = \cup_{p\in\mathbb{N}} \mathrm{int}(\cap_{k\geq p} \mathrm{int}M_k); \\ (\mathrm{scc}) & \mathbb{R}^n \setminus \mathrm{int}M & = \limsup(\mathbb{R}^n \setminus \mathrm{int}M_k) = \liminf(\mathbb{R}^n \setminus \mathrm{int}M_k). \end{array}$$

Proof of Claim 3.1. We first prove assertion (int).[4] Since

$$\cup_{p\in\mathbb{N}} \cap_{k\geq p} M_k \subset \liminf M_k = M,$$

the inclusion $\cup_{p\in\mathbb{N}} \mathrm{int}(\cap_{k\geq p} \mathrm{int}M_k) \subset \mathrm{int}M$ is immediate. Let us consider

$$x \in \mathrm{int}M \setminus \cup_{p\in\mathbb{N}} \mathrm{int}(\cap_{k\geq p} \mathrm{int}M_k) = \mathrm{int}M \cap [\cap_{p\in\mathbb{N}} \mathrm{cl}(\mathbb{R}^n \setminus \cap_{k\geq p} \mathrm{int}M_k)].$$

Then there is a sequence (x_k) in $\mathbb{R}^n$ converging to x such that, for all $k \in \mathbb{N}$, $x_k \notin \cap_{l\geq k} \mathrm{int}M_l$. Without any loss of generality, we may assume that there is an increasing map $\varphi : \mathbb{N} \to \mathbb{N}$ such that, for all k, $x_k \notin \mathrm{int}M_{\varphi(k)}$. Since $x \in \mathrm{int}M \subset \liminf M_k$, there is a sequence (y_k) in $\mathbb{R}^n$ converging to x such that for all $k \in \mathbb{N}$, $y_k \in M_k$. Since $y_{\varphi(k)} \in M_{\varphi(k)}$, there is $z_k \in \mathrm{bd}M_{\varphi(k)} \cap [x_k, y_{\varphi(k)}]$ and there is $v_k \in N_{M_{\varphi(k)}}(z_k) \cap S$ (see Clarke [4]). The sequence (z_k) converges to x, and we may assume without any loss of generality that v_k converges to some $v \in S$. Hence $(x, v) \in \limsup G(N_{M_k}) \subset G(N_M)$, which implies that $v \in N_M(x)$. The fact that $N_M(x) \neq \{0\}$ contradicts that $x \in \mathrm{int}M$.

Let us now prove assertion (scc). Since $M = \limsup M_k = \cap_{p\in\mathbb{N}} \mathrm{cl}(\cup_{k\geq p} M_k)$, we get that

$$\mathbb{R}^n \setminus M = \cup_{p\in\mathbb{N}} \mathrm{int}(\cap_{k\geq p} \mathbb{R}^n \setminus M_k) \subset \cup_{p\in\mathbb{N}} \cap_{k\geq p} \mathbb{R}^n \setminus \mathrm{int}M_k \subset \liminf(\mathbb{R}^n \setminus \mathrm{int}M_k),$$

hence that $\mathbb{R}^n \setminus \mathrm{int}M \subset \liminf(\mathbb{R}^n \setminus \mathrm{int}M_k) \subset \limsup(\mathbb{R}^n \setminus \mathrm{int}M_k)$. From assertion (int), we get that $\mathbb{R}^n \setminus \mathrm{int}M = \cap_{p\in\mathbb{N}} \mathrm{cl}(\cup_{k\geq p} \mathbb{R}^n \setminus \mathrm{int}M_k) = \limsup(\mathbb{R}^n \setminus \mathrm{int}M_k)$. □

Proof of Proposition 3.1. Note that, without any loss of generality, we only need to prove the implication [(i) ⇒ (ii)]. The implication [(ii) ⇒ (i)] can then be deduced

[4] In fact, we prove it in a more general setting later in this paper with a longer proof (Lemma 4.3).

from [(i) ⇒ (ii)], applying the result to the set $\mathbb{R}^n \setminus \operatorname{int} M$, since $\mathbb{R}^n \setminus \operatorname{int} M$ is epi-Lipschitzian, since $\mathbb{R}^n \setminus \operatorname{int}(\mathbb{R}^n \setminus \operatorname{int} M) = \overline{\operatorname{int} M} = M$[5] and $\mathbb{R}^n \setminus \operatorname{int}(\mathbb{R}^n \setminus \operatorname{int} M_k) = \overline{\operatorname{int} M_k} = M_k$ for all k.

In view of Claim 3.1, the sequence $(\mathbb{R}^n \setminus \operatorname{int} M_k)$ and the set $\mathbb{R}^n \setminus \operatorname{int} M$ clearly satisfy assertion (sc) of Definition 2.1. Let us prove that they satisfy assertion (nc). Let $(x,p) \in \limsup G(N_{\mathbb{R}^n \setminus \operatorname{int} M_k})$. Without any loss of generality, we may assume that $p \neq 0$, hence that $x \in \limsup \operatorname{bd}(\mathbb{R}^n \setminus \operatorname{int} M_k) = \limsup \operatorname{bd} M_k$, since $\operatorname{bd}(\mathbb{R}^n \setminus \operatorname{int} M_k) = (\mathbb{R}^n \setminus \operatorname{int} M_k) \setminus \operatorname{int}(\mathbb{R}^n \setminus \operatorname{int} M_k) = \overline{\operatorname{int} M_k} \setminus \operatorname{int} M_k = \operatorname{bd} M_k$ for all k (since M_k is epi-Lipschitzian). This implies that $x \in M \cap (\mathbb{R}^n \setminus \operatorname{int} M) = \operatorname{bd} M = \operatorname{bd}(\mathbb{R}^n \setminus \operatorname{int} M)$. Then, since M_k is epi-Lipschitzian, $N_{\mathbb{R}^n \setminus \operatorname{int} M_k}(x') = -N_{M_k}(x')$ for all $x' \in \operatorname{bd} M_k$, hence $(x,-p) \in \limsup G(N_{M_k})$, hence $(x,-p) \in G(N_M)$, which implies that $(x,p) \in G(N_{\mathbb{R}^n \setminus \operatorname{int} M})$, since $N_{\mathbb{R}^n \setminus \operatorname{int} M}(x) = -N_M(x)$. We proved that $(\mathbb{R}^n \setminus \operatorname{int} M_k)$ is an approximation of $\mathbb{R}^n \setminus \operatorname{int} M$. If (M_k) is a smooth (resp., internal, resp., external, resp., satisfying (lip), resp., (lip^c), resp., (L)) approximation of M, then $(\mathbb{R}^n \setminus \operatorname{int} M_k)$ is clearly a smooth (resp., external, resp., internal, resp., satisfying (lip), resp., (lip^c), resp., (L)) approximation of $\mathbb{R}^n \setminus \operatorname{int} M$. □

3.2. A representation theorem.

We first state a representation theorem of M when $\operatorname{bd} M$ is compact.

THEOREM 3.1. *Let M be a closed epi-Lipschitzian subset of $\mathbb{R}^n$ with compact boundary $\operatorname{bd} M$. Then there is a function $f_M : \mathbb{R}^n \to \mathbb{R}$ which is a quasi-smooth inequality representation of M in the following sense:*

(i) *f_M is locally Lipschitzian on $\mathbb{R}^n$ and C^∞ on $\mathbb{R}^n \setminus \operatorname{bd} M$;*

(ii) *$M = \{x \in \mathbb{R}^n | f_M(x) \le 0\}$;*

(iii) *$\operatorname{bd} M = \{x \in \mathbb{R}^n | f_M(x) = 0\}$;*[6]

(iv) *$0 \notin \partial f_M(x)$ if $f_M(x) = 0$;*

(v) *$N_M(x) \cap -N_{\mathbb{R}^n \setminus \operatorname{int} M}(x) = \cup_{\lambda \ge 0} \lambda \partial f_M(x)$ for all $x \in \operatorname{bd} M$.*

Furthermore, one can assume that for some $\varepsilon > 0$:

(vi) *$f_M^{-1}([-\varepsilon,\varepsilon])$ is compact;*

(vii) *$\forall x \in f_M^{-1}([-\varepsilon,\varepsilon])$, $\operatorname{co}\partial f_M(\overline{B}(x,\varepsilon)) \cap \overline{B}(0,\varepsilon) = \emptyset$.*

Proof of Theorem 3.1. The existence of a function f satisfying assertions (i)–(v) is exactly Theorem 2.1 of Cornet and Czarnecki [6] (in which the closed set M is not assumed to have a compact boundary).

Let f be a quasi-smooth representation of M on $\mathbb{R}^n$ (i.e., satisfying assertions (i)–(v)). Let $\alpha : \mathbb{R}^n \to [0,1]$ be a C^∞ function such that $\alpha(x) = 0$ if $x \in \overline{B}(\operatorname{bd} M, 1/2)$ and $\alpha(x) = 1$ if $x \notin B(\operatorname{bd} M, 1)$. We define the function $f_M : \mathbb{R}^n \to \mathbb{R}$ for all $x \in \mathbb{R}^n$ by

$$\begin{array}{lll} f_M(x) = (1-\alpha(x))f(x) + \alpha(x)\operatorname{sgn} f(x) & \text{if} & f(x) \neq 0, \\ f_M(x) = 0 & \text{if} & f(x) = 0, \end{array}$$

denoting $\operatorname{sgn} t = t/|t|$ if $t \in \mathbb{R} \setminus \{0\}$.

Proof of (i)–(iii). The function f_M clearly satisfies assertions (i), (ii), and (iii) of Theorem 3.1.

Proof of (iv) *and* (v). Let $x \in \operatorname{bd} M$. Since $\alpha = 0$ on a neighborhood of x, one gets that $\partial f_M(x) = \partial f(x)$; hence assertions (iv) and (v) of Theorem 3.1 are satisfied.

[5] This is a classical result on epi-Lipschitzian sets (see, for example, Cornet and Czarnecki [6]).

[6] This assertion is a consequence of assertions (ii) and (iv).

Proof of (vi). Since $f_M^{-1}((-1,1)) \subset B(\mathrm{bd}M, 1)$, then $\mathrm{cl}(f_M^{-1}((-1,1)))$ is compact.

Proof of (vii). It is a consequence of the following lemma (taking $m = n$, $K = \mathrm{bd}M = f_M^{-1}(\{0\})$, and $\Phi = \partial f_M$) and of the fact that $f_M^{-1}([-\varepsilon,\varepsilon]) \subset B(\mathrm{bd}M, r)$ for some $\varepsilon \in (0,1]$ (since $B(\mathrm{bd}M,r)$ is an open set containing the intersection of compact sets $\cap_{\varepsilon\in(0,1]} f_M^{-1}([-\varepsilon,\varepsilon])$).

LEMMA 3.1. *Let* K *be a compact subset of* $\mathbb{R}^n$ *and let* Φ *be a u.s.c. correspondence from* $\mathbb{R}^n$ *to* $\mathbb{R}^m$*, with nonempty compact convex values, such that* $0 \notin \Phi(x)$ *for every* $x \in K$*. Then there exists* $r > 0$ *such that*

$$\mathrm{co}\Phi(\overline{B}(x,r)) \cap \overline{B}(0,r) = \emptyset, \textit{ for all } x \in B(K,r).$$

Proof of Lemma 3.1 *(by contraposition).* Suppose that there exists a sequence (x^k) in $\mathbb{R}^n$ such that, for all k, $x^k \in B(K,1/k)$ and

$$\mathrm{co}\Phi(\overline{B}(x^k,1/k)) \cap \overline{B}(0,1/k) = \emptyset.$$

From Carathéodory's theorem, there exist $n+1$ elements $(x_i^k, y_i^k, \lambda_i^k)$ $(i = 1,\ldots,n+1)$ in $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_+$ such that $x_i^k \in \overline{B}(x^k,1/k)$, $y_i^k \in \Phi(x_i^k)$, $\sum_{i=1}^{m+1} \lambda_i^k = 1$, and

$$\left\| \sum_{i=1}^{m+1} \lambda_i^k y_i^k \right\| \leq 1/k.$$

Without loss of generality, we assume that the sequence $(x^k, \lambda_1^k, \ldots, \lambda_{m+1}^k, y_1^k, \ldots, y_{m+1}^k)$ converges to some element $(x^*, \lambda_1^*, \ldots, \lambda_{m+1}^{k*}, y_1^*, \ldots, y_{m+1}^*) \in K \times \Sigma \times \mathbb{R}^{m(m+1)}$, since the sequence belongs to the compact set $\overline{B}(K,1) \times \Sigma \times \Phi(\overline{B}(K,1))^{m+1}$, where Σ is the unit simplex of $\mathbb{R}^{m+1}$ and the set $\Phi(\overline{B}(K,1))$ is clearly bounded (since $\Phi(\overline{B}(K,1))$ is the image of the compact set $\overline{B}(K,1)$ by the u.s.c. correspondence Φ). However, for all $i \in \{1,\ldots,m+1\}$, the sequence (x_i^k) also converges to x^* (since from above $\|x_i^k - x^k\| \leq 1/k$).

Taking the limit when $k \to \infty$, we get $0 = \sum_{i=1}^{m+1} \lambda_i^* y_i^*$ and $y_i^* \in \Phi(x^*)$ for all $i \in \{1,\ldots,m+1\}$, since the correspondence Φ is u.s.c. Consequently, $0 \in \Phi(x^*)$ since $\Phi(x^*)$ is convex. Since $x^* \in K$, this contradicts the assumption $0 \notin \Phi(x^*)$. □

Remark 3.2. When $\mathrm{bd}M$ is not compact, assertion (vi) is clearly false for any quasi-smooth inequality representation of M. The following example shows that assertion (vii) may not be true if $\mathrm{bd}M$ is not assumed to be compact. Consider the set $M = \{(x,y) \in \mathbb{R}^2 | (x \leq 0) \text{ or } (x > 0 \text{ and } y \in [-1/x, 1/x])\}$.

3.3. Proof of Theorem 2.1. In view of Proposition 3.1, we only need to show the existence of smooth internal approximations of epi-Lipschitzian sets. Let f_M be a quasi-smooth representation of M satisfying the conclusions of Theorem 3.1 for some $\varepsilon > 0$; we let

$$M_k = \{x \in \mathbb{R}^n | f_M(x) \leq -\varepsilon/k\}$$

for every integer $k \geq 1$. We shall prove that (M_k) is a smooth internal approximation of M satisfying the conclusions of Theorem 2.1.

It is clearly an increasing sequence, and the set-convergence assertion (sc) is immediate. The sets M_k are clearly smooth, since f_M is C^∞ on $\{x \in \mathbb{R}^n | f_M(x) \neq 0\}$ and since $\nabla f_M(x) \neq 0$ when $x \in \mathrm{bd}M_k$ (in that case, $f_M(x) = -\varepsilon/k$, $f_M(x) \in [-\varepsilon,\varepsilon]$ and $0 \notin \partial f_M(x) = \{\nabla f_M(x)\}$ from Theorem 3.1, assertion (vii)).

Proof of the normal convergence assertion (nc). Let $(x, v) \in \limsup G(N_{M_k})$. Then there exist two sequences (x_k) and (v_k) in $\mathbb{R}^n$ and an increasing function $\varphi : \mathbb{N} \to \mathbb{N}$, such that (x_k) converges to x, (v_k) converges to v, and, for all k, $x_k \in M_{\varphi(k)}$ and $v_k \in N_{M_{\varphi(k)}}(x_k)$. Then, for all k, there is $\lambda_k \geq 0$ such that $v_k = \lambda_k \nabla f_M(x_k)$. Since f_M is Lipschitzian on a neighborhood of x, the sequence $(\nabla f_M(x_k))$ is bounded. Without any loss of generality, we may assume that it converges to some $u \in \mathbb{R}^n$. Since the correspondence ∂f_M is u.s.c., $u \in \partial f_M(x)$, hence $u \neq 0$ since $0 \notin \partial f_M(x)$. This implies that the sequence (λ_k) converges to $\lambda = \|v\|/\|u\|$ and $v = \lambda u$ with $\lambda \geq 0$ and $u \in \partial f_M(x)$. Hence $v \in N_M(x)$.

Proof of the lipeomorphism assertions (lip) *and* (lip$^{\rm c}$). Let us now prove that the sets M and M_k are lipeomorphic for all k. In view of Bonnisseau–Cornet [2, Theorem 2.5],[7] assertion (lip) is a clear consequence of the following facts:

$$\begin{aligned}
&M_k = \{x \in \mathbb{R}^n | f_M(x) \leq -\varepsilon/k\} \text{ for every } k,\\
&M = \{x \in \mathbb{R}^n | f_M(x) \leq 0\},\\
&f_M^{-1}([-\varepsilon/k, 0]) \text{ is compact,}\\
&0 \notin \partial f_M(x) \text{ when } f_M(x) \in [-\varepsilon/k, 0].
\end{aligned}$$

The proof that the sets $\mathbb{R}^n \setminus \mathrm{int} M_k$ and $\mathbb{R}^n \setminus \mathrm{int} M$ are lipeomorphic is a consequence of the same result, since $\mathbb{R}^n \setminus \mathrm{int} M_k = \{x \in \mathbb{R}^n | f_M(x) \geq -\varepsilon/k\}$ and $\mathbb{R}^n \setminus \mathrm{int} M = \{x \in \mathbb{R}^n | f_M(x) \leq 0\}$.

Proof of (L). In view of assertion (sc) of Definition 2.1, since the sequence (M_k) is increasing, and since $\mathrm{bd} M$ is compact, there clearly is an integer k_0 such that $\mathrm{bd} M_k \subset B(\mathrm{bd} M, \varepsilon)$ for all $k \geq k_0$. Let us consider $k \geq k_0$, and let L be the Lipschitz constant of f_M on $B(\mathrm{bd} M, \varepsilon)$. We first prove that

$$\varepsilon[(1/k) - 1/(k+1)] \leq L \min\{\|x - y\|, x \in \mathrm{bd} M_k, y \in \mathrm{bd} M_{k+1}\}.$$

Indeed, if $x \in \mathrm{bd} M_k$ and $y \in \mathrm{bd} M_{k+1}$, then $f_M(x) = -\varepsilon/k$ and $f_M(y) = -\varepsilon/(k+1)$. Then $\varepsilon[(1/k) - 1/(k+1)] = |f_M(x) - f_M(y)| \leq L\|x - y\|$. We now end the proof of Theorem 2.1 by proving that

$$\delta(\mathrm{bd} M_k, \mathrm{bd} M_{k+1}) \leq [(1/k) - 1/(k+1)].$$

We first consider $x \in \mathrm{bd} M_k$. Since f_M satisfies Theorem 3.1, assertion (vii), we can separate the compact convex sets $\mathrm{co}\partial f_M(\overline{B}(x, \varepsilon))$ and $\overline{B}(0, \varepsilon)$. Hence there is $p \in S(0, 1)$ such that $(p|y) > \varepsilon$ for all $y \in \mathrm{co}\partial f_M(\overline{B}(x, \varepsilon))$. Then, considering the map $t \mapsto f_M(x + tp)$, from the mean-value theorem (see Clarke [4]) one easily proves that $f_M(x + tp) - f_M(x) \geq \varepsilon t$ for all $t \in [0, \varepsilon]$. Then $f_M(x + \varepsilon p) \geq -\varepsilon/k + \varepsilon^2 > 0$ (if k is large enough), hence there is $t \in [0, \varepsilon]$ such that $f_M(x + tp) = -\varepsilon/(k + 1)$. Hence $\varepsilon t \leq f_M(x + tp) - f_M(x) = \varepsilon[(1/k) - 1/(k + 1)]$. This implies that $d(x, \mathrm{bd} M_{k+1}) \leq \|tp\| \leq [(1/k) - 1/(k + 1)]$. One proves that $d(x, \mathrm{bd} M_k) \leq [(1/k) - 1/(k + 1)]$ for all $x \in \mathrm{bd} M_{k+1}$ in the same way. Hence $\delta(\mathrm{bd} M_k, \mathrm{bd} M_{k+1}) \leq [(1/k) - 1/(k + 1)]$. □

Remark 3.3. From the above proof, we note that assertion (L) is satisfied by the sequence $M_k = \{x \in \mathbb{R}^n | f_M(x) \leq \varepsilon_k\}$, where $f_M : \mathbb{R}^n \to \mathbb{R}$ is a quasi-smooth representation of M and (ε_k) is a strictly decreasing sequence of positive real numbers converging to zero. However, there may exist normal converging sequences which cannot be represented as above. Consider the example in Remark 2.5.

[7] Bonnisseau and Cornet [2] prove a homeomorphism result, but only a slight change in their proof gives a lipeomorphism result.

4. Proof of the lipeomorphism result. In view of Proposition 3.1, the proof of Theorem 2.3 has three steps. We first show that the sets M and M_k are epi-Lipschitzian for k large enough. In the second step, we show the existence of a Lipschitzian transverse field between M and M_k, in a sense that will be explained later. In the third step of the proof we show that, if there is a such transverse field between two sets M and N, then M and N are (epi-Lipschitzian and) lipeomorphic. These three steps are proved successively in the following three sections.

4.1. The sets M and M_k are epi-Lipschitzian.

PROPOSITION 4.1. *Let (M_k) be a sequence of closed subsets of $\mathbb{R}^n$ such that $\mathrm{bd} M_k \subset K$ for all k, for some fixed compact subset $K \subset \mathbb{R}^n$, and such that*

$(**)$ $\quad \mathrm{co}\limsup_{x'\to x,k\to\infty} N_{M_k}(x')$ *is pointed for all $x \in \limsup M_k$.*

Then the set $M = \limsup M_k$ and the set M_k, for k large enough, are epi-Lipschitzian.

Proof of Proposition 4.1. We first prove that M_k is epi-Lipschitzian for k large enough. Assume that it is not true. Then there is a sequence (x_k) in $\mathbb{R}^n$, a sequence (p_k) in S, and an increasing map $\varphi : \mathbb{N} \to \mathbb{N}$, such that, for all $k \in \mathbb{N}$, $x_k \in M_{\varphi(k)}$, $p_k \in N_{M_{\varphi(k)}}(x_k)$, and $-p_k \in N_{M_{\varphi(k)}}(x_k)$. Then $x_k \in \mathrm{bd} M_{\varphi(k)} \subset K$; hence we may assume without any loss of generality that the sequence x_k converges to some $x \in K$ and that the sequence p_k converges to some $p \in S$. Then $x \in M$ (since $M = \limsup M_k$) and p and $-p$ belong to $\limsup_{x'\to x,k\to\infty} N_{M_k}(x')$, which contradicts the fact that $\mathrm{co}\limsup_{x'\to x,k\to\infty} N_{M_k}(x')$ is pointed for every $x \in M$.

We now prove that M is epi-Lipschitzian. Let $x \in M$. Then, since the set $\mathrm{co}\limsup_{x'\to x,k\to\infty} N_{M_k}(x')$ is pointed, it is sufficient to prove that $N_M(x) \subset \mathrm{co}\limsup_{x'\to x,k\to\infty} N_{M_k}(x')$ for all $x \in M$, which we do in the next lemma, which generalizes Lemma 6.2 from Benoist [1] (see also Kruger and Mordukhovich [11, Theorem P.3] and Ioffe [10, Theorem 3]).

LEMMA 4.1. *Let (M_k) be a sequence of closed subsets of $\mathbb{R}^n$, and let $M = \limsup M_k$. Then, for all $x \in M$*

(i) $\quad \perp_M(x) \subset \limsup_{x'\to x,k\to\infty} N_{M_k}(x')$;[8]

(ii) $\quad N_M(x) \subset \mathrm{cl}\ (\mathrm{co}\limsup_{x'\to x,k\to\infty} N_{M_k}(x'))$;

(iii) $\quad$ *if we additionally assume that the set $\mathrm{co}\limsup_{x'\to x,k\to\infty} N_{M_k}(x')$ is pointed, then we can suppress* cl *in the above assertion, i.e., formally*

$$N_M(x) \subset \mathrm{co} \limsup_{x'\to x,k\to\infty} N_{M_k}(x').$$

Proof of Lemma 4.1. Proof of (i). Let $\overline{x} \in M$ and $\overline{p} \in \perp_M(\overline{x}) \setminus \{0\}$. One easily notices from the definition of $\perp_M(\overline{x})$ that, for $\mu > 0$ small enough,

$$\overline{B}(\overline{x} + \mu\overline{p}, \mu\|\overline{p}\|) \cap M = \{\overline{x}\}. \tag{2}$$

We define the function $\psi : \mathbb{R}^n \to \mathbb{R}$ by $\psi(x) = (1/2)\|x - (\overline{x} + \mu\overline{p})\|^2$. Then, for all integer k there is a solution $x_k \in \mathbb{R}^n$ of the following minimization problem:

$$(P_k)\left\{\begin{array}{ll} \text{minimize} & \psi(x), \\ x \in M_k. & \end{array}\right.$$

Then, from (2), $\overline{x} + \mu\overline{p} \notin M$. Hence, since $M = \limsup M_k$, $\overline{x} + \mu\overline{p} \notin M_k$ for k

[8] One can easily replace assertion (i) with $\perp_M(x) \subset \limsup_{x'\to x,k\to\infty} \perp_{M_k}(x')$ or, equivalently, $G(\perp_M) \subset \limsup G(\perp_{M_k})$. However, without the convexification of the right-hand side, assertion (iii) does not hold in general.

large enough; hence $x_k \in \mathrm{bd}M_k$. Then x_k satisfies the following first-order necessary condition associated with (P_k) (see Clarke [4]):

$$-x_k + \overline{x} + \mu\overline{p} = -\nabla\psi(x_k) \in N_{M_k}(x_k). \tag{3}$$

Let us show that the sequence (x_k) admits a bounded subsequence. Since $\overline{x} \in M = \limsup M_k$, there is a sequence $(\overline{x}_k)$ converging to $\overline{x}$ and an increasing map $\varphi : \mathbb{N} \to \mathbb{N}$, such that $\overline{x}_k \in M_{\varphi(k)}$ for all k. Then, for every k, $\overline{x}_k$ satisfies the constraint of $(P_{\varphi(k)})$; hence

$$\psi(x_{\varphi(k)}) = (1/2)\|x_{\varphi(k)} - \overline{x} - \mu\overline{p}\|^2 \leq \psi(\overline{x}_k) = (1/2)\|\overline{x}_k - \overline{x} - \mu\overline{p}\|^2. \tag{4}$$

Because the sequence $(\overline{x}_k)$ is convergent, hence bounded, this implies that $(x_{\varphi(k)})$ is bounded. Without any loss of generality, we may assume that the sequence $(x_{\varphi(k)})$ converges to some $x \in \mathbb{R}^n$. Since $x_{\varphi(k)} \in M_{\varphi(k)}$ for all $k \in \mathbb{N}$, we get that $x \in M$. Since the sequence $(\overline{x}_k)$ converges to $\overline{x}$, $\psi(\overline{x}_k)$ converges to $(1/2)\mu^2\|\overline{p}\|^2$, hence (4) implies that $\psi(x) \leq (1/2)\mu^2\|\overline{p}\|^2$ and hence that $x \in \overline{B}(\overline{x} + \mu\overline{p}, \mu\|\overline{p}\|)$. In view of (2), since additionally $x \in M$, we get that $x = \overline{x}$. Letting $p_{\varphi(k)} = (1/\mu)(-x_{\varphi(k)} + \overline{x} + \mu\overline{p})$, we proved that $\overline{p} = \lim_{k\to\infty} p_{\varphi(k)}$ with $p_{\varphi(k)} \in N_{M_{\varphi(k)}}(x_{\varphi(k)})$.

Proof of (ii) *and* (iii). Since $A = \limsup_{x'\to\overline{x},k\to\infty} N_{M_k}(x')$ is a cone, since the correspondence $x \mapsto \limsup_{x'\to x,k\to\infty} N_{M_k}(x')$ is closed at $\overline{x}$, we get that $N_M(\overline{x}) \subset \mathrm{cl}(\mathrm{co}A)$, which proves (ii). If we additionally assume that $\mathrm{co}A$ is pointed, then $\mathrm{cl}(\mathrm{co}A) = \mathrm{co}A$, since A is closed (recalling that $\mathrm{co}A$ is closed, when A is a closed cone such that $\mathrm{co}A$ is pointed). □

4.2. A transverse field between M and M_k.

PROPOSITION 4.2. *Let M and (M_k) satisfy the hypothesis of Theorem* 2.3*. Then, for k large enough, there exists a Lipschitzian transverse field between the two sets M and $N = M_k$ in the following sense:*

$$(\mathrm{T}) \left\{ \begin{array}{ll} \multicolumn{2}{l}{\textit{There is bounded Lipschitzian map } F : \mathbb{R}^n \to \mathbb{R}^n\textit{, such that}} \\ \text{(a)} & M \setminus \mathrm{int}N \textit{ and } N \setminus \mathrm{int}M \textit{ are compact;} \\ \text{(b)} & \forall x \in \mathrm{bd}M \ (\textit{resp.}, \mathrm{bd}N),\ F(x) \in \mathrm{int}T_M(x),\ (\textit{resp.}, F(x) \in \mathrm{int}T_N(x)); \\ \text{(c)} & \textit{for all } x \textit{ in an open neighborhood } U \textit{ of } M \setminus \mathrm{int}N \cup N \setminus \mathrm{int}M, \\ & \exists (t,t') \in \mathbb{R}^2, \varphi(t,x) \in \mathrm{bd}M, \varphi(t',x) \in \mathrm{bd}N.^{9} \end{array} \right.$$

Remark 4.1. From [5], the assertion $\forall x \in \mathrm{bd}M$, $F(x) \in \mathrm{int}T_M(x)$ is equivalent to the weaker one, $\forall x \in \mathrm{bd}M$, $F(x) \in \mathrm{int}\mathcal{T}_M(x)$, where $\mathcal{T}_M(x)$ is Bouligand's tangent cone to M at x.

Note that from Proposition 4.1 the sets M and M_k are epi-Lipschitzian for k large enough. Before proving Proposition 4.2, we prove some preliminary lemmas.

4.2.1. Preliminary lemmas.

The first lemma gives an extended Gauss correspondence, an essential step to the construction of the transverse field.

LEMMA 4.2. *There is $r > 0$ and a correspondence G from $\mathbb{R}^n$ to $\mathbb{R}^n$ which is u.s.c., with nonempty compact convex values and such that*

(i) *$G_M(x) = N_M(x) \cap S \subset G(x)$ for all $x \in \mathrm{bd}M$;*

(ii) *$G_{M_k}(x) = N_{M_k}(x) \cap S \subset G(x)$ for all $x \in \mathrm{bd}M_k$, for k large enough;*

[9] Where φ is the flow of the following differential equation:

$$(E)\ \dot{x}(t) = F(x(t)),\ x(0) = x,$$

i.e., $\varphi : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ is the unique C^1 map such that $t \mapsto \varphi(t,x)$ is the (unique) maximal solution of (E) (and is defined on $\mathbb{R}$ since F is bounded).

(iii) $0 \notin G(x)$ *for all* $x \in B(\limsup \mathrm{bd} M_k, r)$,
recalling that S *is the unit sphere of* $\mathbb{R}^n$.

Proof of Lemma 4.2. We let

$$G_\infty(x) = \mathrm{co}\Big(\mathrm{co}\Big(\limsup_{x' \to x, k \to \infty} N_{M_k}(x')\Big) \cap S\Big) \text{ for all } x \in \limsup \mathrm{bd} M_k.$$

Claim 4.1. *The correspondence* G_∞, *from* $\limsup \mathrm{bd} M_k$ *to* $\mathbb{R}^n$, *is u.s.c. with nonempty compact convex values and* $0 \notin G_\infty(x)$ *for every* $x \in \limsup \mathrm{bd} M_k$.

Proof of Claim 4.1. The correspondence $x \mapsto \limsup_{x' \to x, k \to \infty} N_{M_k}(x')$ has a closed graph, hence the correspondence $x \mapsto \limsup_{x' \to x, k \to \infty} N_{M_k}(x') \cap S$ is u.s.c., with compact values. We now show that it has nonempty values. Let $x \in \limsup \mathrm{bd} M_k$. Then there is a sequence (x_k) in $\mathbb{R}^n$ converging to x and an increasing map $\varphi : \mathbb{N} \to \mathbb{N}$ such that $x_k \in \mathrm{bd} M_{\varphi(k)}$ for all k. Then there is a sequence (p_k) in S such that $p_k \in N_{M_{\varphi(k)}}(x_k)$ for all k (see Clarke [4]). Let p be a cluster point of the sequence (p_k). Then $p \in \limsup_{x' \to x, k \to \infty} N_{M_k}(x') \cap S \subset G_\infty(x)$. We proved that the correspondence G_∞ is u.s.c., with nonempty compact convex values. Finally $0 \notin G_\infty(x)$, for all $x \in \limsup \mathrm{bd} M_k \subset M$, since $\mathrm{co} \limsup_{x' \to x, k \to \infty} N_{M_k}(x')$ is pointed. □

In view of the extension theorem of Cellina [3],[10] we let $\widehat{G_\infty}$ be an extension of G_∞ on $\mathbb{R}^n$, which is u.s.c., with nonempty compact convex values. For $r > 0$ we define the correspondence G_r from $\mathbb{R}^n$ to $\mathbb{R}^n$ by

$$G_r(x) = \mathrm{co}\, \overline{B}\Big(\widehat{G_\infty}(\overline{B}(x, r)), r\Big).$$

It has clearly nonempty convex values. Also, the correspondence G_r has compact values (since it is the sum of a compact set and of the convex hull of the image of a compact set by a u.s.c. correspondence). The correspondence G_r is clearly u.s.c. (recalling that, if Φ is a u.s.c. correspondence with convex values from $\mathbb{R}^n$ to $\mathbb{R}^n$ and if $r > 0$, then the correspondence $\mathrm{co}\Phi$, and the correspondences Φ_1 and Φ_2, defined by $\Phi_1(x) = \Phi(\overline{B}(x, r))$ and $\Phi_2(x) = \overline{B}(\Phi(x), r)$, respectively, are also u.s.c.).

Proof of (i). Since $M = \limsup M_k$ and since $\mathrm{co} \limsup_{x' \to x, k \to \infty} N_{M_k}(x')$ is pointed, then $N_M(x) \subset \mathrm{co} \limsup_{x' \to x, k \to \infty} N_{M_k}(x')$ for all $x \in \mathrm{bd} M$ (Lemma 4.1). Since $M = \limsup M_k$, then, from Lemma 4.1, $\mathrm{bd} M \subset \limsup \mathrm{bd} M_k$ (this can also be proved more directly). Then assertion (i) is a direct consequence of the definition of G_r.

Proof of (ii). We now prove that, for k large enough, $N_{M_k}(x) \cap S \subset G_r(x)$ for all $x \in \mathrm{bd} M_k$. Suppose that it is not true, then we may assume without any loss of generality that there are two sequences (x_k) and (p_k) in $\mathbb{R}^n$ such that, for all k, $x_k \in \mathrm{bd} M_k$, $p_k \in N_{M_k}(x) \cap S$, and $p_k \notin G_r(x_k)$. Since the sequence (x_k, p_k) belongs to the compact set $K \times S$, without any loss of generality, we may assume that it converges to an element $(x, p) \in K \times S$. Then $x \in M$ (since $M = \limsup M_k$) and $p \in \limsup_{x' \to x, k \to \infty} N_{M_k}(x')$. For k large enough, $x_k \in \overline{B}(x, r)$ and $p_k \in \overline{B}(p, r)$, hence $p_k \in \widehat{G_\infty}(\overline{B}(x_k, r)) + \overline{B}(0, r) \subset G_r(x_k)$, which is a contradiction. □

Proof of (iii). Since $\limsup \mathrm{bd} M_k \subset K$, it is compact. The end of the proof consists of choosing $r > 0$ as in Lemma 3.1 (taking $m = n$, considering the compact

[10] Let Φ be a u.s.c. correspondence, with nonempty compact convex values defined on a closed set $X \subset \mathbb{R}^n$, with values in $\mathbb{R}^m$. Then there is a u.s.c. correspondence $\hat{\Phi}$, with nonempty compact convex values, defined on $\mathbb{R}^n$ with values in $\mathbb{R}^m$, such that $\hat{\Phi}|_X = \Phi$ and such that $\hat{\Phi}(\mathbb{R}^n) \subset \mathrm{co}\Phi(X)$.

set $\limsup \mathrm{bd} M_k$ and the correspondence G_∞), recalling that $0 \notin G_\infty(x)$ if $x \in \limsup \mathrm{bd} M_k$ (Claim 4.1). □

The second lemma is a consequence of Lemma 4.2 and eliminates the situation where $M = \overline{B}(0,1)$ and $M_k = M \setminus B(0, 1/k)$ for $k \geq 1$ (see Remark 2.3).

LEMMA 4.3.

$$\text{(int)} \qquad \mathrm{int} M = \cup_{p \in \mathbb{N}} \mathrm{int}(\cap_{k \geq p} \mathrm{int} M_k).$$

Proof of Lemma 4.3. We recall that the inclusion $\cup_{p\in\mathbb{N}} \mathrm{int}(\cap_{k\geq p} \mathrm{int} M_k) \subset \mathrm{int} M$ is an immediate consequence of the equality $\liminf M_k = M$. We now prove the converse inclusion. Let $x \in \mathrm{int} M$. There is $\varepsilon > 0$ such that $B(x, \varepsilon) \subset M$, and from Lemma 3.1, such that $0 \notin \mathrm{co} G(B(x, \varepsilon))$. Assume that $x \notin \cup_{p\in\mathbb{N}} \mathrm{int}(\cap_{k\geq p} \mathrm{int} M_k)$. Then there is a sequence (x_k) converging to x and a subsequence $(M_{\varphi(k)})$ such that, for all k, $x_k \in \mathrm{bd} M_{\varphi(k)}$ (see the proof of Claim 3.1). Hence $x \in \limsup \mathrm{bd} M_{\varphi(k)} \subset \limsup \mathrm{bd} M_k$. From a classical separation argument, there is $p \in S$ and a real number $a > 0$ such that $(p|y) > a$ for all $y \in \mathrm{co} G(\overline{B}(x, \varepsilon))$. We may assume without any loss of generality that $a = \varepsilon$. Then, since $\mathrm{co} G(\overline{B}(x, \varepsilon))$ is bounded (it is the convex hull of the image of a compact set by a u.s.c. correspondence), there is $\varepsilon' > 0$ such that $(p'|y) > 0$ for all $p' \in B(p, \varepsilon')$ and for all $y \in \mathrm{co} G(B(x, \varepsilon))$. Without any loss of generality, we may assume that $\varepsilon' = \varepsilon$. Then $-p' \in \mathrm{int} T_{M_k}(x')$ for all $x' \in B(x, \varepsilon) \cap \mathrm{bd} M_k$ and for all $p' \in B(p, \varepsilon')$, if k is large enough since, from Lemma 4.2, $N_{M_k}(x') \cap S \subset G(x')$.

Then the following claim implies that, for k large enough, $x_k + tp' \notin M_{\varphi(k)}$ for all $p' \in B(p, \varepsilon/2)$ and all $t \in (0, \varepsilon/2)$, hence that $B(x + (\varepsilon/3)p, (\varepsilon^2/6)) \cap M_{\varphi(k)} = \emptyset$, hence that $x + (\varepsilon/3)p \notin \limsup M_{\varphi(k)}$, contradicting the fact that $M = \limsup M_{\varphi(k)}$.[11] □

CLAIM 4.2. *Let M be a closed epi-Lipschitzian subset of $\mathbb{R}^n$, let $x \notin \mathrm{int} M$, $\varepsilon > 0$, and $p \in S$ such that $-p \in \mathrm{int} T_M(x')$ for all $x' \in B(x, \varepsilon) \cap M$. Then $x + tp \notin M$ for $t \in (0, \varepsilon)$.*

Proof of Claim 4.2. Assume that $x + tp \in M$ for some $t \in [0, \varepsilon)$. Since $x \notin \mathrm{int} M$, we may assume without any loss of generality that $x + tp \in \mathrm{bd} M$. Then $x + tp \in B(x, \varepsilon)$, hence $-p \in \mathrm{int} T_M(x + tp)$. We recall that from Rockafellar [16]

$$\begin{aligned} \mathrm{int} T_M(x + tp) = \{\, v \in \mathbb{R}^n | \exists \alpha > 0, y + \lambda w \in M \\ \text{for all } (y, w, \lambda) \in (B(x + tp, \alpha) \cap M) \times B(w, \alpha) \times [0, \alpha)\}. \end{aligned}$$

Let $\alpha > 0$ be chosen as above. Then $(x + tp) - (\alpha/2)p \in M$. Hence $(x + tp) - (\alpha/2)p + (\alpha/2)p' \in M$ for all $p' \in B(p, \alpha)$; hence $x + tp \in \mathrm{int} M$, which is a contradiction. □

4.2.2. Proof of Proposition 4.2. We now prove that, for k large enough, there is a Lipschitzian transverse field between the sets M and M_k. We recall that $\mathrm{bd} M \subset \limsup \mathrm{bd} M_k \subset K$, hence that it is compact. We let $U = B(\mathrm{bd} M, \varepsilon)$ for a given real number $\varepsilon > 0$.

Proof of (a). Since $B(M, \varepsilon) = M \cup B(\mathrm{bd} M, \varepsilon)$ and since $B(\mathbb{R}^n \setminus \mathrm{int} M, \varepsilon) = (\mathbb{R}^n \setminus \mathrm{int} M) \cup B(\mathrm{bd} M, \varepsilon)$, the following claim clearly implies that $M_k \setminus \mathrm{int} M \subset B(\mathrm{bd} M, \varepsilon) = U$ and that $M \setminus \mathrm{int} M_k \subset U$, hence that $M_k \setminus \mathrm{int} M$ and $M \setminus \mathrm{int} M_k$ are compact, since $\overline{B}(\mathrm{bd} M, \varepsilon)$ is clearly compact.

[11] Recalling that the equality $M = \limsup M_k = \liminf M_k$ implies that $M = \limsup M_{\varphi(k)} = \liminf M_{\varphi(k)}$.

Claim 4.3. *Let ε be a positive real number. Then, for k large enough,*[12]

$$\begin{array}{lcl} M_k & \subset & B(M,\varepsilon); \\ \mathbb{R}^n \setminus \mathrm{int} M_k & \subset & B(\mathbb{R}^n \setminus \mathrm{int} M,\varepsilon). \end{array}$$

Proof of Claim 4.3. Assume that the inclusion $M_k \subset B(M,\varepsilon)$ does not hold for k large enough. Then, without loss of generality, we assume that there is a sequence (x_k) in $\mathbb{R}^n$ such that, for all k, $x_k \in M_k$ and $d(x_k,M) \geq \varepsilon$. If $\{x|d(x,M)=\varepsilon\} \in M_k$: $S(M,\varepsilon) \cap M_k \neq \emptyset$, we let $y_k \in S(M,\varepsilon)\cap M_k$. If $S(M,\varepsilon)\cap M_k = \emptyset$, there is $y_k \in \mathrm{bd}M_k$ such that $d(y_k,M) > \varepsilon$. Indeed, let $x \in S(M,\varepsilon)$ such that $\|x_k - x\| = d(x_k, B(M,\varepsilon))$. Then $x \notin M_k$, $x_k \in M_k$ and hence there is $y_k \in (x,x_k] \cap \mathrm{bd}M_k$, and $d(y_k, B(M,\varepsilon)) > 0$; hence $d(y_k,M) > \varepsilon$. Then the sequence (y_k) is in the compact set $S(M,\varepsilon)\cup K$ (since $S(M,\varepsilon) \subset S(\mathrm{bd}M,\varepsilon)$, which is compact, and since $\mathrm{bd}M_k \subset K$). We may assume without any loss of generality that it converges to some $x \in S(M,\varepsilon)\cup K$. Then $d(x,M) \geq \varepsilon$, but, since $y_k \in M_k$ for all k, and since $M = \limsup M_k$, we get that $x \in M$, which is a contradiction. To get the second inclusion, apply the first result to the sets $\mathbb{R}^n \setminus \mathrm{int}M$ and $\mathbb{R}^n \setminus \mathrm{int}M_k$, noticing that assertion (int) (see Lemma 4.3) implies that $\mathbb{R}^n \setminus \mathrm{int}M = \limsup(\mathbb{R}^n \setminus \mathrm{int}M_k)$, that $\mathrm{bd}(\mathbb{R}^n \setminus \mathrm{int}M) = \overline{\mathrm{int}M} \setminus \mathrm{int}M \subset \mathrm{bd}M$, and that $\mathrm{bd}(\mathbb{R}^n \setminus \mathrm{int}M_k) = \overline{\mathrm{int}M_k} \setminus \mathrm{int}M_k \subset \mathrm{bd}M_k \subset K$. □

Proof of (b). We let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a map satisfying the conclusions of the following lemma, which is a slightly different version of Lemma 3.1 of Bonnisseau–Cornet [2] (its proof is left to the reader). Then, in view of Lemma 4.2, if we choose $\varepsilon < r'$, assertion (b) follows.

Lemma 4.4. *There is $r' > 0$ and a bounded locally Lipschitzian map $F : \mathbb{R}^n \to \mathbb{R}^n$, such that*

$$\forall\ x \in B(\mathrm{bd}M, r'),\ \forall\ y \in G(x),\ (F(x)|y) > r'.$$

Proof of (c). We let $f : \mathbb{R}^n \to \mathbb{R}$ be a quasi-smooth representation of M satisfying the conclusions of Theorem 3.1, i.e., such that $f^{-1}([-\varepsilon_0,\varepsilon_0])$ is compact for some $\varepsilon_0 > 0$. Then from Lemma 4.4, since $\partial f(x) \subset N_M(x)$ for all $x \in \mathrm{bd}M$, since the correspondence ∂f is u.s.c., with nonempty convex compact values, and since $\mathrm{bd}M$ is compact, there is $r'' > 0$ such that

$$(F(x)|y) > r'' \quad \forall\ x \in \mathrm{bd}M \text{ and } \forall\ y \in \partial f(x).$$

Then assertion (c) holds, if we choose $\varepsilon > 0$ given by the following claim. □

Claim 4.4. *Let f be a quasi-smooth representation of an epi-Lipschitzian subset $M \subset \mathbb{R}^n$ such that $f^{-1}([-r,r])$ is compact for some $r > 0$, and such that $(F(x)|y) > r$ for all $x \in \mathrm{bd}M$ and for all $y \in \partial f(x)$. Then there are two real numbers $\varepsilon > 0$ and $\alpha > 0$ such that*

(i) $B(\mathrm{bd}M,\varepsilon) \subset f^{-1}([-\alpha,\alpha]) \subset B(\mathrm{bd}M,r)$;

(ii) $\big(\partial f(\varphi(t,x))|(\partial\varphi/\partial t)(t,x)\big) \subset [r,+\infty)\ \forall\ x \in B(\mathrm{bd}M,r)$ *and* $\forall\ t \in \mathbb{R}$ *such that* $|f(\varphi(t,x))| \leq \alpha$;

(iii) *the function* $f \circ \varphi(.,x)$ *is strictly increasing on* $\{t \in \mathbb{R}|\ |f(\varphi(t,x))| < \alpha\}\ \forall\ x\ \in B(\mathrm{bd}M,r)$;

(iv) *there are* t *and* t' *in* $\mathbb{R}$ *such that* $f(\varphi(t,x)) \leq -\alpha$ *and* $f(\varphi(t',x)) \geq \alpha\ \ \forall\ x \in f^{-1}([-\alpha,\alpha])$,

[12] Note that these inclusions imply that $\delta(M,M_k) \leq \varepsilon$ and that $\delta(\mathrm{bd}M,\mathrm{bd}M_k) \leq \varepsilon$, defining $\delta(X,Y) = \max\{\sup_{x\in X} d(x,Y), \sup_{y \in Y} d(y,X)\}$ if X and Y are two subsets of $\mathbb{R}^n$ (not necessarily nonempty compact), hence that $\delta(M,M_k) \to 0$ and that $\delta(\mathrm{bd}M,\mathrm{bd}M_k) \to 0$. Conversely, the assumption that $\delta(M,M_k) \in \mathbb{R}$ and converges to 0 implies that $M = \liminf M_k = \limsup M_k$.

recalling that $\varphi : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ *is the flow of the differential equation* (E) $\dot{x}(t) = F(x(t))$, $x(0) = x$, *where* $\varphi(., x)$ *is defined on* $\mathbb{R}$.

Proof of Claim 4.4. Since the correspondence ∂f is u.s.c. with compact convex values and since the map F is continuous, the correspondence H defined by $H(x) = (F(x)|\partial f(x))$ is u.s.c., with compact convex values. Then, from Lemma 3.1, there is $\beta > 0$ such that $B(\mathrm{bd}M, \beta) \subset U$ and such that

$$(F(x)|y) \quad > \quad r \qquad \text{for all } x \in B(\mathrm{bd}M, \beta) \text{ and for all } y \in \partial f(x).$$

Then there is $\alpha > 0$ such that $f^{-1}([-\alpha, \alpha]) \subset B(\mathrm{bd}M, \beta)$ (since $B(\mathrm{bd}M, \beta)$ is an open set containing the intersection of compact sets $\cap_{\alpha \in (0, \varepsilon_0]} f^{-1}([-\alpha, \alpha))$. Without any loss of generality, we may assume that $f^{-1}([-\alpha, \alpha]) \subset B(\mathrm{bd}M, r)$. Since $f^{-1}((-\alpha, \alpha))$ is an open set containing the intersection of compact sets $\cap_{\epsilon > 0} \overline{B}(\mathrm{bd}M, \epsilon)$, there is $\varepsilon > 0$ such that $B(\mathrm{bd}M, \varepsilon) \subset f^{-1}([-\alpha, \alpha])$, which proves (i). Let $x \in B(\mathrm{bd}M, r)$. We define the function $h_x : \mathbb{R} \to \mathbb{R}$ by $h_x(t) = f(\varphi(t, x))$. Then, from Clarke [4],

$$\partial h_x(t) \subset \Big(\partial f(\varphi(t, x)) | (\partial \varphi / \partial t)(t, x)\Big),$$

i.e., $\partial h_x(t) \subset \Big(\partial f(\varphi(t, x)) | F(\varphi(t, x))\Big)$, which is minorized by r if $|f(\varphi(t, x))| < \alpha$; this proves (ii). Hence from the mean-value theorem, the function $f \circ \varphi(., x) = h_x$ is strictly increasing on $\{t \in \mathbb{R} |\, |f(\varphi(t, x))| < \alpha\}$, which proves (iii). If we additionally assume that $x \in f^{-1}([-\alpha, \alpha])$, then (see, for example, Hirsch and Smale [9]; the fact can also be proved directly) $\varphi(t, x) \notin f^{-1}([-\alpha, \alpha])$ when t is large enough; hence, from (ii), $f(\varphi(t, x)) \geq \alpha$ when $t \to +\infty$. In the same way, $f(\varphi(t, x)) \leq -\alpha$ when $t \to -\infty$. □

4.3. The sets M and M_k are lipeomorphic. In view of Propositions 4.1 and 4.2, the proof of Theorem 2.3 is finished if we prove the next proposition.

PROPOSITION 4.3. *Let M and N be two closed subsets of* $\mathbb{R}^n$ *admitting a Lipschitzian transverse field in the sense of* (T). *Then M and N are epi-Lipschitzian and lipeomorphic.*

Remark 4.2. One could reduce Proposition 4.3 to the smooth case. Indeed, since $\mathrm{bd}M$ and $\mathrm{bd}N$ are compact, there are two smooth normal approximations (M_k) and (N_k) of M and N, respectively. Then, for k large enough, there is a Lipschitzian transverse field between the two sets M_k and N_k in the sense of (T). But the proof would be identical. Proposition 4.3 can be proved directly, without using the notion of normal approximation, as we do in the following.

Proof of Proposition 4.3. The set M is clearly epi-Lipschitzian, since the transversality condition (T) implies that, for every $x \in \mathrm{bd}M$, $\mathrm{int}T_M(x) \neq \emptyset$, hence that $N_M(x)$ is pointed. Similarly, the set N is also epi-Lipschitzian. Let f_M (resp., f_N) be an inequality representation of M (resp., N) satisfying the conclusions of Theorem 3.1, i.e., such that $f_M^{-1}([-\varepsilon_0, \varepsilon_0])$ (resp., $f_N^{-1}([-\varepsilon_0, \varepsilon_0])$) is compact for some $\varepsilon_0 > 0$. The following lemma is a different version of Lemma 3.2 of Bonnisseau–Cornet [2] that we prove for the sake of completeness.

LEMMA 4.5. *There is a real number $\beta > 0$ and two Lipschitzian functions τ and θ, defined from $U \times [-\beta, \beta]$ to* $\mathbb{R}^n$, *such that*

$$\text{(i)} \quad \Big\{x \in \mathbb{R}^n |\min\{f_M(x), f_N(x)\} \leq \beta, \max\{f_M(x), f_N(x)\} \geq -\beta\Big\} \subset U$$

and such that, for all $(x, \delta, t) \in U \times [-\beta, \beta] \times \mathbb{R}$, then

(ii) $f_M(\varphi(x,t)) = \delta \Leftrightarrow t = \tau(x,\delta)$;
(iii) $f_N(\varphi(x,t)) = \delta \Leftrightarrow t = \theta(x,\delta)$.

Proof of Lemma 4.5. Note that, without any loss of generality, we only need to prove assertions (i) and (ii). Since the set

$$\Big\{x \in \mathbb{R}^n | \min\{f_M(x), f_N(x)\} \leq 0, \max\{f_M(x), f_N(x)\} \geq 0\Big\}$$

is equal to the set $[M \setminus \mathrm{int}N] \cup [N \setminus \mathrm{int}M]$, which is included in the set U, we get the inclusion

$$\cap_{\beta\in(0,\varepsilon_0]}\Big\{x \in \mathbb{R}^n | \min\{f_M(x), f_N(x)\} \leq \beta, \max\{f_M(x), f_N(x)\} \geq -\beta\Big\} \subset U.$$

This implies assertion (i) for some $\beta > 0$ (since we have an intersection of compact sets included in U). Then from the assumption (T), (b), there is $r > 0$ such that $(F(x)|y) > r$ for all $x \in \mathrm{bd}M$ and for all $y \in \partial f_M(x)$. We may assume that β is small enough and we thus satisfy the conclusions of Claim 4.4, given the function f_M. Then $f_M^{-1}([-\beta,\beta]) \subset U$, and the function $f_M \circ \varphi(.,x)$ is strictly increasing on $\{t \in \mathbb{R}| \, |f_M(\varphi(t,x))| < \beta\}$ for every $x \in \mathbb{R}^n$. Let us now consider $x \in U$. From the assumption (T), (c), there is $t \in \mathbb{R}$ such that $\varphi(x,t) \in \mathrm{bd}M$, i.e., $f_M(\varphi(x,t)) = 0$. Let $(\delta,t) \in [-\beta,\beta]\times\mathbb{R}$. Then there is one and only one $t' \in \mathbb{R}$ such that $f_M(\varphi(t',x)) = \delta$. We let $t' = \tau(x,\delta)$. Let us now prove that the function τ is Lipschitzian. We let

$$\Omega = \{(x,t,\delta) \in \Omega\times\mathbb{R}| \, |f_M(\varphi(t,x))| < \beta\},$$

and we define $H : \Omega \to \mathbb{R}$ by $H(x,t,\delta) = f_M(\varphi(t,x)) - \delta$. The function H is clearly Lipschitzian. The fact that τ is Lipschitzian around some (x,δ) is a direct consequence of the implicit function theorem [4, p. 255], if we prove that $t^* \neq 0$, for every element $(x^*,t^*,\delta^*) \in \partial H(x,t,\delta)$, where $t \in \mathbb{R}$ satisfies $H(x,t,\delta) = 0$. In other words, if $\pi_t : \mathbb{R}^n\times\mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is the projection defined by $\pi_t(x,t,\delta) = t$, if $0 \notin \pi_t[\partial H(x,t,\delta)]$. From Clarke [4, Proposition 2.6.2], we get that

$$\pi_t[\partial H(x,t,\delta)] \subset \Big(\partial f_M(\varphi(t,x))|F(\varphi(t,x))\Big),$$

which is minorized by β if we chose β small enough (Claim 4.4). □

We then get the following property on the functions τ and θ.

Claim 4.5. *For all* $(x,\delta,t) \in U\times[-\beta,\beta]\times\mathbb{R}$

(i) $f_M(\varphi(x,t)) \geq \delta \Leftrightarrow t \geq \tau(x,\delta)$;
(ii) $f_N(\varphi(x,t)) \geq \delta \Leftrightarrow t \geq \theta(x,\delta)$;

If $y \in U$ *is such that* $\varphi(x,t) = y$ *for some* t, *then* $\tau(x,\delta) = \tau(y,\delta) + t$ *and* $\theta(x,\delta) = \theta(y,\delta) + t$ *for all* δ.

Proof of Claim 4.5. The proof comes from Lemma 4.5 and from the Cauchy–Lipschitz theorem for φ, since $f_M\circ\varphi(x,.)$ is strictly increasing on $\{t \in \mathbb{R}||f_M(\varphi(t,x))| < \beta\}$. □

We are now able to build the lipeomorphism between M and N. Let us define the map $h : M \to N$ by

$$\begin{array}{lcll} h(x) & = & x & \text{if } x \in \mathrm{int}(M_{-\beta}\cap N_{-\beta}),\\ h(x) & = & \varphi\Big(x, i_x(\tau_x - \theta_x)/(\tau_x - i_x)\Big) & \text{if } x \in M \setminus \mathrm{int}(M_{-\beta}\cap N_{-\beta}), \end{array}$$

where we let $M_{-\beta} = \{x \in \mathbb{R}^n | f_M(x) \leq -\beta\}$, $N_{-\beta} = \{x \in \mathbb{R}^n | f_N(x) \leq -\beta\}$, $\tau_x = \tau(x,0)$, $\theta_x = \theta(x,0)$, and $i_x = \inf\{\tau(x,-\beta), \theta(x,-\beta)\}$. Let $x \in M \setminus \mathrm{int}(M_{-\beta} \cap N_{-\beta})$; then $f_M(x) \leq 0$ and $\max\{f_M(x), f_N(x)\} \geq -\beta$; hence $x \in U$ from Lemma 4.5. Hence the map h is well defined. The map h has its values in N. Indeed, let $x \in M \setminus \mathrm{int}(M_{-\beta} \cap N_{-\beta})$. Then, recalling that $x = \varphi(x,0)$,

$$\begin{array}{rl} & f_M[\varphi(x,0)] = f_M(x) \geq -\beta = f_M[\varphi(x,\tau(x,-\beta))], \\ \text{or} & f_N[\varphi(x,0)] = f_N(x) \geq -\beta = f_N[\varphi(x,\theta(x,-\beta))]; \end{array}$$

hence, from Claim 4.5, and $0 \geq \tau(x,-\beta)$ or $0 \geq \theta(x,-\beta)$, which implies that $i_x \leq 0$. The same claim implies that $i_x < \tau_x$ (since $f_M[\varphi(x,\tau(x,-\beta))] < f_M[\varphi(x,\tau_x)]$), $i_x < \theta_x$ (since $f_N[\varphi(x,\theta(x,-\beta))] < f_N[\varphi(x,\theta_x)]$), and $\tau_x \geq 0$ (since $f_M[\varphi(x,0)] = f_M(x) \leq 0 = f_M[\varphi(x,\tau_x)]$). Then we get that $i_x(\tau_x - \theta_x)/(\tau_x - i_x) \leq \theta_x$, which proves that $f_N(h(x)) \leq 0$, i.e., $h(x) \in N$.

Claim 4.6. *The map h is locally Lipschitzian.*

Proof of Claim 4.6. The map h is clearly Lipschitzian on $\mathrm{int}(M_{-\beta} \cap N_{-\beta})$. From Lemma 4.5, the map $x \mapsto \inf\{\tau(x,-\beta), \theta(x,-\beta)\}$ is Lipschitzian and $\tau(x,0) - \inf\{\tau(x,-\beta), \theta(x,-\beta)\} \neq 0$ for all $x \in M \setminus (M_{-\beta} \cap N_{-\beta})$. Then clearly h is Lipschitzian on $M \setminus \mathrm{int}(M_{-\beta} \cap N_{-\beta})$. If $x \in \mathrm{bd}(M_{-\beta} \cap N_{-\beta})$, then $f_M(x) = -\beta$ or $f_N(x) = -\beta$; besides, $f_M(x) \leq -\beta$ and $f_N(x) \leq -\beta$. This implies that $\tau(x,-\beta) = 0$ or $\theta(x,-\beta) = 0$, and that $\tau(x,-\beta) \geq 0$ and $\theta(x,-\beta) \geq 0$. Hence $\inf\{\tau(x,-\beta), \theta(x,-\beta)\} = 0$ and $h(x) = x$. This proves that h is Lipschitzian on $M_{-\beta} \cap N_{-\beta}$. Hence h is Lipschitzian on M. □

Claim 4.7. *The map h is one to one, and h^{-1} is Lipschitzian.*

Proof of Claim 4.7. Let us define the map $k : N \to M$ by

$$\begin{array}{rcll} k(x) & = & x & \text{if } x \in \mathrm{int}(M_{-\beta} \cap N_{-\beta}); \\ k(x) & = & \varphi\Big(x, i_x(\theta_x - \tau_x)/(\theta_x - i_x)\Big) & \text{if } x \in N \setminus \mathrm{int}(M_{-\beta} \cap N_{-\beta}). \end{array}$$

We let the reader check that k has its values in M and that it is Lipschitzian. Let us now prove that $h \circ k = id_N$. Let us consider $x \in M$. We let $y = k(x)$; then $i_y = i_x - t$, $\tau_y = \tau_x - t$, and $\theta_y = \theta_x - t$, where $t = i_x(\theta_x - \tau_x)/(\theta_x - i_x)$. Hence $i_y(\tau_y - \theta_y)/(\tau_y - i_y) = i_x(\tau_x - \theta_x)/(\theta_x - i_x)$, and hence $h(y) = \varphi(y, i_y(\tau_y - \theta_y)/(\tau_y - i_y)) = x$. The fact that $k \circ h = id_M$ goes in the same way. □

Acknowledgment. We wish to thank an anonymous referee for suggesting that we consider the link between normal convergence and the one considered in section 2.4.2.

REFERENCES

[1] J. Benoist, *Approximation and regularization of arbitrary sets in finite dimension*, Set-Valued Anal., 2 (1994), pp. 95–115.

[2] J.-M. Bonnisseau and B. Cornet, *Fixed-point theorem and Morse's lemma for Lipschitzian functions*, J. Math. Anal. Appl., 146 (1990), pp. 318–322.

[3] A. Cellina, *A theorem on the approximation of compact multi-valued mappings*, Atti Accad. Naz. Lincei Cl. Sci. Fis., Mat. Natur. Rend. (8), 47 (1969), pp. 429–433.

[4] F. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983. Reprinted in Classics Appl. Math. 5 (1989), SIAM, Philadelphia, PA, 1989.

[5] B. Cornet, *Regularity properties of open tangent cones*, Math. Prog. Study, 30 (1987), pp. 17–33.

[6] B. Cornet and M.-O. Czarnecki, *Smooth representations of epi-Lipschitzian subsets of* $\mathbb{R}^n$, Nonlinear Anal. (1998), to appear. French abridged version in C. R. Acad. Sci. Sér. 1, 325 (1997), pp. 475–480.

[7] B. Cornet and M.-O. Czarnecki, *Existence of (generalized) equilibria on an epi-Lipschitzian domain: A necessary and sufficient condition*, Cahier Eco-Maths 95-55, Université de Paris 1, Paris, France, 1995.

[8] P. Doktor, *Approximation of domains with Lipschitzian boundary*, Casopis Pest. Mat., 101 (1976), pp. 237–255.

[9] M.W. Hirsch and S. Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1960.

[10] A.D. Ioffe, *Approximate subdifferentials and applications* I: *The finite dimensional theory*, Trans. Amer. Math. Soc., 281 (1984), pp. 389–416.

[11] A.Y. Kruger and B.S. Mordukhovich, *Generalized normals and derivatives, and necessary optimality conditions in nondifferentiable programming, Depon. VINITI*, Nos. 408–80 and 494–80, Moscow, 1980 (in Russian).

[12] K. Kuratowski, *Topologie*, Państwowe Wydawnictwo Naukowe, Warszawa, 1958.

[13] U. Massari and L. Pepe, *Sull'approssimazione degli aperti lipschitziani di* $\mathbb{R}^n$ *con varietà differenziabili*, Boll. Un. Mat. Ital., Ser. 4, 10 (1974), pp. 532–544.

[14] B.S. Mordukhovich, *Maximum principle in the problem of time optimal response with nonsmooth constraints*, J. Appl. Math. Mech. 40 (1976), pp. 960–969.

[15] J. Nečas, *On type* $\mathcal{M}$ *domains*, Czechoslovak Math. J., 12 (1962), pp. 274–287 (in Russian).

[16] R.T. Rockafellar, *Clarke's tangent cones and the boundaries of closed sets in* $\mathbb{R}^n$, Nonlinear Anal., 3 (1979), pp. 145–154.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 3, pp. 731–764

GLOBAL CONVERGENCE OF TRUST-REGION INTERIOR-POINT ALGORITHMS FOR INFINITE-DIMENSIONAL NONCONVEX MINIMIZATION SUBJECT TO POINTWISE BOUNDS*

MICHAEL ULBRICH†, STEFAN ULBRICH‡, AND MATTHIAS HEINKENSCHLOSS§

Abstract. A class of interior-point trust-region algorithms for infinite-dimensional nonlinear optimization subject to pointwise bounds in L^p-Banach spaces, $2 \le p \le \infty$, is formulated and analyzed. The problem formulation is motivated by optimal control problems with L^p-controls and pointwise control constraints. The interior-point trust-region algorithms are generalizations of those recently introduced by Coleman and Li [*SIAM J. Optim.*, 6 (1996), pp. 418–445] for finite-dimensional problems. Many of the generalizations derived in this paper are also important in the finite-dimensional context. All first- and second-order global convergence results known for trust-region methods in the finite-dimensional setting are extended to the infinite-dimensional framework of this paper.

Key words. infinite-dimensional optimization, bound constraints, affine scaling, interior-point algorithms, trust-region methods, global convergence, optimal control, nonlinear programming

AMS subject classifications. 49M37, 65K05, 90C30, 90C48

PII. S0363012997319541

1. Introduction.

This paper is concerned with the development and analysis of a class of interior-point trust-region algorithms for the solution of the following infinite-dimensional nonlinear programming problem:

$$\text{(P)} \qquad \begin{array}{ll} \text{minimize} & f(u) \\ \text{subject to} & u \in \mathcal{B} \overset{\text{def}}{=} \{u \in L^p(\Omega) \,:\, a(x) \le u(x) \le b(x),\ x \in \Omega\}. \end{array}$$

Here $\Omega \subset \mathbb{R}^n$ is a domain with positive and finite Lebesgue measure $0 < \mu(\Omega) < \infty$, and the objective function $f : \mathcal{D} \longrightarrow \mathbb{R}$ is continuous on an open neighborhood $\mathcal{D} \subset L^p(\Omega)$ of $\mathcal{B}$. All pointwise statements on measurable functions are meant to hold μ-almost everywhere. The lower and upper bound functions $a, b \in L^\infty(\Omega)$, are assumed to have a distance of at least $\nu > 0$ from each other. More precisely, $b(x) - a(x) \ge \nu$ on Ω.

Problems of type (P) arise, for instance, when the black-box approach is applied to optimal control problems with bound-constrained L^p-control. A typical example is the boundary control of a heat equation with a Stefan–Boltzmann boundary condition

*Received by the editors April 7, 1997; accepted for publication (in revised form) June 18, 1998; published electronically February 19, 1999.

http://www.siam.org/journals/sicon/37-4/31954.html

†Lehrstuhl für Angewandte Mathematik und Mathematische Statistik, Technische Universität München, D-80290 München, Germany (mulbrich@statistik.tu-muenchen.de). The research of this author was supported by the DFG under grant Ul157/1–1 and by NATO under grant CRG 960945.

‡Lehrstuhl für Angewandte Mathematik und Mathematische Statistik, Technische Universität München, D-80290 München, Germany (sulbrich@statistik.tu-muenchen.de). The research of this author was supported by the DFG under grant Ul158/1–1 and by NATO under grant CRG 960945.

§Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005–1892 (heinken@rice.edu). The research of this author was supported by the NSF under grant DMS–9403699, the DOE under grant DE-FG03-95ER25257, the AFSOR under grant F49620–96–1–0329, and NATO under grant CRG 960945.

given by

$$\begin{array}{ll} \text{minimize} & f(u) \stackrel{\text{def}}{=} \frac{1}{2}\|y(u)(T,.) - y_d\|_2^2 + \frac{\alpha}{2}\|u\|_2^2 \\ \text{subject to} & 0 \le u(t) \le 1,\ t \in [0,T], \end{array}$$

where $\alpha \ge 0$ is a given parameter and $y = y(u)$ is the solution of

$$\begin{aligned} y_t(t,x) &= y_{xx}(t,x), \quad (t,x) \in (0,T) \times (0,1) \\ y(0,x) &= 0, \quad x \in (0,1) \\ y_x(t,0) &= 0, \quad y_x(t,1) = -y(t,1)^4 + u(t), \quad t \in (0,T). \end{aligned} \tag{1.1}$$

Another example is an optimal control problem of Bolza type given by

$$\begin{array}{ll} \text{minimize} & f(u) \stackrel{\text{def}}{=} P(y(u)(1)) + \int_0^1 h_0(x, y(u)(x), u(x))\, dx \\ \text{subject to} & u \in \mathcal{B}, \end{array}$$

where $y = y(u)$ is the solution of

$$\frac{dy}{dx}(x) = h(x, y(x), u(x)), \quad y(0) = y^0. \tag{1.2}$$

These control problems and similar ones are studied, e.g., by Burger and Pogu [3], Kelley and Sachs [14], Sachs [18], and Tian and Dunn [21]. We will return to these two specific examples in section 8.

The algorithms in this paper are extensions of the interior-point trust-region algorithms for bound constrained problems in $\mathbb{R}^N$ introduced by Coleman and Li [6]. Algorithmic enhancements of these methods have been proposed and analyzed in the finite-dimensional context in Branch, Coleman, and Li [2], Coleman and Li [5], and Dennis and Vicente [11]. Dennis, Heinkenschloss, and Vicente [10] and Heinkenschloss and Vicente [13] extend these methods to solve a class of finite-dimensional constrained optimization problems with bound constraints on parts of the variables. See also Vicente [24]. The interior-point trust-region methods in [6] are based on the reformulation of the Karush–Kuhn–Tucker (KKT) necessary optimality conditions as a system of nonlinear equations using a diagonal matrix D. This affine-scaling matrix is computed using the sign of the gradient components and the distance of the variables to the bounds (see section 2). The nonlinear system is then solved by an affine-scaling interior-point method in which the trust region is scaled by $D^{-\frac{1}{2}}$. These methods enjoy strong theoretical convergence properties as well as a good numerical behavior. The latter is documented in [2], [6], [10], [11], where these algorithms have been applied to various standard finite-dimensional test problems and to some discretized optimal control problems.

The present work is motivated by the application of interior-point trust-region algorithms to optimal control problems with bounds on the controls. Even though the numerical solution of these problems requires a discretization and allows the application of the previously mentioned algorithms to the resulting finite-dimensional problems, it is known that the infinite-dimensional setting dominates the convergence behavior if the discretization becomes sufficiently small. If the algorithm can be applied to the infinite-dimensional problem and convergence can be proven in the infinite-dimensional setting, asymptotically the same convergence behavior can be expected if the algorithm is applied to the finite-dimensional discretized problems. Otherwise, the convergence behavior might—and usually does—deteriorate fast as the discretization is refined.

In the present context, the formulation of the interior-point trust-region algorithms for the solution of the infinite-dimensional problem (P) requires a careful statement of the problem and of the requirements on the function f. This will be done in section 4. The infinite-dimensional problem setting in this paper is similar to the ones in [12], [14], [15], [21]. The general structure of the interior-point trust-region algorithms presented here is closely related to the finite-dimensional algorithms in [6]. However, the statement and analysis of the algorithm in the infinite-dimensional context is more delicate and has motivated generalizations and extensions which are also relevant in the finite-dimensional context. The analysis performed in this paper allows for a greater variety of choices for the affine-scaling matrix and the scaling of the trust region than those presented previously in [6], [11]. Our convergence analysis is more comprehensive than the ones in [5], [6], [11], [24]. In particular, we adapt techniques proposed in Shultz, Schnabel, and Byrd [19] to prove that under mild assumptions every accumulation point satisfies the second-order necessary optimality conditions. Moreover, the convergence results proven in this paper extend all the finite-dimensional ones stated in [17], [19], [20] to our infinite-dimensional context with bound constraints. In the follow-up paper [23] we present a local convergence analysis of a superlinearly convergent affine-scaling interior-point Newton method which is based on (5.3), and we prove under appropriate assumptions that in a neighborhood of the solution the generated trial steps are accepted by our trust-region algorithms. There a projection onto the set $\mathcal{B}$ is used in the computation of trial steps.

Trust-region methods for infinite-dimensional problems like (P) have also been investigated by Kelley and Sachs [15] and Toint [22]. In both papers the constraints are handled by projections. The paper [22] considers trust-region algorithms for minimization on closed convex bounded sets in Hilbert space. They are extensions of the finite-dimensional algorithms by Conn, Gould, and Toint [7]. It is proven that the projected gradient converges to zero. A comprehensive finite-dimensional analysis of trust-region methods closely related to those introduced by Toint can be found in Burke, Moré, and Toraldo [4]. In contrast to the results in [22], our convergence analysis is also applicable to objective functions that are merely differentiable on a Banach space $L^p(\Omega)$, $p \in (2, \infty]$, which reduces the differentiability requirements substantially compared to the L^2-Hilbert space framework. Furthermore, for the problem class under consideration our convergence results are more comprehensive than the ones in [22]. The infinite-dimensional setting used in [15] fits into the framework of this paper but is more restrictive. The formulation of their algorithm depends on the presence of a penalty term $\alpha \int_\Omega u^2(x)dx$ in the objective function f, and they assume that $\Omega \subset \mathbb{R}$ is an interval. Their algorithm also includes a postsmoothing step, which is performed after the trust-region step is computed. The presence of the postsmoothing step ensures that existing local convergence results can be applied. Such postsmoothing is not needed in the global analysis of this paper.

This paper is organized as follows. In the next section we review the basics of the finite-dimensional interior-point trust-region algorithms in [6]. This is used in section 3 to motivate the infinite-dimensional setting applied in this paper. This section also contains some basic technical results. In section 4 we formulate the necessary optimality conditions in the framework needed for the interior-point trust-region algorithms, which are introduced in section 5. The main convergence results are given in section 6, which concerns the global convergence to points satisfying the first-order necessary optimality conditions, and in section 7, which concerns the global convergence to points satisfying the second-order necessary optimality conditions. These convergence results extend all the known convergence results for trust-region methods in finite dimensions to the infinite-dimensional setting of this paper. Some optimal control examples from the literature to which our analysis applies are discussed in section 8. The local convergence analysis of these algorithms is given in the follow-up paper [23], which also contains numerical examples illustrating the theoretical findings of this paper.

2. Review of the finite-dimensional algorithm.

We briefly review the main ingredients of the affine-scaling interior-point trust-region method introduced in [6]. We refer to that paper for more details. The algorithm solves finite-dimensional problems of the form

$$(\mathrm{P_N}) \qquad \begin{aligned} &\text{minimize} \quad f(u) \\ &\text{subject to} \quad u \in \mathcal{B}_N \overset{\text{def}}{=} \left\{ u \in \mathbb{R}^N \, : \, a \le u \le b \right\}, \end{aligned}$$

where $f : \mathbb{R}^N \longrightarrow \mathbb{R}$ is a twice continuously differentiable function and $a < b$ are given vectors in $\mathbb{R}^N$. (One can allow components of a and b to be $-\infty$ or ∞, respectively. This is excluded here to simplify the presentation.) Inequalities are understood componentwise.

The necessary optimality conditions for $(\mathrm{P_N})$ are given by

$$\begin{aligned} &\nabla f(\bar{u}) - \bar{\mu}^a + \bar{\mu}^b = 0, \\ &a \le \bar{u} \le b, \\ &(\bar{u} - a)^T \bar{\mu}^a + (b - \bar{u})^T \bar{\mu}^b = 0, \\ &\bar{\mu}^a \ge 0, \ \bar{\mu}^b \ge 0. \end{aligned}$$

With the diagonal matrix defined by

$$(2.1) \qquad \big(D(u)\big)_{ii} \overset{\text{def}}{=} \begin{cases} (b-u)_i & \text{if} \quad \big(\nabla f(u)\big)_i < 0, \\ (u-a)_i & \text{if} \quad \big(\nabla f(u)\big)_i \ge 0 \end{cases}$$

for $i = 1, \ldots, N$, the necessary optimality conditions can be rewritten as

$$(2.2) \qquad \begin{gathered} D(\bar{u})^r \nabla f(\bar{u}) = 0, \\ a \le \bar{u} \le b, \end{gathered}$$

where the power $r > 0$ is applied to the diagonal elements. This form of the necessary optimality conditions—we choose $r = 1$—can now be solved using Newton's method. The ith component of the function $D(u)$ is differentiable except at points where $(\nabla f(u))_i = 0$. However, this lack of smoothness is benign since $D(u)$ is multiplied by $\nabla f(u)$. One can use

$$(2.3) \qquad D(u)\nabla^2 f(u) + E(u)$$

as the approximate derivative of $D(u)\nabla f(u)$, where $E(u) = D'(u)\mathrm{diag}(\nabla f(u))$ and $D'(u)$ is the diagonal matrix

$$\big(D'(u)\big)_{ii} \overset{\text{def}}{=} \begin{cases} -1 & \text{if} \quad \big(\nabla f(u)\big)_i < 0, \\ 1 & \text{if} \quad \big(\nabla f(u)\big)_i > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Here $D'(u)\text{diag}(v)$, $v \in \mathbb{R}^N$, plays the role of the generally nonexistent derivative of $D(u)v$ with respect to u, but the prime is not meant to indicate differentiation.

After symmetrization, one obtains

$$\hat{M}(u) = D(u)^{1/2}\nabla^2 f(u) D(u)^{1/2} + E(u). \tag{2.4}$$

One can show that the standard second-order necessary optimality conditions are equivalent to (2.2) and the positive semidefiniteness of $\hat{M}(\bar{u})$. The standard second-order sufficient optimality conditions are equivalent to (2.2) and the positive definiteness of $\hat{M}(\bar{u})$.

A point satisfying the necessary optimality conditions (2.2) is now computed using the iteration $u_{k+1} = u_k + s_k$, where for a given u_k with $a < u_k < b$, the trial step $s_k = D_k^{1/2}\hat{s}_k$ satisfies $a < u_k + s_k < b$, and $\hat{s}_k$ is an approximate solution of

$$\min \hat{\psi}_k(\hat{s}) \quad \text{subject to} \quad \|\hat{s}\|_2 \le \Delta_k \ , \quad u_k + D_k^{1/2}\hat{s} \in \mathcal{B}_N, \tag{2.5}$$

with $\hat{\psi}_k(\hat{s}) \overset{\text{def}}{=} \hat{g}_k^T \hat{s} + \frac{1}{2}\hat{s}^T \hat{M}_k \hat{s}$, $\hat{g}_k \overset{\text{def}}{=} D_k^{1/2}\nabla f_k$. The trust-region radius Δ_k is updated from iteration to iteration in the usual fashion. In (2.5) the Hessian $\nabla^2 f(u_k)$ might be replaced by a symmetric approximation B_k. If the approximate solution $\hat{s}_k$ of (2.5) satisfies a fraction of Cauchy decrease condition

$$\begin{aligned} &\hat{\psi}_k(\hat{s}_k) < \beta \min\left\{\hat{\psi}_k(\hat{s}) \ : \ \hat{s} = -t\hat{g}_k,\ t \ge 0,\ \|\hat{s}\|_2 \le \Delta_k,\ u_k + D_k^{1/2}\hat{s} \in \mathcal{B}_N\right\}, \\ &\|\hat{s}_k\|_2 \le \beta_0 \Delta_k, \end{aligned} \tag{2.6}$$

then under appropriate, standard conditions one can show the basic trust-region convergence result

$$\liminf_{k\to\infty} \|D(u_k)^{1/2}\nabla f(u_k)\| = 0. \tag{2.7}$$

Stronger convergence results can be proven if the assumptions on the function f and on the step computation $\hat{s}_k$ are strengthened appropriately. See [6], [5], [11].

Coleman and Li [6] show that close to nondegenerate KKT points one obtains trial steps $\hat{s}_k$ which meet these requirements if one first computes an approximate solution of (2.5) ignoring the bound constraints and then satisfies the interior-point condition $a < u_k + s_k < b$ by a step-size rule. A careful analysis of the proofs in [6] reveals that the same holds true for nearly arbitrary trust-region scalings. It becomes apparent that the crucial role of the affine scaling does *not* consist of the scaling of the trust-region but rather in leading to the additional term E_k in the Hessian $\hat{M}_k$ of $\hat{\psi}_k$. Near nondegenerate KKT points this positive semidefinite diagonal matrix shapes the level sets of $\hat{\psi}_k$ in such a way that all bad directions $\hat{s}$ which allow only for small step-sizes to the boundary of the box cannot minimize $\hat{\psi}_k$ on any reasonable trust region. The trust-region scaling in (2.5) and (2.6) tends to equilibrate the distance of the origin to the bounding box constraints $\{\hat{s} : u_k + D_k^{1/2}\hat{s} \in \mathcal{B}_N\}$. However, for this feature the equivalence of 2- and ∞-norm is indispensable and thus it does not carry over to our infinite-dimensional framework. In fact, in the infinite-dimensional setting the affine-scaled trust-region $\{\|\hat{s}\|_p \le \Delta_k\}$ no longer enjoys the property of reflecting the distance to the bounding box constraints. Therefore, we will allow for a very general class of trust-region scalings in our analysis. See also [11]. Since, as mentioned above, the term E_k in the Hessian $\hat{M}_k$ plays the crucial role in this affine-scaling interior-point method, all convergence results in [6] remain valid.

In the following section we state the assumptions that allow us to generalize the affine-scaling interior-point algorithm to the infinite-dimensional problems (P).

3. Infinite-dimensional problem setting.

3.1. Notation. We set

$$U \overset{\text{def}}{=} L^p(\Omega), \quad 2 \le p \le \infty, \quad V \overset{\text{def}}{=} L^\infty(\Omega).$$

In this notation, our optimization problem (P) is written as

$$\text{(P)} \qquad \begin{aligned} &\text{minimize} \quad f(u) \\ &\text{subject to} \ \ u \in \mathcal{B} \overset{\text{def}}{=} \{u \in U \ : \ a(x) \le u(x) \le b(x), \ x \in \Omega\}. \end{aligned}$$

We use the following notations. $\mathcal{L}(X, Y)$ is the space of linear bounded operators from a Banach space X into a Banach space Y. By $\|\cdot\|_q$ we denote the norm of the Lebesgue space $L^q(\Omega)$, $1 \le q \le \infty$, and we write $(\cdot,\cdot)_2$ for the inner product of the Hilbert space $H \overset{\text{def}}{=} L^2(\Omega)$. For $(v, w) \in (L^q(\Omega), L^q(\Omega)^*)$, with $L^q(\Omega)^*$ denoting the dual space of $L^q(\Omega)$, we use the canonical dual pairing $\langle v, w\rangle \overset{\text{def}}{=} \int_\Omega v(x)w(x)\,dx$, for which, if $q < \infty$, the dual space $L^q(\Omega)^*$ is given by $L^{q'}(\Omega)$, $1/q + 1/q' = 1$. (In the case $q = 1$ this means $q' = \infty$.) Especially, if $q = 2$, we have $L^2(\Omega)^* = L^2(\Omega)$ and $\langle\cdot,\cdot\rangle$ coincides with $(\cdot,\cdot)_2$.

Finally, we set $U' \overset{\text{def}}{=} L^{p'}(\Omega)$, $1/p + 1/p' = 1$, which is the same as U^*, if $p < \infty$. Moreover, it is easily seen that $w \longmapsto \langle\cdot, w\rangle$ defines a linear norm-preserving injection from $L^1(\Omega)$ into $L^\infty(\Omega)^*$. Therefore, we may always interpret U' as subspace of U^*. Lemma 3.2 guarantees that $L^{q_2}(\Omega) \subset L^{q_1}(\Omega)$ for $1 \le q_1 \le q_2 \le \infty$. As a consequence we get the following chain of continuous imbeddings:

$$V \hookrightarrow U \hookrightarrow H = H^* \hookrightarrow U' \hookrightarrow U^* \hookrightarrow V^*.$$

Throughout this paper we will work with differentiability in the Fréchet sense. We write $g(u) \overset{\text{def}}{=} \nabla f(u) \in U^*$ for the gradient and $\nabla^2 f(u) \in \mathcal{L}(U, U^*)$ for the second derivative of f at $u \in \mathcal{B}$ if they exist. The $\|\cdot\|_\infty$-interior of $\mathcal{B}$ is denoted by $\mathcal{B}^\circ$:

$$\mathcal{B}^\circ \overset{\text{def}}{=} \bigcup_{\delta > 0} \mathcal{B}_\delta\ , \quad \mathcal{B}_\delta \overset{\text{def}}{=} \{u \in U \ : \ a(x) + \delta \le u(x) \le b(x) - \delta, \ x \in \Omega\}.$$

We often write $f_k, g_k, \ldots$ for $f(u_k), g(u_k), \ldots$.

Finally, we recall that all pointwise statements on measurable functions are meant to hold μ-almost everywhere. Similarly, we call two sets equal if they differ at most by a set of measure zero.

3.2. Basic problem setting. The finite-dimensional convergence analysis heavily relies on the equivalency of norms in $\mathbb{R}^N$. This analysis is used, for example, to obtain pointwise ($\|\cdot\|_\infty$) estimates from $\|\cdot\|_2$ estimates. In the infinite-dimensional context the formulation of the algorithm and the proof of its convergence is more delicate. The following assumptions on the function f are needed to state the algorithm and to prove its basic global convergence property, which corresponds to (2.7). Additional assumptions are needed to derive refined convergence results and will be stated later.

The basic assumptions are as follows:

(A1) $f : \mathcal{D} \longrightarrow \mathbb{R}$ is differentiable on $\mathcal{D}$ with g mapping $\mathcal{B} \subset U$ continuously into U'.

(A2) The gradient g satisfies $g(\mathcal{B}) \subset V$.

(A3) There exists $c_1 > 0$ such that $\|g(u)\|_\infty \le c_1$ for all $u \in \mathcal{B}$.

The formulation of second-order necessary optimality conditions requires the following assumption, which is also the basis for our second-order convergence results.

(A4) f is twice continuously differentiable on $\mathcal{D}$. If $p = \infty$, then $\nabla^2 f(u) \in \mathcal{L}(U, U')$ for all $u \in \mathcal{B}$, and if $(h_k) \subset V$ converges to zero in all spaces $L^q(\Omega)$, $1 \le q < \infty$, then $\nabla^2 f(u) h_k$ tends to zero in U'.

For $p \in [2, \infty)$ the assumptions (A1) and (A4) simply say that f is continuously Fréchet differentiable or that f is twice continuously Fréchet differentiable, respectively. If $p = \infty$, then the requirements that $g(u), \nabla^2 f(u) h \in U' = L^1(\Omega) \ne U^*$ for $u \in \mathcal{B}$, $h \in V$ is a further condition. It allows us to use estimates like $\langle v, g(u) \rangle \le \|g(u)\|_{p'} \|v\|_p$ for $p \in [2, \infty)$ and $p = \infty$. Moreover, since on $L^1(\Omega)$ the L^1- and $(L^\infty)^*$-norms coincide, assumption (A4) implies that $\nabla^2 f : \mathcal{B} \subset U \longrightarrow \mathcal{L}(U, U')$ is continuous also for $p = \infty$. Finally, (A1) ensures that the gradient $g(u)$ is always at least an L^1-function which will be essential for many reasons, e.g., to allow the definition of a function space analogue for the scaling matrix D.

Remark 3.1. From (A1), (A3), the boundedness of $\mathcal{B}$, and the mean value theorem it follows that f is bounded on $\mathcal{B}$.

The above conditions limit the optimal control problems that fit into this framework. However, a large and important class of optimal control problems with L^p-controls satisfies these conditions. In section 8 we will discuss the validity of these assumptions for the control problems stated in the introduction.

3.3. Norm estimates. In this section we collect several useful norm estimates for L^q-spaces. The first lemma states that $\|\cdot\|_{q_1}$ is majorizable by a multiple of $\|\cdot\|_{q_2}$ if $q_2 \ge q_1$.

LEMMA 3.2. *For all* $1 \le q_1 \le q_2 \le \infty$ *and* $v \in L^{q_2}(\Omega)$ *we have*

$$\|v\|_{q_1} \le m_{q_1,q_2} \|v\|_{q_2}$$

with $m_{q_1,q_2} = \mu(\Omega)^{\frac{1}{q_1} - \frac{1}{q_2}}$. *Here* $1/\infty$ *is to be interpreted as zero.*

Proof. See, e.g., [1, Thm. 2.8]. □

As a consequence of Hölder's inequality we obtain the following result, which allows us to apply the principle of boundedness in the high norm and convergence in the low norm.

LEMMA 3.3 (interpolation inequality). *Given* $1 \le q_1 \le q_2 \le \infty$ *and* $0 \le \theta \le 1$, *let* $1 \le q \le \infty$ *satisfy* $1/q = \theta/q_1 + (1-\theta)/q_2$. *Then for all* $v \in L^{q_2}(\Omega)$ *the following is true:*

$$\|v\|_q \le \|v\|_{q_1}^{\theta} \|v\|_{q_2}^{1-\theta}. \tag{3.1}$$

Proof. In the nontrivial cases $0 < \theta < 1$ and $q < \infty$ observe that $[q_1/(\theta q)]^{-1} + [q_2/((1-\theta)q)]^{-1} = 1$ and apply Hölder's inequality:

$$\|v\|_q^q = \big\||v|^{\theta q} |v|^{(1-\theta)q}\big\|_1 \le \big\||v|^{\theta q}\big\|_{\frac{q_1}{\theta q}} \big\||v|^{(1-\theta)q}\big\|_{\frac{q_2}{(1-\theta)q}} = \|v\|_{q_1}^{(\theta q)} \|v\|_{q_2}^{(1-\theta)q}. \qquad \square$$

The next lemma will be used in the proof of Lemma 7.1.

LEMMA 3.4. *For* $v \in L^q(\Omega)$, $1 \le q < \infty$ *and all* $\delta > 0$ *holds*

$$\mu(\{x \in \Omega \,:\, |v(x)| \ge \delta\}) \le \delta^{-q} \|v\|_q^q.$$

Proof.

$$\|v\|_q^q = \||v|^q\|_1 \ge \|\chi_{\{|v| \ge \delta\}} |v|^q\|_1 \ge \mu(\{|v| \ge \delta\}) \delta^q. \qquad \square$$

4. Necessary optimality conditions and affine scaling.

The problem under consideration belongs to the class of cone constrained optimization problems in Banach space for which optimality conditions are available [16]. However, we believe that for our particular problem an elementary derivation of the necessary optimality conditions for problem (P) not only is simpler but also is more transparent than the application of the general theory. This derivation also helps us to motivate the choice of the affine scaling which is used to reformulate the optimality condition and which is the basis for the interior-point method.

4.1. First-order necessary conditions.

The first-order necessary optimality conditions in Theorem 4.1 are completely analogous to those for finite-dimensional problems with simple bounds ([6], Sect. 2). We have only to replace coordinatewise statements with pointwise statements and to ensure that the gradient $g(\bar{u})$ is a measurable function.

THEOREM 4.1 (first-order necessary optimality conditions). *Let $\bar{u}$ be a local minimizer of problem* (P) *and assume that f is differentiable at $\bar{u}$ with $g(\bar{u}) \in U'$. Then*

$$\bar{u} \in \mathcal{B}, \tag{O1}$$

$$g(\bar{u})(x) \begin{cases} = 0 & \text{for } x \in \Omega \text{ with } a(x) < \bar{u}(x) < b(x), \\ \geq 0 & \text{for } x \in \Omega \text{ with } \bar{u}(x) = a(x), \\ \leq 0 & \text{for } x \in \Omega \text{ with } \bar{u}(x) = b(x) \end{cases} \tag{O2}$$

are satisfied.

The first-order optimality conditions for infinite-dimensional problems are typically not stated in this form. The conditions (O1), (O2) can be deduced from other first order conditions such as those in [16]. For completeness, we provide an elementary proof of Theorem 4.1.

Proof. Condition (O1) is trivially satisfied. To verify (O2), define

$$A_- = \{x \in \Omega \;:\; \bar{u}(x) = a(x),\ g(\bar{u})(x) < 0\}\,, \quad A_-^k = \{x \in A_- \;:\; g(\bar{u})(x) \leq -1/k\}\,,$$

and assume that A_- has positive measure $\mu(A_-) > \varepsilon > 0$. Since μ is continuous from below and $A_-^k \uparrow A_-$, there exists $l > 0$ with $\mu(A_-^l) \geq \varepsilon$. This yields a contradiction because $\bar{u} + \tau s \in \mathcal{B}$, $s = \chi_{A_-}(b - a)$, for $0 \leq \tau \leq 1$, and

$$\frac{d}{d\tau} f(\bar{u} + \tau s)|_{\tau=0} = \langle s, g(\bar{u}) \rangle \leq -\frac{\varepsilon \nu}{l} < 0.$$

Hence, we must have $\mu(A_-) = 0$. In the same way we can show that $\mu(A_+) = 0$ for $A_+ = \{x \in \Omega \;:\; \bar{u}(x) = b(x),\ g(\bar{u})(x) > 0\}$. Finally, we look at

$$I = \{x \in \Omega \;:\; a(x) < \bar{u}(x) < b(x),\ g(\bar{u})(x) \neq 0\}\,.$$

Assume that $\mu(I) > \varepsilon > 0$. Since $I^k \uparrow I$ with

$$I^k = \{x \in \Omega \;:\; a(x) + 1/k \leq \bar{u}(x) \leq b(x) - 1/k,\ |g(\bar{u})(x)| \geq 1/k\}\,,$$

we can find $l > 0$ with $\mu(I^l) \geq \varepsilon$ and obtain for

$$s = -\chi_{I^l} \frac{g(\bar{u})}{|g(\bar{u})|}$$

that $\bar{u} + \tau s \in \mathcal{B}$, $0 \le \tau \le 1/l$, and

$$\frac{d}{d\tau} f(\bar{u} + \tau s)|_{\tau=0} = \langle s, g(\bar{u}) \rangle \le -\frac{\varepsilon}{l} < 0,$$

a contradiction to the local optimality of $\bar{u}$. Hence $\mu(I) = \mu(A_-) = \mu(A_+) = 0$, which means that (O2) holds. □

4.2. Affine scaling. Let assumption (A1) hold. Our algorithm will be based on the following equivalent affine-scaling formulation of (O2):

$$d^r(\bar{u}) g(\bar{u}) = 0, \tag{4.1}$$

where $r > 0$ is arbitrary and $d(u) \in V$, $u \in \mathcal{B}$, is a scaling function which is assumed to satisfy

$$d(u)(x) \begin{cases} = 0 & \text{if } u(x) = a(x) \text{ and } g(u)(x) \ge 0, \\ = 0 & \text{if } u(x) = b(x) \text{ and } g(u)(x) \le 0, \\ > 0 & \text{else}, \end{cases} \tag{4.2}$$

for all $x \in \Omega$. The equivalence of (O2) and (4.1) will be stated and proved in Lemma 4.2. Before we do this, we give two examples of proper choices for d. The first choice $d = d_{\mathrm{I}}$ is motivated by the scaling matrices used in [6] (see (2.1)). Except for points x with $g(u)(x) = 0$ it equals those used in [6] and [11]:

$$d_{\mathrm{I}}(u)(x) \overset{\text{def}}{=} \begin{cases} u(x) - a(x) & \text{if } g(u)(x) > 0 \text{ or} \\ & g(u)(x) = 0 \text{ and } u(x) - a(x) \le b(x) - u(x), \\ b(x) - u(x) & \text{if } g(u)(x) < 0 \text{ or} \\ & g(u)(x) = 0 \text{ and } b(x) - u(x) < u(x) - a(x). \end{cases} \tag{4.3}$$

This slight modification in comparison to (2.1) lets d_{I} satisfy (4.2). This will enable us to prove second-order global convergence without a nondegeneracy assumption, which is needed by Coleman and Li [6].

While the global analysis could be carried out entirely with this choice, the discontinuous response of $d(u)(x)$ to sign changes of $g(u)(x)$ raises difficulties for the design of superlinearly convergent algorithms in infinite dimensions. These can be circumvented by the choice $d = d_{\mathrm{II}}$, where

$$d_{\mathrm{II}}(u)(x) \overset{\text{def}}{=} \begin{cases} \min\{|g(u)(x)|, c(x)\} & \text{if } -g(u)(x) > u(x) - a(x) \\ & \text{and } u(x) - a(x) \le b(x) - u(x), \\ \min\{|g(u)(x)|, c(x)\} & \text{if } g(u)(x) > b(x) - u(x) \\ & \text{and } b(x) - u(x) \le u(x) - a(x), \\ \min\{u(x) - a(x), & \\ \qquad b(x) - u(x), c(x)\} & \text{else}. \end{cases} \tag{4.4}$$

Here $c : x \in \Omega \longmapsto \min\{\zeta(b(x) - a(x)), \kappa\}$ with $\zeta \in (0, 1/2]$ and $\kappa \ge 1$.

It is easily seen that $d = d_{\mathrm{I}}$ and $d = d_{\mathrm{II}}$ both satisfy (4.2).

LEMMA 4.2. *Let* (A1) *hold and* $\bar{u} \in \mathcal{B}$. *Then* (O2) *is equivalent to* (4.1) *for all* $r > 0$ *and all* d *satisfying* (4.2).

Proof. Since d^r, $r > 0$, also satisfies (4.2), we may restrict ourselves to the case $r = 1$. First assume that (O2) holds. For all $x \in \Omega$ with $g(\bar{u})(x) = 0$ we also have $d(\bar{u})(x) g(\bar{u})(x) = 0$. If $g(\bar{u})(x) > 0$, then by (O2) $\bar{u}(x) = a(x)$ and if $g(\bar{u})(x) < 0$ then $\bar{u}(x) = b(x)$. In both cases $d(\bar{u})(x) = 0$ and hence $d(\bar{u})(x) g(\bar{u})(x) = 0$. On the other hand, let $d(\bar{u})g(\bar{u}) = 0$ hold. For all $x \in \Omega$ with $a(x) < \bar{u}(x) < b(x)$ we have $d(\bar{u})(x) > 0$ which implies $g(\bar{u})(x) = 0$. For all $x \in \Omega$ with $\bar{u}(x) = a(x)$ we obtain $g(\bar{u})(x) \geq 0$ since $g(\bar{u})(x) < 0$ would yield the contradiction $d(\bar{u})(x) > 0$. Analogously, we see that $g(\bar{u})(x) \leq 0$ for all $x \in \Omega$ with $\bar{u}(x) = b(x)$. Therefore, (O2) holds. □

4.3. Second-order conditions. If assumption (A4) holds, we can derive second-order conditions which are satisfied at all local solutions of (P). These are also analogous to the well-known conditions for finite-dimensional problems.

THEOREM 4.3 (second-order necessary optimality conditions). *Let* (A4) *be satisfied and let $g(\bar{u}) \in U'$ hold at the local minimizer $\bar{u}$ of problem* (P). *Then* (O1), (O2), *and*

$$\text{(O3)} \qquad \langle s, \nabla^2 f(\bar{u})s \rangle \geq 0 \textit{ for all } s \in T(\mathcal{B}, \bar{u})$$

are satisfied, where

$$T(\mathcal{B}, \bar{u}) \overset{\text{def}}{=} \{s \in V \;:\; s(x) = 0 \textit{ for all } x \in \Omega \textit{ with } \bar{u}(x) \in \{a(x), b(x)\}\}$$

denotes the tangent space of the active constraints.

As in the case of Theorem 4.1, these second-order necessary optimality conditions are related to the general conditions in, e.g., [16]. However, they are tailored to our problem and are slightly different. For completeness, we provide an elementary proof.

Proof. Let the assumptions hold. As shown in Theorem 4.1, (O1) and (O2) are satisfied. In particular, we have that $sg(\bar{u}) = 0$ for all $s \in T(\mathcal{B}, \bar{u})$. Now assume the existence of $s \in T(\mathcal{B}, \bar{u})$ and $\varepsilon > 0$ with $\langle s, \nabla^2 f(\bar{u})s \rangle < -\varepsilon$. Let $I = \{x \in \Omega \;:\; a(x) < \bar{u}(x) < b(x)\}$,

$$I_k = \{x \in \Omega \;:\; a(x) + 1/k \leq \bar{u}(x) \leq b(x) - 1/k\}, \tag{4.5}$$

and define restrictions $s^k = \chi_{I_k} s \in V$. Since $I_k \uparrow I$ and $s = \chi_I s$, we get $\|s^k - s\|_q^q \leq \mu(I \backslash I_k)\|s\|_\infty^q$. Hence, the restrictions s^k converge to s in all spaces $L^q(\Omega)$, $1 \leq q < \infty$. Therefore, $\nabla^2 f(\bar{u})(s - s^k)$ tends to zero in U' by (A4) and, using the symmetry of $\nabla^2 f(\bar{u})$,

$$\begin{aligned}
\langle s^k, \nabla^2 f(\bar{u})s^k \rangle &= \langle s, \nabla^2 f(\bar{u})s \rangle - \langle s + s^k, \nabla^2 f(\bar{u})(s - s^k) \rangle \\
&< 2\,\|\nabla^2 f(\bar{u})(s - s^k)\|_{p'} \|s\|_p - \varepsilon \\
&\leq -\varepsilon/2
\end{aligned}$$

for all sufficiently large k. Let $l > 0$ be such that $\langle s^l, \nabla^2 f(\bar{u})s^l \rangle \leq -\varepsilon/2$. The observations that $s^l \in T(\mathcal{B}, \bar{u})$ and $\bar{u} + \tau s^l \in \mathcal{B}$ for $0 \leq \tau \leq 1/(l\|s\|_\infty)$ now yield the desired contradiction:

$$\frac{d}{d\tau} f(\bar{u} + \tau s^l)|_{\tau=0} = \langle s, g(\bar{u}) \rangle = 0,$$

$$\frac{d^2}{d\tau^2} f(\bar{u} + \tau s^l)|_{\tau=0} = \langle s^l, \nabla^2 f(\bar{u})s^l \rangle \leq -\varepsilon/2 < 0.$$

This readily shows that (O3) holds. □

5. The algorithm.

5.1. A Newton-like iteration. The key idea of the method to be developed consists in solving the equation $d(u)g(u) = 0$ by means of a Newton-like method augmented by a trust-region globalization. The bound constraints on u are enforced by, e.g., a scaling of the Newton-like step. In particular, all iterates will be strictly feasible with respect to the bounds $u_k \in \mathcal{B}^\circ$.

In general it is not possible to find a function d satisfying (4.2) that depends smoothly on u. For an efficient method, however, we need a suitable substitute for the derivative of dg. Formal application of the product rule suggests choosing an approximate derivative of the form

$$D(u)\nabla^2 f(u) + D_u(u)g(u), \ u \in \mathcal{B}^\circ,$$

with $D_u(u)w \in \mathcal{L}(U, U')$, $w \in U'$, replacing the generally nonexistent derivative of $u \in \mathcal{B} \longmapsto d(u)w \in U'$ at u. Here and in the sequel the linear operator $D^r(u)$, $r \geq 0$, denotes the pointwise multiplication operator associated with $d^r(u)$, i.e.,

$$D^r(u) : v \longmapsto d(u)^r v.$$

Since $d^r(u) \in V$, $D^r(u)$ maps $L^q(\Omega)$, $1 \leq q \leq \infty$, continuously into itself. Moreover, if the assumption (D2) below is satisfied and $u \in \mathcal{B}^\circ$, then $D^r(u)$ defines an automorphism of $L^q(\Omega)$, $1 \leq q \leq \infty$, with inverse $D^{-r}(u)$. In fact, for all $u \in \mathcal{B}^\circ$ there exists $0 < \delta \leq \delta_d$ such that $u \in \mathcal{B}_\delta$, and thus $d(u)(x) \geq \varepsilon_d(\delta)$ on Ω by (D2). If we look at the special case $d = d_{\mathrm{I}}$, the choice $D_u(u)w = d'_{\mathrm{I}}(u)w$ with

$$d'_{\mathrm{I}}(u)(x) \overset{\mathrm{def}}{=} \begin{cases} 1 & \text{if } g(u)(x) > 0 \text{ or} \\ & g(u)(x) = 0 \text{ and } u(x) - a(x) \leq b(x) - u(x) \\ -1 & \text{if } g(u)(x) < 0 \text{ or} \\ & g(u)(x) = 0 \text{ and } b(x) - u(x) < u(x) - a(x) \end{cases} \tag{5.1}$$

for $u \in \mathcal{B}, \ x \in \Omega$ seems to be the most natural. A suitable choice for d'_{II} is

$$d'_{\mathrm{II}}(u)(x) = \chi_{\{d'_{\mathrm{II}}(u)<c\}}(x) d'_{\mathrm{I}}(u)(x). \tag{5.2}$$

For the general case this suggests the choice

$$D(u)\nabla^2 f(u) + E(u),$$

where

$$E(u) : v \longmapsto e(u)v$$

is a multiplication operator characterized by $e(u) \in V$, which approximates $D_u(u)g(u)$. Properties of E will be specified below.

We are now able to formulate the following Newton-like iteration for the solution of $d(u)g(u) = 0$.

Given $u_k \in \mathcal{B}^\circ$, compute the new iterate $u_{k+1} := u_k + s_k \in \mathcal{B}^\circ$, where $s_k \in U$ solves

$$(D_k B_k + E_k)s_k = -d_k g_k \tag{5.3}$$

and B_k denotes a symmetric approximation of (or replacement for) $\nabla^2 f(u_k)$, i.e., $\langle v, B_k w\rangle = \langle w, B_k v\rangle$ for all $v, w \in U$.

We assume that B_k satisfies the following condition:

(A5) The norms $\|B_k\|_{U,U'}$ are uniformly bounded by a constant $c_2 > 0$.

In the following, we will not restrict our investigations to special choices of d and e. Rather, we will develop an algorithm that is globally convergent for all affine scalings d and corresponding e satisfying the assumptions (D1)–(D5):

(D1) The scaling d satisfies (4.2) for all $u \in \mathcal{B}$.

(D2) There exists $\delta_d > 0$ such that for all $\delta \in (0, \delta_d]$ there is $\varepsilon_d = \varepsilon_d(\delta) > 0$ such that $d(u)(x) \geq \varepsilon_d$ for all $u \in \mathcal{B}$ and all $x \in \Omega$ with $a(x) + \delta \leq u(x) \leq b(x) - \delta$.

(D3) The scaling satisfies $d(u)(x) \leq d_{\mathrm{I}}(u)(x)$ for all $u \in \mathcal{B}$, $x \in \Omega$ and d_{I} given by (4.3). In particular, $d(u)(x) \leq c_d$ for some $c_d > 0$.

(D4) For all $u \in \mathcal{B}$ the function $e(u)$ satisfies $0 \leq e(u)(x) \leq c_e$ for all $x \in \Omega$ and $g(u)(x) = 0$ implies $e(u)(x) = 0$.

(D5) The function $e(u)$ is given by $e(u) = d'(u)g(u)$, where $d'(u)$ satisfies $|d'(u)(x)| \leq c_{d'}$ for all $u \in \mathcal{B}$ and $x \in \Omega$.

We have seen that assumption (D1) is essential for the reformulation of the first-order necessary optimality conditions and that (D2) ensures the continuous invertibility of the scaling operator $D(u)$ for $u \in \mathcal{B}^\circ$. Furthermore, assumption (D2) will be used in the second-order convergence analysis. The assumption (D4), together with (A5), is needed to ensure uniform boundedness of the Hessian approximations $\hat{M}_k$ to be defined in the next section. The assumption (D5) is needed to prove second-order convergence results.

Obviously, (D1)–(D3) hold for either $d = d_{\mathrm{I}}$ and $d = d_{\mathrm{II}}$. The assumption (D4) is satisfied for $e(u) = d'_{\mathrm{I}}(u)g(u)$, where $d'_{\mathrm{I}}(u)$ is given by (5.1), provided that $\|g(u)\|_\infty$ is uniformly bounded on $\mathcal{B}$, i.e., provided that (A3) holds.

5.2. New coordinates and symmetrization. Since neither the global well-definedness nor the global convergence of the Newton-like iteration (5.3) can be ensured, we intend to safeguard and globalize it by means of a closely related trust-region method. To this end we have to transform (5.3) into an equivalent quadratic programming problem. While the iterates are required to stay strictly feasible with respect to the bound constraints, we want to use an affine-scaling interior-point approach to reduce the effect of the interfering bound constraints in the quadratic subproblem as far as possible. The affine scaling can be expressed by a change of coordinates $s \rightsquigarrow \hat{s}$ and has to be performed in such a way that we get enough distance from the boundary of the box $\mathcal{B}$ to be able to impose a useful fraction of Cauchy decrease condition on the trial step. An appropriate change of coordinates $s \rightsquigarrow \hat{s}$ is given by $\hat{s} \overset{\text{def}}{=} d_k^{-r} s$. Here

$$r \geq \frac{1}{2}$$

is arbitrary but fixed throughout the iteration. Performing this transformation and applying D_k^{r-1}, the multiplication operator associated with d_k^{r-1}, from the left to (5.3) leads to the equivalent equation

$$\hat{M}_k \hat{s}_k = -\hat{g}_k, \tag{5.4}$$

with $\hat{g}(u) \overset{\text{def}}{=} d^r(u)g(u)$, $\hat{M}_k \overset{\text{def}}{=} \hat{B}_k + \hat{C}_k$, where $\hat{B}_k \overset{\text{def}}{=} D_k^r B_k D_k^r$ and $\hat{C}_k \overset{\text{def}}{=} E_k D_k^{2r-1}$.

Remark 5.1. Assumptions (D3), (D4), and (A5) imply that $\|\hat{M}_k\|_{U,U'}$ are uniformly bounded by a constant $c_3 > 0$.

Since $\hat{M}_k$ is symmetric, $\hat{s}_k$ is a solution of (5.4) if and only if it is a stationary point of the quadratic function

$$\hat{\psi}_k(\hat{s}) \overset{\text{def}}{=} \langle \hat{s}, \hat{g}_k \rangle + \frac{1}{2} \langle \hat{s}, \hat{M}_k \hat{s} \rangle.$$

We will return to this issue later.

5.3. Second-order necessary conditions revisited. If $B_k = \nabla^2 f(u_k)$, then the operator

$$\hat{M}(u) \overset{\text{def}}{=} D(u)^r \nabla^2 f(u) D(u)^r + E(u) D(u)^{2r-1} \tag{5.5}$$

also plays an important role in the second-order necessary optimality conditions. In fact, we will show that if conditions (O1), (O2) hold at $\bar{u}$, then (O3) can be equivalently replaced by

$$\langle s, \hat{M}(\bar{u}) s\rangle \geq 0 \text{ for all } s \in T(\mathcal{B}, \bar{u}) \tag{O3$'$}$$

or even

$$\langle s, \hat{M}(\bar{u}) s\rangle \geq 0 \text{ for all } s \in V. \tag{O3$''$}$$

The proof requires the following two lemmas.

Lemma 5.2. *Let f be differentiable at $\bar{u}$ with $g(\bar{u}) \in U'$,* (D1) *be satisfied, and suppose that* (O1), (O2) *hold at $\bar{u}$. Then*

$$I^* \overset{\text{def}}{=} \{x \in \Omega \,:\, d(\bar{u})(x) > 0\} = \{x \in \Omega \,:\, a(x) < \bar{u}(x) < b(x)\} \overset{\text{def}}{=} I. \tag{5.6}$$

Proof. The inclusion $I \subset I^*$ is obvious from (4.2) and (O2). Now let $x \in I^*$ be given. Then $g(\bar{u})(x) = 0$ by (O2) and Lemma 4.2. From (4.2) we conclude $\bar{u}(x) \notin \{a(x), b(x)\}$, i.e., $x \in I$. □

Lemma 5.3. *Let f be twice continuously differentiable at $\bar{u}$ with $g(\bar{u}) \in U'$ and $\nabla^2 f(\bar{u}) \in \mathcal{L}(U, U')$. Assume that* (D1) *and* (D4) *are satisfied and* (O1), (O2) *hold at $\bar{u}$. Then the statements* (O3$'$) *and* (O3$''$) *are equivalent.*

Proof. Obviously (O3$''$) implies (O3$'$). To show the opposite direction, assume that (O3$'$) holds. Set $A = \Omega \backslash I$, where I is the set defined in (5.6). For arbitrary $s \in V$ we perform the splitting $s = s_I + s_A$, $s_I = \chi_I s \in T(\mathcal{B}, \bar{u})$, $s_A = \chi_A s$. Lemma 5.2 implies that $d^r(\bar{u}) s_A = 0$. Moreover, since s_I and s_A have disjoint support, integrals $\langle \cdot, \cdot \rangle$ involving both s_I and s_A are zero. With these observations we obtain

$$\begin{aligned}
\langle s, \hat{M}(\bar{u}) s\rangle &= \langle s_I, \hat{M}(\bar{u}) s_I\rangle + 2\langle s_A, e(\bar{u}) d^{2r-1}(\bar{u}) s_I\rangle \\
&\quad + 2\langle d^r(\bar{u}) s_A, \nabla^2 f(\bar{u}) d^r(\bar{u}) s_I\rangle + \langle d^r(\bar{u}) s_A, \nabla^2 f(\bar{u}) d^r(\bar{u}) s_A\rangle \\
&\quad + \langle s_A, e(\bar{u}) d^{2r-1}(\bar{u}) s_A\rangle \\
&= \langle s_I, \hat{M}(\bar{u}) s_I\rangle + \langle s_A, e(\bar{u}) d^{2r-1}(\bar{u}) s_A\rangle \geq \langle s_I, \hat{M}(\bar{u}) s_I\rangle \geq 0.
\end{aligned} \tag{5.7}$$

This completes the proof. □

Theorem 5.4. *Let* (D1), (D2), *and* (D4) *be satisfied. Then in Theorem* 4.3 *condition* (O3) *can be equivalently replaced by* (O3$'$) *or* (O3$''$).

Proof. Since the conditions of Theorem 4.3 and Lemma 5.3 guarantee that (O3$'$) and (O3$''$) are equivalent, we only need to show that (O3) can be replaced by (O3$'$).

Let (O1), (O2) be satisfied and let $s \in T(\mathcal{B}, \bar{u})$ be arbitrary. Condition (O2) implies $sg(\bar{u}) = 0$. Hence $\{s \neq 0\} \subset \{g(\bar{u}) = 0\}$ and, by (D4), $\{s \neq 0\} \subset \{e(\bar{u}) = 0\}$. This shows that

$$E(\bar{u}) s = 0 \quad \text{for all } s \in T(\mathcal{B}, \bar{u}). \tag{5.8}$$

To show that (O3) implies (O3′), let $s \in T(\mathcal{B}, \bar{u})$ be arbitrary. Since $d(\bar{u}) \geq 0$, $h = d^r(\bar{u})s \in T(\mathcal{B}, \bar{u})$. With (5.8) this yields

$$\begin{aligned}\langle s, \hat{M}(\bar{u})s\rangle &= \langle h, \nabla^2 f(\bar{u})h\rangle + \langle s, E(\bar{u})D^{2r-1}(\bar{u})s\rangle \\ &= \langle h, \nabla^2 f(\bar{u})h\rangle + \langle E(\bar{u})s, D^{2r-1}(\bar{u})s\rangle = \langle h, \nabla^2 f(\bar{u})h\rangle \geq 0.\end{aligned}$$

To prove the opposite direction, assume that there exist $s \in T(\mathcal{B}, \bar{u})$ and $\varepsilon > 0$ with $\langle s, \nabla^2 f(\bar{u})s\rangle < -\varepsilon$. As carried out in the proof of Theorem 4.3, we can find $l > 0$ such that $s^l = \chi_{I_l} s \in T(\mathcal{B}, \bar{u})$, I_l, as defined in (4.5), satisfies $\langle s^l, \nabla^2 f(\bar{u})s^l\rangle \leq -\varepsilon/2$. Since $d(\bar{u})$ is bounded away from zero on I_l by assumption (D2), we obtain that $h = \chi_{I_l} d^{-r}(\bar{u})s$ is an element of $T(\mathcal{B}, \bar{u})$. With (5.8) applied to h this yields

$$\begin{aligned}\langle h, \hat{M}(\bar{u})h\rangle &= \langle h, D(\bar{u})^r \nabla^2 f(\bar{u}) D(\bar{u})^r h\rangle + \langle h, E(\bar{u})D(\bar{u})^{2r-1}h\rangle \\ &= \langle s^l, \nabla^2 f(\bar{u})s^l\rangle \leq -\varepsilon/2,\end{aligned}$$

a contradiction to (O3′). □

The previous results show that $\hat{\psi}_k(\hat{s})$ is convex and admits a global minimum at $\hat{s} = 0$ if $B_k = \nabla^2 f(u_k)$ and $u_k = \bar{u}$ is a local solution of (P).

5.4. Trust-region globalization. The results on the second-order conditions in the previous section indicate that the Newton-like iteration (5.4) can be used locally under appropriate conditions on B_k. To globalize the iteration, we minimize $\hat{\psi}_k(\hat{s})$ over the intersection of the ball $\|\hat{w}_k\hat{s}\|_p \leq \Delta_k$ and the box $\mathcal{B}$ which leads to the following trust-region subproblem: Compute an approximate solution $\hat{s}_k$ with $u_k + d_k^r\hat{s}_k \in \mathcal{B}^\circ$ of

$$\min \hat{\psi}_k(\hat{s}) \quad \text{subject to} \quad \|\hat{w}_k\hat{s}\|_p \leq \Delta_k\,,\ u_k + d_k^r\hat{s} \in \mathcal{B}. \tag{5.9}$$

Here $\hat{w}_k \in V$ is a positive scaling function for the trust-region. We make the following assumption on $w_k = d_k^{-r}\hat{w}_k$:

(W) There exist $c_w > 0$ and $c_{w'} > 0$ such that $\|d_k^r w_k\|_\infty \leq c_w$ and $\|w_k^{-1}\|_\infty \leq c_{w'}$ for all k.

Examples for w_k are $w_k = d_k^{-r}$, which yields a ball in the $\hat{s}$-variables, and $w_k = 1$, which leads to a ball in the s-variables. Both choices satisfy (W) if (D3) holds. See also [11].

As noted in section 2, the crucial contributions of the affine scaling are the term $E(u)D(u)^{2r-1}$ in the Hessian $\hat{M}(u)$ and the scaling $\hat{g}$ of the gradient. The trust region serves as a tool for globalization. Therefore, general trust-region scalings can be admitted as long as they satisfy (W).

We will work with the original variables in terms of which the above problem reads as follows: Compute s_k with $u_k + s_k \in \mathcal{B}^\circ$ as an approximate solution of

$$\min \psi_k(s) \quad \text{subject to} \quad \|w_k s\|_p \leq \Delta_k\,,\ u_k + s \in \mathcal{B}, \tag{5.10}$$

with $\psi_k(s) = \langle s, g_k\rangle + \frac{1}{2}\langle s, M_k s\rangle$, $M_k = B_k + C_k$, $C_k = E_k D_k^{-1}$, and $w_k = d_k^{-r}\hat{w}_k$. The functions d_k^{-r} and d_k^{-1} are well defined only if $u_k \in \mathcal{B}^\circ$. Therefore, the condition $u_k + d_k^r\hat{s}_k \in \mathcal{B}^\circ$ on the trial iterate is essential. However, it is important to remark that the bound constraints do not need to be strictly enforced when computing $\hat{s}_k$. For example, in the finite-dimensional algorithms in [6], [11], an approximate solution of

$$\min \hat{\psi}_k(\hat{s}) \quad \text{subject to} \quad \|\hat{w}_k\hat{s}\|_p \leq \Delta_k$$

is computed and then scaled by $\tau_k > 0$ so that $u_k + \tau_k d_k^r \hat{s}_k \in \mathcal{B}^\circ$. Similar techniques also apply in the infinite-dimensional framework. Practical choices for the infinite-dimensional algorithm will be discussed in [23].

5.5. Cauchy decrease for the trial steps. An algorithm which is based on the iterative approximate solution of subproblem (5.10) can be expected to converge to a local solution of (P) only if the trial steps s_k produce a sufficiently large decrease of ψ_k. A well-established way to impose such a condition is the requirement that the decrease provided by s_k should be at least a fraction of the Cauchy decrease. Here the Cauchy decrease denotes the maximum possible decrease along the steepest descent direction of ψ_k at $s = 0$ with respect to an appropriate norm (or, equivalently, appropriate coordinates) inside the feasible region of the subproblem. We will see in Lemma 6.1 that the new coordinates $\hat{s} = d_k^{-r} s$ indeed provide enough distance to the boundary of $\mathcal{B}$ to allow the implementation of a useful Cauchy decrease strategy.

Unless as in the Hilbert space case, $p = 2$, the steepest descent direction of $\hat{\psi}_k$ at $\hat{s} = 0$ is *not* given by the negative gradient $-\hat{g}_k$ but rather by any $\hat{s}^d \neq 0$ satisfying $\langle \hat{s}^d, \hat{g}_k \rangle = \|\hat{s}^d\|_p \|\hat{g}_k\|_{p'}$. On the other hand, if $\hat{g}_k \in H$, then $-\nabla\hat{\psi}_k(0) = -\hat{g}_k$ is the $\|\cdot\|_2$-steepest descent direction of $\hat{\psi}_k$ at $\hat{s} = 0$. This is a strong argument for choosing this direction as basis for the Cauchy decrease condition. Of course, this approach is useful only if we ensure that $u_k - \tau d_k^r \hat{g}_k \in \mathcal{B}^\circ$ for all $\tau > 0$ sufficiently small which can be done by imposing condition (A2) on g, which is not very restrictive. Assuming this, we may take $-d_k^r \hat{g}_k = -d_k^{2r} g_k$ as the Cauchy decrease direction of ψ_k and therefore define the following fraction of Cauchy decrease condition: There exist $\beta, \beta_0 > 0$ (fixed for all k) such that s_k is an approximate solution of (5.10) in the sense

$$\|w_k s_k\|_p \le \beta_0 \Delta_k\,,\ u_k + s_k \in \mathcal{B}^\circ\,,\ \text{and } \psi_k(s_k) < \beta \psi_k(s_k^c), \tag{5.11a}$$

where s_k^c is a solution of the one-dimensional problem

$$\min \psi_k(s) \quad \text{subject to} \quad s = -t d_k^{2r} g_k\,,\ t \ge 0\,,\ u_k + s \in \mathcal{B}\,,\ \|w_k s\|_p \le \Delta_k. \tag{5.11b}$$

It is worth mentioning that, as in the finite-dimensional case, an approximate solution $\hat{s}_k$ of (5.10) satisfying (5.11a) can be easily obtained by applying any descent method which starts minimization at $s = 0$ along the Cauchy decrease direction $-d_k^{2r} g_k$.

5.6. Formulation of the algorithm. For the update of the trust-region radius Δ_k and the acceptance of the step we use a very common strategy. It is based on the demand that the actual decrease

$$ared_k(s_k) \overset{\text{def}}{=} f_k - f(u_k + s_k) \tag{5.12}$$

should be a sufficiently large fraction of the predicted decrease

$$pred_k(s_k) \overset{\text{def}}{=} -\langle s_k, g_k \rangle - \frac{1}{2}\langle s_k, B_k s_k \rangle = -\psi_k(s_k) + \frac{1}{2}\langle s_k, C_k s_k \rangle \tag{5.13}$$

promised by the quadratic model. Since the model error is at most $O(\|s_k\|_p^2)$, the decrease ratio

$$\rho_k \overset{\text{def}}{=} \frac{ared_k(s_k)}{pred_k(s_k)} \tag{5.14}$$

will tend to 1 for $s_k \to 0$. This suggests the following strategy for the update of the trust-region radius.

Algorithm 5.5 (update of the trust-region radius Δ_k).

Let $0 < \eta_1 < \eta_2 < \eta_3 < 1$, and $0 < \gamma_1 < 1 < \gamma_2 < \gamma_3$.

1. If $\rho_k \le \eta_1$, then choose $\Delta_{k+1} \in (0, \gamma_1\Delta_k]$.
2. If $\rho_k \in (\eta_1, \eta_2)$, then choose $\Delta_{k+1} \in [\gamma_1\Delta_k, \Delta_k]$.
3. If $\rho_k \in [\eta_2, \eta_3)$, then choose $\Delta_{k+1} \in [\Delta_k, \gamma_2\Delta_k]$.
4. If $\rho_k \ge \eta_3$, then choose $\Delta_{k+1} \in [\gamma_2\Delta_k, \gamma_3\Delta_k]$.

Remark 5.6. The forms of predicted and actual decrease follow the choices used in [11], [24] (and [10] for the constrained case). In [6] the decreases and the ratio are computed as follows:

$$\begin{aligned} pred_k^1(s_k) &\overset{\text{def}}{=} -\psi_k(s_k), \\ ared_k^1(s_k) &\overset{\text{def}}{=} ared_k(s_k) - \tfrac{1}{2}\langle s_k, C_k s_k\rangle, \\ \rho_k^1 &\overset{\text{def}}{=} \frac{ared_k^1(s_k)}{pred_k^1(s_k)}. \end{aligned} \tag{5.15}$$

We restrict the presentation to the choice (5.12), (5.13). However, we already note that all convergence results presented in this paper remain valid if ρ_k is replaced by ρ_k^1 . This will be discussed in more detail in Remarks 6.7 and 7.4 following the convergence results.

The algorithm iteratively computes a trial step s_k satisfying the fraction of Cauchy decrease condition. Depending on the decrease ratio ρ_k the trial step is accepted or rejected, and the trust-region radius is adjusted.

Algorithm 5.7 (trust-region interior-point algorithm).

```
Let $\eta_1 > 0$ as in Algorithm 5.5.
 1. Choose $u_0 \in \mathcal{B}^\circ$ and $\Delta_0 > 0$.
 2. For $k = 0, 1, \dots$
    2.1. Compute the gradient $g_k = g(u_k)$.
    2.2. Compute the scaling $d_k = d(u_k)$.
    2.3. Compute the trust-region scaling $w_k = w(u_k)$.
    2.4. Compute the Hessian approximation $B_k$.
    2.5. Compute $s_k$ satisfying (5.11).
    2.6. If $\|s_k\| = 0$ then stop with result $u_k$.
    2.7. Compute $\rho_k$ as defined in (5.14).
    2.8. If $\rho_k > \eta_1$ then set $u_{k+1} = u_k + s_k$, else set $u_{k+1} = u_k$.
    2.9. Compute $\Delta_{k+1}$ using Algorithm 5.5.
```

In the first-order version of Algorithm 5.7 one can stop after step 2.2 if $\|\hat{g}_k\| = 0$. The stopping criteria as applied above will be important in the second-order version of the algorithm, which will be introduced in section 7.

6. Convergence to first-order optimal points. The convergence of the algorithm is mainly achieved by two ingredients: a lower bound for the predicted decrease for trial steps satisfying the fraction of Cauchy decrease condition, and the relation $ared_k(s_k) > \eta_1 pred_k(s_k)$, which is always satisfied for successful steps s_k. The lower bound on the predicted decrease is established in the following lemma.

Lemma 6.1. *Let the assumptions* (A1), (A2), (D1)–(D4), *and* (W) *hold. Then there exists $c_4 > 0$ such that for all $u_k \in \mathcal{B}^\circ$ with $\hat{g}_k \ne 0$ and all s_k satisfying* (5.11) *the following holds:*

$$pred_k(s_k) \ge -\psi_k(s_k) \ge \frac{1}{2}\beta\|\hat{g}_k\|_2^2 \min\left\{ \frac{\Delta_k}{c_w\|\hat{g}_k\|_p}, \frac{\|\hat{g}_k\|_2^2}{\|\hat{M}_k\|_{U,U'}\|\hat{g}_k\|_p^2}, \frac{c_d^{1-2r}}{\|g_k\|_\infty} \right\}, \tag{6.1}$$

$$\geq c_4 \|\hat{g}_k\|_{p'}^2 \min\left\{ \frac{\Delta_k}{c_w \|\hat{g}_k\|_p}, \frac{\|\hat{g}_k\|_{p'}^2}{\|\hat{M}_k\|_{U,U'} \|\hat{g}_k\|_p^2}, \frac{c_d^{1-2r}}{\|g_k\|_\infty} \right\}. \tag{6.2}$$

If in addition the assumptions (A3) *and* (A5) *hold, then there exists* $c_4' > 0$ *such that for all* $u_k \in \mathcal{B}^\circ$ *with* $\hat{g}_k \neq 0$ *and all* s_k *satisfying* (5.11) *the following holds:*

$$pred_k(s_k) \geq -\psi_k(s_k) \geq c_4' \|\hat{g}_k\|_{p'}^2 \min\left\{ \Delta_k, \|\hat{g}_k\|_{p'}^2, c_d^{1-2r} \right\}. \tag{6.3}$$

Proof. Since C_k is obviously positive by (D4), we have

$$pred_k(s_k) = -\psi_k(s_k) + \frac{1}{2}\langle s_k, C_k s_k \rangle \geq -\psi_k(s_k).$$

From (5.11) we obtain that

$$\psi_k(s_k) \leq \beta \min_{[0,\tau^+]} \phi(\tau), \tag{6.4}$$

where $\phi(\tau) = \psi_k(-\tau d_k^{2r} g_k)$ and $\tau^+ = \min\{\tau_{\mathcal{B}}, \tau_\Delta\}$, with

$$\tau_{\mathcal{B}} = \max\big\{\tau \;:\; b(x) - u_k(x) + \tau d_k^{2r}(x) g_k(x) \geq 0 \text{ and } u_k(x) - a(x) - \tau d_k^{2r}(x) g_k(x) \geq 0 \text{ for all } x \in \Omega\big\}$$

and

$$\tau_\Delta = \frac{\Delta_k}{\|w_k d_k^{2r} g_k\|_p}.$$

We will derive an upper bound for $\min_{[0,\tau^+]} \phi(\tau)$.

First, we bound τ_Δ and $\tau_{\mathcal{B}}$ from below. Using (W), we obtain the bound

$$\tau_\Delta = \frac{\Delta_k}{\|w_k d_k^{2r} g_k\|_p} = \frac{\Delta_k}{\|\hat{w}_k \hat{g}_k\|_p} \geq \frac{\Delta_k}{c_w \|\hat{g}_k\|_p},$$

and using the positivity of d_k, the definition (4.3) of d_{I}, (D3), and $r \geq 1/2$, we obtain the following bound for $\tau_{\mathcal{B}}$:

$$\tau_{\mathcal{B}} = \min\left\{ \inf_{\{g_k(x)<0\}} \frac{b(x) - u_k(x)}{-d_k^{2r}(x) g_k(x)}, \inf_{\{g_k(x)>0\}} \frac{u_k(x) - a(x)}{d_k^{2r}(x) g_k(x)} \right\} = \inf_{\{g_k(x)\neq 0\}} \frac{d_{\mathrm{I}}(u_k)(x)}{d_k^{2r}(x)|g_k(x)|}$$
$$\geq \inf_{\{g_k(x)\neq 0\}} \frac{d_k^{1-2r}(x)}{|g_k(x)|} \geq \inf_{\{g_k(x)\neq 0\}} \frac{c_d^{1-2r}}{|g_k(x)|} \geq \frac{c_d^{1-2r}}{\|g_k\|_\infty}.$$

We have $\phi(\tau) = -\kappa_1 \tau + \frac{1}{2}\kappa_2 \tau^2$ with

$$\kappa_1 = \langle d_k^{2r} g_k, g_k \rangle = \|\hat{g}_k\|_2^2, \quad \kappa_2 = \langle d_k^{2r} g_k, M_k d_k^{2r} g_k \rangle = \langle \hat{g}_k, \hat{M}_k \hat{g}_k \rangle$$

and observe that $|\kappa_2| \leq \|\hat{M}_k\|_{U,U'} \|\hat{g}_k\|_p^2$. Let τ^* be a minimizer for ϕ on $[0, \tau^+]$.

If $\tau^* < \tau^+$, then $\kappa_2 > 0$, $\tau^* = \kappa_1/\kappa_2$, and

$$\phi(\tau^*) = -\frac{1}{2}\frac{\kappa_1^2}{\kappa_2} \leq -\frac{1}{2}\frac{\|\hat{g}_k\|_2^4}{\|\hat{M}_k\|_{U,U'} \|\hat{g}_k\|_p^2}. \tag{6.5}$$

If $\tau^* = \tau_\Delta$ and $\kappa_2 > 0$, then $\kappa_1/\kappa_2 \ge \tau_\Delta$ and

$$\phi(\tau^*) = -\kappa_1\tau_\Delta + \frac{\kappa_2}{2}\tau_\Delta^2 \le -\frac{\kappa_1}{2}\tau_\Delta \le -\frac{1}{2}\frac{\|\hat{g}_k\|_2^2}{c_w\|\hat{g}_k\|_p}\Delta_k. \tag{6.6}$$

If $\tau^* = \tau_\Delta$ and $\kappa_2 \le 0$, then even

$$\phi(\tau^*) \le -\kappa_1\tau_\Delta < -\frac{1}{2}\frac{\|\hat{g}_k\|_2^2}{c_w\|\hat{g}_k\|_p}\Delta_k. \tag{6.7}$$

For $\tau^* = \tau_\mathcal{B}$ analogous arguments show that

$$\phi(\tau^*) \le -\frac{\kappa_1}{2}\tau_\mathcal{B} \le -\frac{1}{2}c_d^{1-2r}\frac{\|\hat{g}_k\|_2^2}{\|g_k\|_\infty}. \tag{6.8}$$

The first inequality (6.1) now follows from the estimates (6.5)–(6.8) and (6.4). The second inequality (6.2) follows from (6.1) and the application

$$\|\hat{g}_k\|_{p'} \le m_{p',2}\|\hat{g}_k\|_2$$

of Lemma 3.2. Note that $p \ge 2$ and $1/p + 1/p' = 1$ yield $p' \le 2$.

The inequality (6.3) follows from (6.2) by applying the following observations. Assumptions (A3) and (D3) yield $\|\hat{g}_k\|_p = \|d_k^r g_k\|_p \le c_d^r \mu(\Omega)^{1/p}\|g_k\|_\infty \le c_d^r c_1 \mu(\Omega)^{1/p}$. Assumptions (D3), (D4), and (A5) imply that $\|\hat{M}_k\|_{U,U'}$ are uniformly bounded by a constant $c_3 > 0$ (see Remark 5.1). □

Let the assumptions of Lemma 6.1 hold. If the kth iteration of Algorithm 5.7 is successful, i.e., $\rho_k > \eta_1$ (or equivalently $u_{k+1} \ne u_k$), then (6.3) guarantees that with $c_5 = \eta_1 c_4' > 0$ the following estimate for the actual decrease holds:

$$f_k - f_{k+1} > c_5\|\hat{g}_k\|_{p'}^2 \min\left\{\Delta_k, \|\hat{g}_k\|_{p'}^2, c_d^{1-2r}\right\}. \tag{6.9}$$

The next statement is trivial.

Lemma 6.2. *Let* (Δ_k) *and* (ρ_k) *be generated by Algorithm* 5.7. *If* $\rho_k \ge \eta_2$ *for sufficiently large* k *then* (Δ_k) *is bounded away from zero.*

Now we can prove a first global convergence result.

Theorem 6.3. *Let assumptions* (A1)–(A3), (A5), (D1)–(D4), *and* (W) *hold. Let the sequence* (u_k) *be generated by Algorithm* 5.7. *Then*

$$\liminf_{k\to\infty}\|d_k^r g_k\|_{p'} = 0.$$

Even more,

$$\liminf_{k\to\infty}\|d_k^r g_k\|_q = 0 \quad \textit{for all } 1 \le q < \infty.$$

Proof. Assume that there are $K > 0$ and $\varepsilon > 0$ with $\|\hat{g}_k\|_{p'} \ge \varepsilon$ for all $k \ge K$. First we will show that this implies $\sum_{k=0}^\infty \Delta_k < \infty$. If there is only a finite number of successful steps then $\Delta_{k+1} \le \gamma_1\Delta_k$ for large k and we are done. Otherwise, if the sequence (k_i) of successful steps does not terminate, we conclude from $f_k \downarrow$ and the boundedness of f (see Remark 3.1) that $\sum_{k=0}^\infty (f_k - f_{k+1}) < \infty$.

For all $k = k_i$ we may use (6.9) and obtain, since $\|\hat{g}_{k_i}\|_{p'} \geq \varepsilon$ for $k_i \geq K$, that Δ_{k_i} tends to zero and, moreover, obeys the inequality

$$\Delta_{k_i} < \frac{1}{c_5 \varepsilon^2}(f_{k_i} - f_{k_i+1})$$

for all k_i sufficiently large. This shows $\sum_{i=0}^{\infty} \Delta_{k_i} < \infty$. Since for all successful steps $k \in \{k_i\}$ we have $\Delta_{k+1} \leq \gamma_2 \Delta_k$ and for all others $\Delta_{k+1} \leq \gamma_1 \Delta_k$, we conclude

$$\sum_{k=0}^{\infty} \Delta_k \leq \sum_{i=0}^{\infty} \Delta_{k_i} \left(1 + \frac{\gamma_2}{1-\gamma_1}\right) < \infty. \tag{6.10}$$

In a second step we will show that $|\rho_k - 1| \to 0$. Due to

$$\|u_{k+1} - u_k\|_p \leq \|s_k\|_p \leq \beta_0 \|w_k^{-1}\|_\infty \Delta_k \leq \beta_0 c_{w'} \Delta_k \tag{6.11}$$

and (6.10), (u_k) is a Cauchy sequence in U. Furthermore,

$$\begin{aligned} 2\left|\psi_k(s_k) - \langle s_k, g_k\rangle - \frac{1}{2}\langle s_k, C_k s_k\rangle\right| &= |\langle s_k, B_k s_k\rangle| \leq \|B_k\|_{U,U'} \|s_k\|_p^2 \\ &\leq c_2 \beta_0^2 c_{w'}^2 \Delta_k^2. \end{aligned}$$

The mean value theorem yields $f(u_k + s_k) - f_k = \langle s_k, \bar{g}_k\rangle$ for some $\tau_k \in [0,1]$ and $\bar{g}_k = g(u_k + \tau_k s_k)$, and hence

$$\begin{aligned} |pred_k(s_k)||\rho_k - 1| &= \left|f(u_k + s_k) - f_k + \frac{1}{2}\langle s_k, C_k s_k\rangle - \psi_k(s_k)\right| \\ &\leq \left|\langle s_k, g_k\rangle + \frac{1}{2}\langle s_k, C_k s_k\rangle - \psi_k(s_k)\right| + |\langle s_k, \bar{g}_k - g_k\rangle| \\ &\leq \left(\frac{c_2}{2}\beta_0^2 c_{w'}^2 \Delta_k + \beta_0 c_{w'} \|\bar{g}_k - g_k\|_{p'}\right) \Delta_k. \end{aligned}$$

Since (u_k) converges in the closed set $\mathcal{B}$, g is continuous, and (Δ_k) as well as $(\|s_k\|_p)$ (see (6.11)) tend to zero, the first factor in the last expression converges to zero, too. Equation (6.3) guarantees that $|pred_k(s_k)|/\Delta_k$ is uniformly bounded away from zero for $k \geq K$, since by assumption $\|\hat{g}_k\|_{p'} \geq \varepsilon$. This shows $|\rho_k - 1| \to 0$. But now Lemma 6.2 yields a contradiction to $\Delta_k \to 0$. Therefore, the assumption is wrong and the first part of the assertion holds.

The second part follows from Lemma 3.2 for $1 \leq q \leq p'$ and from (A3) and the interpolation inequality (3.1) for $p' < q < \infty$. □

Now we will show that if $\hat{g}$ is uniformly continuous, the limits inferior in Theorem 6.3 can be replaced by limits.

We introduce the following assumption:

(A6) The scaled gradient $\hat{g} = d^r g : \mathcal{B} \subset U \longrightarrow U'$ is uniformly continuous.

Condition (A6) is not as easy to verify for most choices of d. With Lemma 6.4, however, we provide a very helpful tool to check the validity of (A6). Moreover, we show in Lemma 6.5 that the following, more convenient, condition implies (A6).

(A6′) The gradient $g : \mathcal{B} \subset U \longrightarrow U'$ is uniformly continuous and $d = d_{\mathrm{I}}$ or $d = d_{\mathrm{II}}$.

The proofs of both lemmas can be found in the appendix. As a by-product of our investigations we get the valuable result that $\hat{g}$ inherits the continuity of g if we choose $d = d_{\mathrm{I}}$ or $d = d_{\mathrm{II}}$. We will derive the results concerning continuity and uniform continuity of $\hat{g}$ simultaneously. Additional requirements for the uniform continuity are written in parentheses.

LEMMA 6.4. *Let* (A1)–(A3), (D3) *hold and* $g : \mathcal{B} \subset U \longrightarrow U'$ *be* (*uniformly*) *continuous. Assume that* $\|\chi_{\{g(u)g(\tilde{u})>0\}}(d(u) - d(\tilde{u}))\|_{p'}$ *tends to zero* (*uniformly in* $u \in \mathcal{B} \subset U$) *for* $\tilde{u} \to u$ *in* $\mathcal{B} \subset U$. *Then* $\hat{g} = d^r g : \mathcal{B} \subset U \longrightarrow U'$ *is* (*uniformly*) *continuous.*

Proof. See appendix. □

The previous lemma is now applicable to the choices $d = d_{\mathrm{I}}$ and $d = d_{\mathrm{II}}$, as follows.

LEMMA 6.5. *Let* (A1)–(A3) *hold and* $d = d_{\mathrm{I}}$ *or* $d = d_{\mathrm{II}}$. *Then* $\hat{g} = d^r g : \mathcal{B} \subset U \longrightarrow U'$ *is continuous. If, in addition, g is uniformly continuous, then the same is true for* $\hat{g}$.

Proof. See appendix. □

Now we state the promised variant of Theorem 6.3.

THEOREM 6.6. *Let assumptions* (A1)–(A3), (A5), (D1)–(D4), (W), *and* (A6) *or* (A6′) *hold. Then the sequence* (u_k) *generated by Algorithm* 5.7 *satisfies*

$$\lim_{k\to\infty} \|d_k^r g_k\|_{p'} = 0. \tag{6.12}$$

Even more,

$$\lim_{k\to\infty} \|d_k^r g_k\|_q = 0 \quad \text{for all } 1 \le q < \infty. \tag{6.13}$$

Proof. Since, due to Lemma 6.5, $\hat{g} = d^r g$ is uniformly continuous, it suffices to show that under the assumption $\|\hat{g}_k\|_{p'} \ge \varepsilon_1 > 0$ for an infinite number of iterations k there exists a sequence of index pairs (m_i, l_i) with $\|\hat{g}_{m_i} - \hat{g}_{l_i}\|_{p'} \ge \delta > 0$ but $\|u_{m_i} - u_{l_i}\|_p \to 0$, which is a contradiction to the uniform continuity of $\hat{g}$.

Let us assume that (6.12) does not hold. Then there is $\varepsilon_1 > 0$ and a sequence (m_i) with $\|\hat{g}_{m_i}\|_{p'} \ge \varepsilon_1$. Theorem 6.3 yields a sequence (k_i) with $\|\hat{g}_{k_i}\|_{p'} \to 0$. For arbitrary $0 < \varepsilon_2 < \varepsilon_1$ we can thus find a sequence (l_i) such that

$$\|\hat{g}_k\|_{p'} \ge \varepsilon_2, \quad m_i \le k < l_i, \quad \|\hat{g}_{l_i}\|_{p'} < \varepsilon_2.$$

Since $\hat{g}_{l_i} \ne \hat{g}_{l_i - 1}$, iteration $l_i - 1$ is successful and one has for all successful iterations k, $m_i \le k < l_i$, by Lemma 6.1 and (6.9)

$$f_k - f_{k+1} > c_5 \varepsilon_2^2 \min\left\{\Delta_k, \varepsilon_2^2, c_d^{1-2r}\right\}. \tag{6.14}$$

The left-hand side converges to zero, because (f_k) is nonincreasing and bounded from below, i.e., it is a Cauchy sequence. We conclude that Δ_k tends to zero for successful steps $m_i \le k < l_i$ and with (6.11) we get that

$$f_k - f_{k+1} \ge c_5 \varepsilon_2^2 \Delta_k \ge \frac{c_5 \varepsilon_2^2}{\beta_0 c_{w'}} \|u_{k+1} - u_k\|_p \overset{\text{def}}{=} c_6 \|u_{k+1} - u_k\|_p,$$

which is clearly true also for unsuccessful iterations. Summing and using the triangle inequality yields

$$f_{m_i} - f_{l_i} \ge c_6 \|u_{m_i} - u_{l_i}\|_p.$$

Since (f_k) is a Cauchy sequence, the left-hand side converges to zero for $i \to \infty$. Hence, $\|u_{m_i} - u_{l_i}\|_p \longrightarrow 0$ but

$$\|\hat{g}_{m_i} - \hat{g}_{l_i}\|_{p'} \geq \|\hat{g}_{m_i}\|_{p'} - \|\hat{g}_{l_i}\|_{p'} \geq \varepsilon_1 - \varepsilon_2 > 0.$$

This is a contradiction to the uniform continuity of $\hat{g}$. The second assertion follows as in the proof of Theorem 6.3. □

Remark 6.7. The convergence results in Theorems 6.3 and 6.6 remain valid if the definitions (5.12), (5.13) of predicted reduction $pred_k$ and actual reduction $ared_k$ are replaced by (5.15), the choices in [6].

To see that this is true, note that the crucial estimate (6.2) remains valid for $pred_k^1(s_k)$. Under (D4), $\langle s_k, C_k s_k\rangle \geq 0$ and, consequently,

$$ared_k^1(s_k) \leq ared_k(s_k).$$

Hence, (6.9) remains valid for successful iterations if ρ_k is replaced by ρ_k^1. Finally, since

$$pred_k^1(s_k)(\rho_k^1 - 1) = pred_k(s_k)(\rho_k - 1),$$

the proof of Theorem 6.3 remains valid if $pred_k$ and ρ_k are replaced by $pred_k^1$ and ρ_k^1, respectively. Since the proof of Theorem 6.6 only depends on Theorem 6.3 and on (6.9) but not explicitly on $pred_k$ and $ared_k$, its proof remains valid without any changes.

7. Convergence to second-order optimal points. The first-order convergence results in the previous section could be shown under rather weak conditions on the trust-region step s_k and for arbitrary symmetric and bounded Hessian approximations. If stronger assumptions are imposed on B_k and on s_k, then it can be shown that every accumulation point of (u_k) satisfies the second-order necessary optimality conditions. This will be done in this section. We need the following assumption on the Hessian approximation:

(A7) For all accumulation points $\bar{u} \in U$ of (u_k) and all $\varepsilon > 0$ there is $\delta = \delta(\bar{u}, \varepsilon) > 0$ such that $\|u_k - \bar{u}\|_p \leq \delta$ implies $\|B_k - \nabla^2 f(\bar{u})\|_{U,U'} \leq \varepsilon$.

Obviously (A7) is satisfied if $B_k = \nabla^2 f(u_k)$ and if (A4) holds. However, (A7) also applies in other important situations. For example, (A7) applies if f is a least squares functional, $f(\bar{u}) = 0$, and B_k is the Gauss–Newton approximation of the Hessian.

The fraction of Cauchy decrease condition does not take into account any properties of the quadratic part of ψ_k. Apparently, this condition is too weak to guarantee the positivity of $\hat{M}(\bar{u})$ at accumulation points of (u_k). The decrease condition has to be strengthened in such a way that for $\bar{u}$ satisfying (O1) and (O2) but not (O3$''$) there are $\alpha, \varepsilon, c > 0$ such that $\psi_k(s_k) \leq -c \min\{\Delta_k^2, \alpha^2\}$ for all iterates u_k with $\|u_k - \bar{u}\|_p \leq \varepsilon$. For the finite-dimensional problem one can establish such an inequality near nondegenerate points $\bar{u}$ by using techniques similar to those of Coleman and Li [6] if the s_k satisfy a finite-dimensional fraction of optimal decrease condition of the form

$$\|w_k s_k\|_2 \leq \beta_0 \Delta_k\,,\ u_k + s_k \in \mathcal{B}^\circ\,,\ \text{and } \psi_k(s_k) < \beta \psi_k^o, \tag{7.1a}$$

where $\psi_k^o = \psi_k(\tau_k s_k^o)$, $\tau_k = \max\{\tau \geq 0 \,:\, u_k + \tau s_k^o \in \mathcal{B}\}$, and s_k^o solves

$$\min \psi_k(s) \quad \text{subject to} \quad \|w_k s\|_2 \leq \Delta_k. \tag{7.1b}$$

This approach is not directly transferable to our setting because the example

$$\min \quad -\int_0^1 ts^2(t)\,dt \quad \text{subject to} \quad \|s\|_2 \le \Delta \tag{7.2}$$

shows that even in a Hilbert space s_k^o may not exist. Moreover, the proofs in [6] use extensively a convenient characterization of s_k^o derived from the KKT conditions [20] and the equivalence of 2-norm and ∞-norm in $\mathbb{R}^N$. Since, as shown by (7.2), in Banach space the quadratic subproblem may not have a solution, this is not applicable in our framework. Our convergence proof requires that the trial steps yield a fraction of the Cauchy decrease and, moreover, a fraction of the decrease achievable along directions of negative curvature of ψ_k at $s = 0$. For convenience and simplicity of notation, however, we favor a more intuitive but stronger condition which is formulated in the flavor of (7.1). Our fraction of optimal decrease condition is given by the following. There exist $\beta, \beta_0 > 0$ (fixed for all k) such that

$$\|w_k s_k\|_p \le \beta_0 \Delta_k\,, \; u_k + s_k \in \mathcal{B}^\circ\,, \text{ and } \psi_k(s_k) < \beta\psi_k^o, \tag{7.3a}$$

where

$$\psi_k^o \overset{\text{def}}{=} \inf \psi_k(s) \quad \text{subject to} \; u_k + s \in \mathcal{B}\,, \; \|w_k s\|_p \le \Delta_k. \tag{7.3b}$$

There are two important differences between (7.1) and (7.3). Both arise in the definition of ψ_k^o. First, the min in (7.1b) is replaced by an inf in (7.3b) for the reasons explained above. Second, in the computation of ψ_k^o in (7.3b), the bound constraints are explicitly enforced. In the finite-dimensional context (7.1b) this is not necessary, since E_k tends to equilibrate the distance to the bounds along the optimal decrease direction s_k^o. As a consequence, the bound constraints can be enforced by simple scaling of the solution s_k^o of (7.1b) by τ_k. This equilibration property, however, makes use of the equivalence of the 2-norm and the ∞-norm in finite dimensions. Recall the discussion at the end of section 2.

To obtain global convergence results toward points satisfying second-order necessary optimality conditions, the fraction of optimal decrease condition has to be incorporated into Algorithm 5.7. Step 2.5 of Algorithm 5.7 has to be replaced by

2.5.′ Compute s_k satisfying (7.3).

When we refer to Algorithm 5.7 in this section we assume that step 2.5 is replaced by step 2.5′.

In the next lemma we show that in a neighborhood of an accumulation point $\bar{u}$ of (u_k) at which (O1), (O2) but not (O3″) hold, one can find a direction of negative curvature h^n of ψ_k such that $u_k \pm h^n \in \mathcal{B}$.

Lemma 7.1. *Let assumptions* (A1), (A2), (A4), (A5), (A7), (D1)–(D5) *hold and let the sequence* (u_k) *be generated by Algorithm* 5.7. *Assume that* $\bar{u} \in \mathcal{B}$ *is an accumulation point of* (u_k) *with* $\hat{g}(\bar{u}) = 0$ *and that there are* $\bar{h} \in V$, $\bar{h} \ne 0$, *and* $\lambda > 0$ *with*

$$\langle \bar{h}, \hat{M}(\bar{u})\bar{h} \rangle \le -\lambda \|\bar{h}\|_p^2. \tag{7.4}$$

Then there exist $\varepsilon, \alpha, \hat{\lambda} > 0$ *such that for all* u_k *with* $\|u_k - \bar{u}\|_p \le \varepsilon$ *one can find* $h \in V$, $\|h\|_p = 1$, *with* $u_k + \tau\alpha d_k^r h \in \mathcal{B}$ *for all* $\tau \in [-1, 1]$ *and*

$$\langle h, \hat{M}_k h \rangle \le -\hat{\lambda} \|h\|_p^2. \tag{7.5}$$

Proof. Since $\bar{u} \in \mathcal{B}$ and $\hat{g}(\bar{u}) = 0$, (O1) and (O2) are satisfied due to Lemma 4.2. Lemma 5.2 yields $I \overset{\text{def}}{=} \{x \in \Omega : \ a(x) < \bar{u}(x) < b(x)\} = \{x \in \Omega : \ d(\bar{u})(x) > 0\} \overset{\text{def}}{=} I^*$. Define $I_\delta = \{x \in \Omega : \ a(x) + \delta \le \bar{u}(x) \le b(x) - \delta\}$ for arbitrary $0 < \delta < 4\delta_d$ with δ_d as in (D2). We write $v_A = \chi_A v$ for measurable functions v and measurable sets $A \subset \Omega$.

We first show that (7.4) implies the existence of $\tilde{h} \in V$ with $\|\tilde{h}\|_p = 1$, $\{\tilde{h} \neq 0\} \subset I_\delta$, and

$$\langle \tilde{h}, \hat{M}(\bar{u})\tilde{h} \rangle \le -\frac{\lambda}{2}. \tag{7.6}$$

In the first step toward (7.6) we use the techniques of the proof of Lemma 5.3. More precisely, applying the inequalities (5.7) with s replaced by $\bar{h}$ we obtain that

$$0 > -\lambda \|\bar{h}\|_p^2 \ge \langle \bar{h}, \hat{M}(\bar{u})\bar{h} \rangle \ge \langle \bar{h}_I, \hat{M}(\bar{u})\bar{h}_I \rangle.$$

Hence, $\bar{h}_I \in V \setminus \{0\}$ and

$$\langle \bar{h}_I, \hat{M}(\bar{u})\bar{h}_I \rangle \le -\lambda \|\bar{h}\|_p^2 \le -\lambda \|\bar{h}_I\|_p^2 < 0, \tag{7.7}$$

i.e., (7.4) holds for $\bar{h}_I$ instead of $\bar{h}$. Furthermore, the symmetry of $\hat{M}(\bar{u})$ and the identity $\bar{h}_I = \bar{h}_{I_\delta} + \bar{h}_{I \setminus I_\delta}$ imply

$$\begin{aligned}
\langle \bar{h}_{I_\delta}, \hat{M}(\bar{u})\bar{h}_{I_\delta} \rangle &= \langle \bar{h}_I, \hat{M}(\bar{u})\bar{h}_I \rangle - \langle \bar{h}_{I_\delta} + \bar{h}_I, \hat{M}(\bar{u})\bar{h}_{I\setminus I_\delta} \rangle \\
&\le -\lambda \|\bar{h}_I\|_p^2 + \|\hat{M}(\bar{u})\bar{h}_{I\setminus I_\delta}\|_{p'} (\|\bar{h}_I\|_p + \|\bar{h}_{I_\delta}\|_p) \\
&\le -\lambda \|\bar{h}_I\|_p^2 + 2\|\hat{M}(\bar{u})\bar{h}_{I\setminus I_\delta}\|_{p'} \|\bar{h}_I\|_p.
\end{aligned}$$

By the definition of $\hat{M}(\bar{u})$ and the fact that $g(\bar{u})$ and thus, by (D4), $e(\bar{u})$ vanishes on I, with (D3) we get

$$\begin{aligned}
\|\hat{M}(\bar{u})\bar{h}_{I\setminus I_\delta}\|_{p'} &= \|(D^r(\bar{u})\nabla^2 f(\bar{u}) D^r(\bar{u}) + E(\bar{u}) D^{2r-1}(\bar{u}))\bar{h}_{I\setminus I_\delta}\|_{p'} \\
&= \|D^r(\bar{u})\nabla^2 f(\bar{u}) D^r(\bar{u})\bar{h}_{I\setminus I_\delta}\|_{p'} \le c_d^r \|\nabla^2 f(\bar{u}) D^r(\bar{u})\bar{h}_{I\setminus I_\delta}\|_{p'}.
\end{aligned}$$

Since the measure of $I \setminus I_\delta$ can be made arbitrarily small by reducing $\delta > 0$ we conclude that $\bar{h}_{I\setminus I_\delta}$ and $D^r(\bar{u})\bar{h}_{I\setminus I_\delta}$ tend to zero for $\delta \to 0$ in all spaces $L^q(\Omega)$, $1 \le q < \infty$. Hence, using (A4), we find $\delta > 0$ with

$$2\|\hat{M}(\bar{u})\bar{h}_{I\setminus I_\delta}\|_{p'} \le \frac{\lambda}{2}\|\bar{h}_I\|_p.$$

Obviously, $\{\tilde{h} \neq 0\} \subset I_\delta$. Thus, (7.6) holds with $\tilde{h} = \bar{h}_{I_\delta} / \|\bar{h}_{I_\delta}\|_p$.

Next, we use $\tilde{h}$ to construct h such that the assertions of the lemma are valid.

For $\varepsilon > 0$ and u_k with $\|u_k - \bar{u}\|_p \le \varepsilon$, define $h \in V$ by

$$h(x) = \begin{cases} \tilde{h}(x)\dfrac{d^r(\bar{u})(x)}{d_k^r(x)} & \text{if} \quad \min\{u_k(x) - a(x), b(x) - u_k(x)\} > \dfrac{\delta}{4}, \\ 0 & \text{otherwise.} \end{cases}$$

We have $I_h \overset{\text{def}}{=} \{h \neq 0\} \subset I_\delta$ and conclude from assumptions (D2) and (D3) that $\varepsilon_d(\delta/4) \le d_k(x) \le c_d$ on I_h and $\varepsilon_d(\delta/4) \le d(\bar{u})(x) \le c_d$ on I_δ, which implies that

$$\|h\|_p \le \gamma \ , \ \|h\|_\infty \le \gamma \|\tilde{h}\|_\infty \ , \ \|h\|_p \ge \frac{1}{\gamma}\|\tilde{h}_{I_h}\|_p \quad \text{with} \quad \gamma = \frac{c_d^r}{\varepsilon_d^r(\delta/4)}. \tag{7.8}$$

From $\bar{u}(x) - a(x), b(x) - \bar{u}(x) \geq \delta$ on I_δ follows

$$I_\delta \setminus I_h \subset \{x \in \Omega \, : \, |u_k(x) - \bar{u}(x)| \geq 3\delta/4\} \, .$$

If $p = \infty$, we achieve $\tilde{h}_{I_\delta \setminus I_h} = 0$ for $\varepsilon < 3\delta/4$. Otherwise, due to Lemma 3.4, we can make $\|\tilde{h}_{I_\delta \setminus I_h}\|_p \leq \mu(I_\delta \setminus I_h)^{1/p} \|\tilde{h}\|_\infty$ arbitrarily small by making $\varepsilon > 0$ small. Hence, in all cases we can reduce ε such that

$$\|h\|_p \geq \frac{1}{\gamma} \|\tilde{h}_{I_h}\|_p \geq \frac{1}{\gamma} \left(\|\tilde{h}\|_p - \|\tilde{h}_{I_\delta \setminus I_h}\|_p \right) \geq \frac{1}{2\gamma}. \tag{7.9}$$

By first using the definition of h, $I_h = \{h \neq 0\}$, then the fact that $g(\bar{u})(x) = 0$ on $I_\delta \supset I_h$, the definitions of B_k, $\hat{M}(\bar{u})$, and Hölder's inequality, and finally (D3), (D5), we get

$$\begin{aligned}
\langle h, \hat{M}_k h \rangle &= \langle h, d_k' g_k d_k^{2r-1} h \rangle + \langle d_k^r h, B_k d_k^r h \rangle \\
&= \langle d_k' d_k^{2r-1}, \chi_{I_h}(g_k h^2) \rangle + \langle d^r(\bar{u}) \tilde{h}_{I_h}, B_k d^r(\bar{u}) \tilde{h}_{I_h} \rangle \\
&= \langle d_k' d_k^{2r-1}, \chi_{I_h}(g_k h^2) \rangle + \langle d^r(\bar{u}) \tilde{h}_{I_h}, B_k d^r(\bar{u}) \tilde{h}_{I_h} \rangle + \langle \tilde{h}_{I_h}, d'(\bar{u}) g(\bar{u}) d^{2r-1}(\bar{u}) \tilde{h}_{I_h} \rangle \\
&\leq \|d_k'\|_\infty \|d_k\|_\infty^{2r-1} \|\chi_{I_h} g_k\|_{p'} \|h\|_p \|h\|_\infty + \langle \tilde{h}_{I_h}, \hat{M}(\bar{u}) \tilde{h}_{I_h} \rangle \\
&\quad + \|d(\bar{u})\|_\infty^{2r} \|B_k - \nabla^2 f(\bar{u})\|_{U,U'} \|\tilde{h}_{I_h}\|_p^2 \\
&\leq c_{d'} c_d^{2r-1} \|\chi_{I_h} g_k\|_{p'} \|h\|_\infty \|h\|_p + \langle \tilde{h}_{I_h}, \hat{M}(\bar{u}) \tilde{h}_{I_h} \rangle + c_d^{2r} \|B_k - \nabla^2 f(\bar{u})\|_{U,U'} \|\tilde{h}_{I_h}\|_p^2.
\end{aligned}$$

From $\tilde{h} = \tilde{h}_{I_\delta} = \tilde{h}_{I_h} + \tilde{h}_{I_\delta \setminus I_h}$ we obtain

$$\langle \tilde{h}_{I_h}, \hat{M}(\bar{u}) \tilde{h}_{I_h} \rangle = \langle \tilde{h}, \hat{M}(\bar{u}) \tilde{h} \rangle - \langle \tilde{h} + \tilde{h}_{I_h}, \hat{M}(\bar{u}) \tilde{h}_{I_\delta \setminus I_h} \rangle.$$

Using this equality, the fact that $g(\bar{u})(x) = 0$ on $I_\delta \supset I_h$, and $\|\tilde{h}_{I_h}\|_p \leq \|\tilde{h}\|_p$ in the previous estimate for $\langle h, \hat{M}_k h \rangle$ gives

$$\begin{aligned}
\langle h, \hat{M}_k h \rangle \leq \; & c_{d'} c_d^{2r-1} \|g_k - g(\bar{u})\|_{p'} \|h\|_\infty \|h\|_p + \langle \tilde{h}, \hat{M}(\bar{u}) \tilde{h} \rangle \\
& - \langle \tilde{h} + \tilde{h}_{I_h}, \hat{M}(\bar{u}) \tilde{h}_{I_\delta \setminus I_h} \rangle + c_d^{2r} \|B_k - \nabla^2 f(\bar{u})\|_{U,U'} \|\tilde{h}_{I_h}\|_p^2 \\
\leq \; & c_{d'} c_d^{2r-1} \|g_k - g(\bar{u})\|_{p'} \|h\|_\infty \|h\|_p + \langle \tilde{h}, \hat{M}(\bar{u}) \tilde{h} \rangle \\
& + 2 \|\tilde{h}\|_p \|\hat{M}(\bar{u})\|_{U,U'} \|\tilde{h}_{I_\delta \setminus I_h}\|_p + c_d^{2r} \|B_k - \nabla^2 f(\bar{u})\|_{U,U'} \|\tilde{h}_{I_h}\|_p^2.
\end{aligned}$$

Using (7.6), (7.8), (7.9), and $\|\tilde{h}\|_p = 1$ we arrive at the estimate

$$\begin{aligned}
\langle h, \hat{M}_k h \rangle \leq \; & \big(2 c_{d'} c_d^{2r-1} \gamma^2 \|\tilde{h}\|_\infty \|g_k - g(\bar{u})\|_{p'} - \frac{\lambda}{2\gamma^2} + 8\gamma^2 \|\hat{M}(\bar{u})\|_{U,U'} \|\tilde{h}_{I_\delta \setminus I_h}\|_p \\
& + c_d^{2r} \gamma^2 \|B_k - \nabla^2 f(\bar{u})\|_{U,U'} \big) \|h\|_p^2.
\end{aligned}$$

We have already shown that $\|\tilde{h}_{I_\delta \setminus I_h}\|_p$ can be made arbitrarily small by making $\varepsilon > 0$ small. By continuity the same is true for $\|g_k - g(\bar{u})\|_{p'}$ and by (A7) for $\|B_k - \nabla^2 f(\bar{u})\|_{U,U'}$ (since $\|u_k - \bar{u}\|_p \leq \varepsilon$). Hence, there exist $\varepsilon > 0$ and $\hat{\lambda} > 0$ such that for all u_k with $\|u_k - \bar{u}\|_p \leq \varepsilon$ we can carry out the above construction to obtain $h \in V \setminus \{0\}$ with

$$\langle h, \hat{M}_k h \rangle \leq -\hat{\lambda} \|h\|_p^2.$$

Since $h \neq 0$, $\|h\|_\infty \leq \gamma \|\tilde{h}\|_\infty$, and $\|h\|_p \geq \frac{1}{2\gamma}$, where γ depends only on δ, we get

$$\frac{\|h\|_\infty}{\|h\|_p} \leq 2\gamma^2 \|\tilde{h}\|_\infty \overset{\text{def}}{=} C.$$

In addition, we have by construction $I_h \subset \{x \in \Omega \,:\, a(x) + \delta/4 \leq u_k(x) \leq b(x) - \delta/4\}$ and consequently

$$u_k + \tau \frac{\delta}{4Cc_d^r} d_k^r \frac{h}{\|h\|_p} \in \mathcal{B} \text{ for all } \tau \in [-1, 1].$$

Setting $\alpha = \delta/4Cc_d^r$ and renorming h to unity completes the proof. □

Now we establish the required decrease estimate.

Lemma 7.2. *Let assumptions* (A1), (D3), (D4), (W) *hold and s_k satisfy* (7.3). *If for u_k there exist $\hat{\lambda}, \alpha > 0$, $h_k \in V$, $\|h_k\|_p = 1$, with $u_k + \tau\alpha d_k^r h_k \in \mathcal{B}$ for all $\tau \in [-1, 1]$ and*

$$\langle h_k, \hat{M}_k h_k \rangle \leq -\hat{\lambda} \|h_k\|_p^2,$$

then

$$pred_k(s_k) \geq -\psi_k(s_k) \geq \frac{\beta\hat{\lambda}}{2} \min\left\{ \frac{\Delta_k^2}{c_w^2}, \alpha^2 \right\}. \tag{7.10}$$

Proof. The first inequality is obvious. Now let $\hat{\lambda}, \alpha > 0$ be given. For all u_k which admit $h_k \in V$, $\|h_k\|_p = 1$, with $u_k \pm \alpha d_k^r h_k \in \mathcal{B}$ and $\langle h_k, \hat{M}_k h_k \rangle \leq -\hat{\lambda} \|h_k\|_p^2$, set

$$\hat{s}_k^n = \pm \min\{\Delta_k / c_w, \alpha\} h_k \text{ and } s_k^n = d_k^r \hat{s}_k^n,$$

and choose the sign such that $\langle \hat{s}_k^n, \hat{g}_k \rangle \leq 0$. Then $\|w_k s_k^n\|_p \leq \Delta_k$ by assumption (W) and $u_k + s_k^n \in \mathcal{B}$. Hence s_k^n is admissible for (7.3b) and can be used to get an upper bound for $\psi_k(s_k)$. The fraction of optimal decrease condition (7.3) gives

$$\begin{aligned} \psi_k(s_k) &\leq \beta\psi_k(s_k^n) = \beta\hat{\psi}_k(\hat{s}_k^n) = \beta\langle \hat{s}_k^n, \hat{g}_k \rangle + \frac{\beta}{2} \langle \hat{s}_k^n, \hat{M}_k \hat{s}_k^n \rangle \leq \frac{\beta}{2} \langle \hat{s}_k^n, \hat{M}_k \hat{s}_k^n \rangle \\ &\leq -\frac{\beta\hat{\lambda}}{2} \|\hat{s}_k^n\|_p^2 = -\frac{\beta\hat{\lambda}}{2} \min\left\{ \frac{\Delta_k^2}{c_w^2}, \alpha^2 \right\}. \qquad \square \end{aligned}$$

For a large class of trust-region algorithms for unconstrained finite-dimensional problems Shultz, Schnabel, and Byrd [19] proposed a very elegant way to prove that all accumulation points of the iterates satisfy the second-order necessary optimality conditions. The key idea is to increase the trust-region radius after exceedingly successful steps (case 4 in Algorithm 5.5). The following convergence theorem is an analogue to [19, Thm. 3.2].

Theorem 7.3. *Let assumptions* (A1)–(A7), (D1)–(D5), *and* (W) *hold. Moreover, let the sequence (u_k) be generated by the Algorithm* 5.7 *and let all s_k satisfy* (7.3). *Then every accumulation point $\bar{u} \in U$ of (u_k) satisfies the second-order necessary conditions* (O1)–(O3).

Proof. Let $\bar{u} \in U$ be an accumulation point of u_k. Then $\bar{u} \in \mathcal{B}$ and, since $\hat{g} : \mathcal{B} \subset U \longrightarrow U$ is continuous, $\hat{g}(\bar{u}) = 0$ by Theorem 6.6. Using Lemma 4.2, this implies (O1) and (O2).

Now assume that (O3) does not hold at $\bar{u}$. Then due to Theorem 5.4 there are $\bar{h} \in V$, $\bar{h} \neq 0$, and $\lambda > 0$ with $\langle \bar{h}, \hat{M}(\bar{u})\bar{h}\rangle \leq -\lambda \|\bar{h}\|_p^2$. Lemmas 7.1 and 7.2 yield $\alpha, c_7, \varepsilon > 0$ with $\mathit{pred}_k(s_k) \geq c_7 \min\{\Delta_k^2, \alpha^2\}$ for all u_k satisfying $\|u_k - \bar{u}\|_p \leq \varepsilon$. By choosing $0 < \Delta \leq \alpha$ we achieve that for all k with $\Delta_k \leq \Delta$ and $\|u_k - \bar{u}\|_p \leq \varepsilon$

$$\mathit{pred}_k(s_k) \geq c_7 \Delta_k^2.$$

Using this estimate, (A4), (A7), and $\|s_k\|_p \leq \beta_0 c_{w'} \Delta_k$ (see (6.11)) we find—possibly after reducing ε and Δ—with appropriate $\tau_k \in [0,1]$

$$\begin{aligned}
\mathit{pred}_k(s_k)\,|\rho_k - 1| &= \left| f(u_k + s_k) - f_k + \frac{1}{2}\langle s_k, C_k s_k\rangle - \psi_k(s_k)\right| \\
&= \frac{1}{2}\left|\langle s_k, (\nabla^2 f(u_k + \tau_k s_k) - B_k)s_k\rangle\right| \\
&\leq \frac{1}{2}\left(\|\nabla^2 f(u_k + \tau_k s_k) - \nabla^2 f(\bar{u})\|_{U,U'} + \|\nabla^2 f(\bar{u}) - B_k\|_{U,U'}\right)\|s_k\|_p^2 \\
&\leq (1-\eta_3)c_7\Delta_k^2 \leq (1-\eta_3)\mathit{pred}_k(s_k).
\end{aligned}$$

This shows $\rho_k \geq \eta_3$ for all k with $\Delta_k \leq \Delta$ and $\|u_k - \bar{u}\|_p \leq \varepsilon$ and hence $\Delta_{k+1} \in [\gamma_2\Delta_k, \gamma_3\Delta_k]$.

For all $K > 0$ there is $l > K$ with $\|u_l - \bar{u}\|_p \leq \varepsilon/2$ and $\rho_l > \eta_1$. In fact, since $\bar{u}$ is an accumulation point of (u_k), we can find $l' > K$ with $\|u_{l'} - \bar{u}\|_p \leq \varepsilon/2$. Now $\rho_k \leq \eta_1$ for all $k \geq l'$ cannot occur, because then $\Delta_k \leq \gamma_1^{k-l'}\Delta_{l'}$ eventually satisfies $\Delta_k \leq \Delta$ and consequently $\rho_k \geq \eta_3 > \eta_1$. Hence, there is $l \geq l' > K$ with $u_l = u_{l'}$ and $\rho_l > \eta_1$.

Since $\Delta_{k+1} \geq \gamma_2\Delta_k$ for all k with $\|u_k - \bar{u}\|_p \leq \varepsilon$ and $\Delta_k \leq \Delta$, it is easily seen that

1. $\Delta_l > \Delta$, or
2. $\Delta_l \leq \Delta$ and there is $m > l$ such that $\|u_k - \bar{u}\|_p \leq \varepsilon$ and $\Delta_k \leq \Delta$ for $l \leq k < m$, and
   2.1. $\Delta_m > \Delta$, or
   2.2. $\Delta_m \leq \Delta$ and $\|u_m - \bar{u}\|_p > \varepsilon$.

In case 1 we get

$$f_l - f_{l+1} > \eta_1 c_7 \min\{\Delta_l^2, \alpha^2\} \geq \eta_1 c_7 \Delta^2.$$

For case 2.1 we have $\Delta \geq \Delta_{m-1} \geq \Delta_m/\gamma_3 > \Delta/\gamma_3$, and $\rho_{m-1} \geq \eta_3$; hence

$$f_{m-1} - f_m \geq \eta_3 c_7 \Delta_{m-1}^2 \geq \eta_3 c_7 \frac{\Delta^2}{\gamma_3^2}.$$

In case 2.2 we get $\Delta_{k+1} \geq \gamma_2\Delta_k$, $k = l, \ldots, m-1$. This implies $\Delta_k \leq \gamma_2^{k-m+1}\Delta_{m-1}$ and

$$\begin{aligned}
\frac{\varepsilon}{2} &\leq \|u_m - \bar{u}\|_p - \|u_l - \bar{u}\|_p \leq \|u_m - u_l\|_p = \left\|\sum_{k=l}^{m-1} s_k\right\|_p \\
&\leq \sum_{k=l}^{m-1}\|s_k\|_p \leq \beta_0 c_{w'}\sum_{k=l}^{m-1}\Delta_k \leq \beta_0 c_{w'}\Delta_{m-1}\sum_{k=l}^{m-1}\gamma_2^{k-m+1} \leq \beta_0 c_{w'}\Delta_{m-1}\frac{\gamma_2}{\gamma_2 - 1}.
\end{aligned}$$

This yields

$$f_{m-1} - f_m \geq \eta_3 c_7 \Delta_{m-1}^2 \geq \eta_3 c_7 \left(\frac{\varepsilon(\gamma_2 - 1)}{2\beta_0 c_{w'} \gamma_2} \right)^2 .$$

Therefore, for infinitely many steps k we get a decrease $f_k - f_{k+1}$ of at least a constant value which yields $f_k \to -\infty$. This contradicts the boundedness of f on $\mathcal{B}$, which follows from (A1)–(A3) (see Remark 3.1). Thus, (O3) must hold at $\bar{u}$. □

Remark 7.4. As for the first-order convergence results, the second-order convergence result in Theorem 7.3 remains valid if the definitions (5.12), (5.13) of predicted reduction $pred_k$ and actual reduction $ared_k$ are replaced by (5.15), the choices in [6]. Since (7.10) remains valid for $pred_k^1$, in the proof of Theorem 7.3 $pred_k$ can be replaced by $pred_k^1$.

8. Examples. In this section we examine the examples presented in the introduction and show how the general class of algorithms presented in this paper can be applied in those cases. We give justifications for the assumptions (A1)–(A3) which are needed to guarantee the first-order convergence result Theorem 6.3. The other assumptions, which mostly address the second derivatives, can be justified in similar ways with proper adjustment of the assumptions on the problem data in the two problems. Since the mere statement of the second-order derivatives is rather lengthy, this has been omitted.

8.1. A parabolic boundary control problem of Stefan–Boltzmann type. If a metal rod is heated by radiation on its right side $x = 1$ with temperature $u(t)$, $t \in [0,T]$, the temperature distribution $y(t,x)$, $(t,x) \in [0,T] \times [0,1]$, satisfies in good approximation the heat equation with the Stefan–Boltzmann boundary condition (1.1). The control $u \in \mathcal{B} \overset{\text{def}}{=} \{u \in L^\infty(0,T) \ : \ 0 \leq u \leq 1\}$ shall be determined in such a way that the temperature $y(T,x)$ at time T follows a given temperature profile $y_d \in C([0,T])$. To this end we want to solve the problem

$$\begin{aligned} &\text{minimize} \quad f(u) \overset{\text{def}}{=} \frac{1}{2}\|y(u)(T,.) - y_d\|_2^2 + \frac{\alpha}{2}\|u\|_2^2 \\ &\text{subject to} \quad u \in \mathcal{B}, \end{aligned} \tag{8.1}$$

where $y = y(u)$ is the solution of (1.1) and $\alpha \geq 0$ is a regularization parameter. For the case $\alpha = 0$ this problem was considered in [18]. This optimization problem with $\alpha > 0$ is also considered in [14] and [15]. By defining mild solutions to (1.1) via a weakly singular integral equation of Volterra type it can be shown with the same techniques as in [18] that for $p > 2$ the mapping $u \in \mathcal{D} \subset L^p([0,T]) \longmapsto y(u) \in C([0,T] \times [0,1])$ is completely continuous and continuously Fréchet differentiable on an open neighborhood $\mathcal{D}$ of $\mathcal{B}$. Consequently, the objective function $f(u)$ of (8.1) is continuously Fréchet differentiable on $\mathcal{D} \subset L^p([0,T])$. Hence, assumption (A1) is shown for $U = L^p([0,T])$, $p > 2$. Since the 2-norm is weakly lower semicontinuous on $L^p([0,T])$, existence of an optimal solution to (8.1) can be proven as in [18], where only the case $\alpha = 0$ is considered. It remains to check (A2) and (A3). As shown in [18] the gradient representation $g(u)$ of $f(u)$ with respect to the L^2 dual pairing, $\langle h, \nabla f(u) \rangle = (h, g(u))_2$, is given by

$$g(u) = z(u,.,1) + \alpha u, \tag{8.2}$$

where $z = z(u) \in C([0,T] \times [0,1])$ is the mild solution of the adjoint equation

$$\begin{aligned} z_t(t,x) &= -z_{xx}(t,x), \quad (t,x) \in (0,T) \times (0,1) \\ z(T,x) &= y(T,x) - y_d(x), \quad x \in (0,1) \\ z_x(t,0) = 0, \quad z_x(t,1) &= -4y(t,1)^3 z(t,1), \quad t \in (0,T). \end{aligned} \tag{8.3}$$

In (8.3) $y = y(u)$ is the solution of (1.1). Using again the Volterra integral formulation of (8.3), it can be shown that $z(u)$ is uniformly bounded in $C([0,T] \times [0,1])$ for $u \in \mathcal{B}$, since this holds for $y(u)$. Hence, $\|g(u)\|_\infty \le \|z(u)\|_\infty + \alpha\|u\|_\infty \le c_1$ for all $u \in \mathcal{B}$. This shows that (A2) and (A3) also are satisfied.

For this example, the computation of the gradient $g_k = g(u_k)$ in step 2.1 of Algorithm 5.7 requires the solution $y = y(u_k)$ of the heat equation (1.1) and the solution $z = z(u_k)$ of the adjoint equation (8.3). In steps 2.2 and 2.3 we can use $d_k = d_{\text{II}}(u_k)$ defined by (4.4), with $\zeta = \kappa = 0.075$, $r = 1$, and $w_k = 1$, the parameter values corresponding to those in the numerical example of [23]. As we have noted, these choices satisfy (D1)–(D4) and (W). The function $d'(u)$ defining $e(u)$ can be chosen to be $d'(u) = d'_{\text{II}}(u)$ defined by (5.2). This choice satisfies (D5). For this example, $c(x) = 0.075 \min\{1-0,1\} = 0.075$ and thus

$$d_{\text{II}}(u)(t) = \begin{cases} \min\{|z(u,t,1) + \alpha u(t)|, 0.075\} & \text{if } -(z(u,t,1) + \alpha u(t)) > u(t) \\ & \text{and } u(t) \le 1 - u(t), \\ \min\{|z(u,t,1) + \alpha u(t)|, 0.075\} & \text{if } z(u,t,1) + \alpha u(t) > 1 - u(t) \\ & \text{and } 1 - u(t) \le u(t), \\ \min\{u(t), 1-u(t), 0.075\} & \text{else.} \end{cases}$$

In the simplest case one could find a step s_k in step 2.5 of the algorithm by one-dimensional minimization of $\psi_k(-t d_k^2 g_k)$. The second-order term $\langle d_k^2 g_k, B_k d_k^2 g_k \rangle$ could be obtained by using a finite difference approximation of $\nabla^2 f(u_k) d_k^2 g_k$.

8.2. A control problem of Bolza type. We consider the control problem of Bolza type

$$\begin{aligned} &\text{minimize} \quad f(u) \overset{\text{def}}{=} P(y(u)(1)) + \int_0^1 h_0(x, y(u)(x), u(x))\,dx \\ &\text{subject to} \quad u \in \mathcal{B}, \end{aligned} \tag{8.4}$$

where $y = y(u)$ is the solution of (1.2). Our presentation is based on Tian and Dunn [21]. We work in $L^2([0,1])$, i.e., $\Omega = [0,1]$, $p = p' = 2$, and $U = H = U' = L^2([0,1])$. Let the function $P : \mathbb{R}^m \longrightarrow \mathbb{R}$ be twice continuously differentiable. The function h_0 and the right-hand side $h = (h_1, \ldots, h_m)^T$ of the state equation (1.2) are given by

$$h_i : [0,1] \times \mathbb{R}^m \times \mathbb{R} \longrightarrow \mathbb{R}\ , \quad h_i(x,y,u) = h_i^0(x,y) + h_i^1(x,y)u + h_i^2(x,y)u^2.$$

Hereby, the functions $h_i^k : [0,1] \times \mathbb{R}^m \longrightarrow \mathbb{R}$ as well as their first and second partial y-derivatives are assumed to be continuous and the derivatives $\nabla_y h_i^k$, $1 \le i \le m$ are assumed to be bounded on $[0,1] \times \mathbb{R}^m$.

For this problem one can prove the following assertions ([21] and the references therein):

(a) For every $u \in L^2([0,1])$ there exists a unique absolutely continuous solution $y(u)$ to (1.2) on $[0,1]$. Moreover, the mapping $u \in \mathcal{B} \subset L^2([0,1])$

$\longmapsto y(u) \in C^0([0,1])^m$ is continuous. In addition, $y(\mathcal{B})$ is a bounded set in $C^0([0,1])^m$.

(b) The objective function f is continuously Fréchet differentiable. The gradient $g(u) \in L^2([0,1])$ of f is given by

$$g(u)(x) = \frac{\partial h_0}{\partial u}(x,y,u) + z(x)^T \frac{\partial h}{\partial u}(x,y,u), \tag{8.5}$$

where $y = y(u)$ is the solution of (1.2) and $z = z(u) : [0,1] \longrightarrow \mathbb{R}^m$ solves the adjoint equation

$$\begin{aligned}\frac{dz}{dx}(x) &= -\nabla_y h_0(x,y(x),u(x)) - \nabla_y h(x,y(x),u(x))z(x),\\ z(1) &= \nabla P(y(1)),\end{aligned}$$

with $y = y(u)$. Here we have used the notation $\nabla_y h = (\nabla_y h_1, \ldots, \nabla_y h_m)$. As for the state one can show that the adjoint state $z(u)$ is uniquely defined and absolutely continuous and that the mapping $u \in L^2([0,1]) \longmapsto z(u) \in C^0([0,1])^m$ is continuous. Furthermore, $z(\mathcal{B})$ is a bounded set in $C^0([0,1])^m$.

We can now verify our assumptions. The validity of (A1) with $\mathcal{D} = L^2([0,1])$ is stated in (b). From the form of the functions h_i, the properties of state and adjoint state, and the gradient representation (8.5) we see that the gradient is obtained by adding and multiplying L^∞-functions which are uniformly bounded for $u \in \mathcal{B} \subset L^\infty([0,1])$. This shows that assumptions (A2) and (A3) hold true. It is also possible to prove that the gradient g is uniformly continuous, i.e., (A6$'$) holds.

One can show that f is even twice continuously Fréchet differentiable with a Hessian representation of the form

$$\langle v, \nabla^2 f(u) w\rangle = \int_0^1 v(x)S(u)(x)w(x)\,dx + \int_0^1\int_0^1 v(x)K(u)(x,\xi)w(\xi)\,d\xi\,dx,$$

where $S(u) \in L^\infty([0,1])$ and $K(u) \in L^2([0,1]\times[0,1])$ are certain functions depending on the problem data. The lengthy expressions for S and K are omitted here. We refer to [21, pp. 535, 536].

A particular version of Algorithm 5.7 for this example can now be obtained as before. One uses a modification of the conjugate gradient algorithm similar to [10] to compute s_k in step 2.5. This will be discussed in detail in [23].

Similar results can be obtained for the more general case

$$h_i(x,y,u) = \sum_{0\le k\le p} h_i^k(x,y)u^k, \quad p \ge 2,$$

if $U = L^p([0,1])$ is used instead of $L^2([0,1])$, which underlines the practical importance of our L^p-framework.

9. Conclusions and future work.

We have introduced and analyzed a globally convergent class of interior-point trust-region algorithms for infinite-dimensional nonlinear optimization subject to pointwise bounds in function space. The methods are generalizations of those presented by Coleman and Li [6] for finite-dimensional problems. We have extended all first- and second-order global convergence results that are available for the finite-dimensional setting to our infinite-dimensional L^p-Banach space framework. The analysis was carried out in a unified way for $2 \le p \le \infty$. The lack of the equivalence of norms required the development of new proof techniques. This is also a valuable contribution to the finite-dimensional theory because our results are derived completely without using norm equivalences and hence are almost independent of the problem dimension. In this sense our convergence theory can be considered to be mesh independent. Moreover, we have carried out our analysis for a very general class of affine-scaling operators and almost arbitrary scaling of the trust-region. This is also new from the finite-dimensional viewpoint.

The work on the trust-region algorithm presented in this paper is continued in [23], where a local convergence analysis is developed. To establish local convergence the general choice of scalings w_k allowed in our global convergence analysis will be important. That paper also contains numerical results for optimal control problems governed by a nonlinear parabolic PDE which proves the efficiency of our algorithms. Another important development of [23] is the incorporation of a projection onto the box into the computation of approximate solutions of the trust-region subproblems.

The results of this paper and [23] represent a first important step toward a rigorous justification for why trust-region interior-point and trust-region interior-point SQP methods perform so well on discretized control problems. See [10], [13], [23] for applications. The extension of our theory to methods with additional equality constraints is in progress.

10. Appendix.

In this section we present proofs of Lemmas 6.4 and 6.5. These proofs require the following three technical results.

LEMMA 10.1. *For $0 < r \le 1$, $1 \le q \le \infty$, $v_1, v_2 \in L^q(\Omega)$, $v_1, v_2 \ge 0$, the following holds:*

$$\|v_1^r - v_2^r\|_q \le m_{q,q/r} \|v_1 - v_2\|_q^r. \tag{10.1}$$

Proof. For $r = 1$ the assertion is trivial. For $\alpha, \beta \ge 0$, $0 < r < 1$, we use the estimate

$$|\alpha^r - \beta^r| \le |\alpha - \beta|^r. \tag{10.2}$$

This estimate can be seen as follows. Due to symmetry we may assume that $\alpha \ge \beta \ge 0$. The function $h(\alpha) = |\alpha - \beta|^r - |\alpha^r - \beta^r|$ satisfies $h(\beta) = 0$,

$$h'(\alpha) = r\left((\alpha - \beta)^{r-1} - \alpha^{r-1}\right) \ge 0 \quad (\alpha > \beta)$$

and, thus, $h(\alpha) \ge 0$ for all $\alpha \ge \beta$.

In the case $q = \infty$ the assertion follows immediately from (10.2). For $1 \le q < \infty$ we use Lemma 3.2 to get

$$\begin{aligned} \|v_1^r - v_2^r\|_q \le m_{q,q/r} \|v_1^r - v_2^r\|_{q/r} &= m_{q,q/r} \left(\int_\Omega |v_1(x)^r - v_2(x)^r|^{q/r} \, dx \right)^{r/q} \\ &\le m_{q,q/r} \left(\int_\Omega |v_1(x) - v_2(x)|^q \, dx \right)^{r/q} = m_{q,q/r} \|v_1 - v_2\|_q^r. \end{aligned}$$

This completes the proof. □

LEMMA 10.2. *For $r \ge 1$, $1 \le q \le \infty$, $v_1, v_2 \in V$, $v_1, v_2 \ge 0$, the following inequality holds:*

$$\|v_1^r - v_2^r\|_q \le r \max\left\{\|v_1\|_\infty, \|v_2\|_\infty\right\}^{r-1} \|v_1 - v_2\|_q. \tag{10.3}$$

Proof. In the case $r = 1$ there is nothing to show. First we prove that for all $r > 1$, $\alpha, \beta \in [0, \gamma]$, $\gamma > 0$, we have $h(\alpha) \overset{\text{def}}{=} r\gamma^{r-1}|\alpha - \beta| - |\alpha^r - \beta^r| \geq 0$. In fact, we may assume $\alpha \geq \beta$ and compute $h(\beta) = 0$,

$$h'(\alpha) = r(\gamma^{r-1} - \alpha^{r-1}) \geq 0 \ \ (\beta \leq \alpha \leq \gamma).$$

Therefore,

$$|v_1^r(x) - v_2^r(x)| \leq r \max\left\{\|v_1\|_\infty, \|v_2\|_\infty\right\}^{r-1} |v_1(x) - v_2(x)| \ \text{ for all } x \in \Omega$$

which immediately implies (10.3). □

LEMMA 10.3. *Let $\alpha_1, \ldots, \alpha_n$, and $\beta_1, \ldots, \beta_n$ be arbitrary real numbers. Then*

$$|\min\{\alpha_1, \ldots, \alpha_n\} - \min\{\beta_1, \ldots, \beta_n\}| \leq \max\{|\alpha_1 - \beta_1|, \ldots, |\alpha_n - \beta_n|\}$$

Proof. Without restriction, let $\beta_k = \min\{\beta_1, \ldots, \beta_n\} \leq \min\{\alpha_1, \ldots, \alpha_n\}$. Then the assertion follows from

$$|\min\{\alpha_1, \ldots, \alpha_n\} - \min\{\beta_1, \ldots, \beta_n\}| = \min\{\alpha_1, \ldots, \alpha_n\} - \beta_k \leq \alpha_k - \beta_k. \qquad \square$$

10.1. Proof of Lemma 6.4. We write $\|\cdot\|_{q,A}$ for $\|\chi_A \cdot\|_q$, $A \subset \Omega$ measurable. For arbitrary $u, \tilde{u} \in \mathcal{B}$ set $N = \{x \in \Omega \ : \ g(u)(x)g(\tilde{u})(x) > 0\}$. The triangle inequality gives the estimate

$$\begin{aligned}
\|\hat{g}(u) - \hat{g}(\tilde{u})\|_{p'} &= \|d^r(u)g(u) - d^r(\tilde{u})g(\tilde{u})\|_{p'} \\
&\leq \|d(u)\|_\infty^r \|g(u) - g(\tilde{u})\|_{p'} + \|(d^r(u) - d^r(\tilde{u}))g(\tilde{u})\|_{p'} \\
&\leq \|d(u)\|_\infty^r \|g(u) - g(\tilde{u})\|_{p'} + \|d^r(u) - d^r(\tilde{u})\|_\infty \|g(\tilde{u})\|_{p',\Omega\setminus N} \\
&\quad + \|g(\tilde{u})\|_\infty \|d^r(u) - d^r(\tilde{u})\|_{p',N}.
\end{aligned}$$

We use the fact that $|g(u) - g(\tilde{u})| \geq |g(\tilde{u})|$ on $\Omega \setminus N$ and obtain

$$\begin{aligned}
\|\hat{g}(u) - \hat{g}(\tilde{u})\|_{p'} &\leq (\|d(u)\|_\infty^r + \|d^r(u) - d^r(\tilde{u})\|_\infty)\|g(u) - g(\tilde{u})\|_{p'} \\
&\quad + \|g(\tilde{u})\|_\infty \|d^r(u) - d^r(\tilde{u})\|_{p',N} \\
&\leq 3c_d^r \|g(u) - g(\tilde{u})\|_{p'} + c_1 \|d^r(u) - d^r(\tilde{u})\|_{p',N}.
\end{aligned}$$

Now the (uniform) continuity of $\hat{g}$ follows from Lemma 10.1, Lemma 10.2, the (uniform) continuity of g, and the assumption $\|\chi_N(d(u) - d(\tilde{u}))\|_{p'} \longrightarrow 0$ (uniformly in u) on the scaling. □

10.2. Proof of Lemma 6.5. We restrict ourselves to the more complicated case $d = d_{\text{II}}$. The result follows from Lemma 6.4 if we verify that

$$\|\chi_{\{g(u)g(\tilde{u})>0\}}(d(u) - d(\tilde{u}))\|_{p'} \longrightarrow 0 \qquad \text{as } \tilde{u} \longrightarrow u \text{ (uniformly in } u\text{)}.$$

Let $u, \tilde{u} \in \mathcal{B}$ be arbitrary. Using symmetries, it is easily seen that we are done if we are able to establish appropriate upper bounds for $|d(u)(x) - d(\tilde{u})(x)|$ for the three cases that $g(u)(x) > 0$, $g(\tilde{u})(x) > 0$, and

(a) $d_{\text{II}}(u)(x)$ and $d_{\text{II}}(\tilde{u})(x)$ are both determined by the second case in (4.4),

(b) $d_{\text{II}}(u)(x)$ and $d_{\text{II}}(\tilde{u})(x)$ are both determined by the else case in (4.4),

(c) $d_{\text{II}}(u)(x)$ is determined by the second and $d_{\text{II}}(\tilde{u})(x)$ by the else case in (4.4).

Set $\rho(x) = |d_{\mathrm{II}}(u)(x) - d_{\mathrm{II}}(\tilde{u})(x)|$. We will use Lemma 10.3 several times.

Case (a):

$$\rho(x) = |\min\{g(u)(x), c(x)\} - \min\{g(\tilde{u})(x), c(x)\}| \le |g(u)(x) - g(\tilde{u})(x)|.$$

Case (b):

$$\begin{aligned}\rho(x) &= |\min\{u(x) - a(x), b(x) - u(x), c(x)\} \\ &\quad - \min\{\tilde{u}(x) - a(x), b(x) - \tilde{u}(x), c(x)\}| \\ &\le |u(x) - \tilde{u}(x)|.\end{aligned}$$

Case (c): From $b(x) - u(x) \le u(x) - a(x)$ follows $u(x) - a(x) \ge c(x)$ and therefore

$$\begin{aligned}d_{\mathrm{II}}(u)(x) &= \min\{u(x) - a(x), g(u)(x), c(x)\} \\ &\ge \min\{u(x) - a(x), b(x) - u(x), c(x)\}.\end{aligned}$$

If $g(\tilde{u})(x) > b(x) - \tilde{u}(x)$ then $b(x) - \tilde{u}(x) > \tilde{u}(x) - a(x)$ and hence

$$d_{\mathrm{II}}(\tilde{u})(x) = \min\{\tilde{u}(x) - a(x), g(\tilde{u})(x), c(x)\}.$$

Therefore, we obtain

$$\rho(x) \le \max\{|u(x) - \tilde{u}(x)|, |g(u)(x) - g(\tilde{u})(x)|\}.$$

Otherwise, if $g(\tilde{u})(x) \le b(x) - \tilde{u}(x)$, we have in the case $d_{\mathrm{II}}(u)(x) \ge d_{\mathrm{II}}(\tilde{u})(x)$ that

$$\begin{aligned}\rho(x) &\le \min\{u(x) - a(x), g(u)(x), c(x)\} - \min\{\tilde{u}(x) - a(x), g(\tilde{u})(x), c(x)\} \\ &\le \max\{|u(x) - \tilde{u}(x)|, |g(u)(x) - g(\tilde{u})(x)|\},\end{aligned}$$

and for $d_{\mathrm{II}}(u)(x) < d_{\mathrm{II}}(\tilde{u})(x)$ we get

$$\begin{aligned}\rho(x) &\le \min\{\tilde{u}(x) - a(x), b(x) - \tilde{u}(x), c(x)\} \\ &\quad - \min\{u(x) - a(x), b(x) - u(x), c(x)\} \\ &\le |u(x) - \tilde{u}(x)|.\end{aligned}$$

Taking all cases together, this shows that

$$\begin{aligned}\|\chi_{\{g(u)g(\tilde{u})>0\}}\rho\|_{p'} &\le \|u - \tilde{u}\|_{p'} + \|g(u) - g(\tilde{u})\|_{p'} \\ &\le m_{p',p}\|u - \tilde{u}\|_{p} + \|g(u) - g(\tilde{u})\|_{p'}.\end{aligned}$$

Now, the application of Lemma 6.4 shows that $\hat{g}$ inherits the (uniform) continuity of g. □

Acknowledgments. This work was done while the first and second authors were visiting the Department of Computational and Applied Mathematics and the Center for Research on Parallel Computation, Rice University. They are greatly indebted to John Dennis for providing the opportunity to work in this excellent research environment.

The authors would like to thank Richard Byrd, Colorado State University, and Philippe Toint, Facultés Universitaires Notre-Dame de la Paix, for pointing us to the second-order convergence result in [19], which led to an improvement of generality and elegance in our presentation. The authors are also grateful to John Dennis, Rice University, and Luís Vicente, Universidade de Coimbra, for their helpful suggestions.

Finally, the authors would like to thank Philippe Toint and another referee for their critical reading and helpful comments that led to improvements in the presentation of the paper.

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] M. A. Branch, T. F. Coleman, and Y. Li, *A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems*, CTC95TR217, Center for Theory and Simulation in Science and Engineering, Cornell University, Ithaca, NY, 1995; also available online from http://www.tc.cornell.edu/Research/Tech.Reports/index.html.

[3] J. Burger and M. Pogu, *Functional and numerical solution of a control problem originating from heat transfer*, J. Optim. Theory Appl., 68 (1991), pp. 49–73.

[4] J. V. Burke, J. J. Moré, and G. Toraldo, *Convergence properties of trust region methods for linear and convex constraints*, Math. Programming, 47 (1990), pp. 305–336.

[5] T. F. Coleman and Y. Li, *On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds*, Math. Programming, 67 (1994), pp. 189–224.

[6] T. F. Coleman and Y. Li, *An interior trust region approach for nonlinear minimization subject to bounds*, SIAM J. Optim., 6 (1996), pp. 418–445.

[7] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Global convergence of a class of trust region algorithms for optimization with simple bounds*, SIAM J. Numer. Anal., 25 (1988), pp. 433–460. See [9].

[8] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Testing a class of methods for solving minimization problems with simple bounds on the variables*, Math. Comp., 50 (1988), pp. 399–430.

[9] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Correction to the paper on global convergence of a class of trust region algorithms for optimization with simple bounds*, SIAM J. Numer. Anal., 26 (1989), pp. 764–767.

[10] J. E. Dennis, M. Heinkenschloss, and L. N. Vicente, *Trust-region interior-point algorithms for a class of nonlinear programming problems*, SIAM J. Control Optim., 36 (1998), pp. 1750–1794.

[11] J. E. Dennis and L. N. Vicente, *Trust-region interior-point algorithms for minimization methods with simple bounds*, in Applied Mathematics and Parallel Computing, Festschrift for Klaus Ritter, H. Fischer, B. Riedmüller, and S. Schäffler, eds., Physica–Verlag, Heidelberg, 1996, pp. 97–107.

[12] J. C. Dunn, *On L^2 sufficient conditions and the gradient projection method for optimal control problems*, SIAM J. Control Optim., 34 (1996), pp. 1270–1290.

[13] M. Heinkenschloss and L. N. Vicente, *Analysis of Inexact Trust-Region Interior-Point SQP Algorithms*, Tech. report TR95–18, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1995; also available online from http://www.caam.rice.edu/~trice/trice_soft.html.

[14] C. T. Kelley and E. W. Sachs, *Multilevel algorithms for constrained compact fixed point problems*, SIAM J. Sci. Comput., 15 (1994), pp. 645–667.

[15] C. T. Kelley and E. W. Sachs, *A Trust Region Method for Parabolic Boundary Control Problems*, Tech. report CRSC–TR96–28, Center for Research in Scientific Computing, North Carolina State University, Raleigh, NC, 1996; also available online from http://www4.ncsu.edu/eos/users/ctkelley/www/pubs.html.

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

[17] J. J. Moré, *Recent developments in algorithms and software for trust region methods*, in Mathematical Programming, The State of The Art, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, Heidelberg, New York, 1983, pp. 258–287.

[18] E. Sachs, *A parabolic control problem with a boundary condition of the Stefan-Boltzmann type*, Z. Angew. Math. Mech., 58 (1978), pp. 443–449.

[19] G. A. Shultz, R. B. Schnabel, and R. H. Byrd, *A family of trust-region-based algorithms for unconstrained minimization with strong global convergence properties*, SIAM J. Numer. Anal., 22 (1985), pp. 47–67.

[20] D. C. Sorensen, *Newton's method with a model trust region modification*, SIAM J. Numer. Anal., 19 (1982), pp. 409–426.

[21] T. Tian and J. C. Dunn, *On the gradient projection method for optimal control problems with nonnegative $\mathcal{L}^2$ inputs*, SIAM J. Control Optim., 32 (1994), pp. 516–537.

[22] P. L. Toint, *Global convergence of a class of trust-region methods for nonconvex minimization in Hilbert space*, IMA J. Numer. Anal., 8 (1988), pp. 231–252.

[23] M. Ulbrich and S. Ulbrich, *Superlinear Convergence of Affine-Scaling Interior-Point Newton Methods for Infinite-Dimensional Nonlinear Problems with Pointwise Bounds*, TR97–05, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1997; also available online from http://www.statistik.tu-muenchen.de/LstAMS/sulbrich/papers/papers.html.

[24] L. N. Vicente, *Trust-Region Interior-Point Algorithms for a Class of Nonlinear Programming Problems*, Ph.D. thesis, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1996; also available online from http://www.mat.uc.pt/∼lvicente/papers/papers.html.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 3, pp. 765–776

A NEW PROJECTION METHOD FOR VARIATIONAL INEQUALITY PROBLEMS*

M. V. SOLODOV† AND B. F. SVAITER†

Abstract. We propose a new projection algorithm for solving the variational inequality problem, where the underlying function is continuous and satisfies a certain generalized monotonicity assumption (e.g., it can be pseudomonotone). The method is simple and admits a nice geometric interpretation. It consists of two steps. First, we construct an appropriate hyperplane which strictly separates the current iterate from the solutions of the problem. This procedure requires a single projection onto the feasible set and employs an Armijo-type linesearch along a feasible direction. Then the next iterate is obtained as the projection of the current iterate onto the intersection of the feasible set with the halfspace containing the solution set. Thus, in contrast with most other projection-type methods, only two projection operations per iteration are needed. The method is shown to be globally convergent to a solution of the variational inequality problem under minimal assumptions. Preliminary computational experience is also reported.

Key words. variational inequalities, projection methods, pseudomonotone maps

AMS subject classifications. 49M45, 90C25, 90C33

PII. S0363012997317475

1. Introduction.

We consider the classical variational inequality problem [1, 3, 7] VI(F, C), which is to find a point x^* such that

$$x^* \in C, \quad \langle F(x^*), x - x^* \rangle \geq 0 \quad \text{for all} \quad x \in C, \tag{1.1}$$

where C is a closed convex subset of $\Re^n$, $\langle \cdot, \cdot \rangle$ denotes the usual inner product in $\Re^n$, and $F : \Re^n \to \Re^n$ is a continuous function. Let S be the solution set of VI(F, C), which we assume to be nonempty. Let x^* be any element of the solution set S. We further assume that

$$\langle F(x), x - x^* \rangle \geq 0 \quad \text{for all} \quad x \in C. \tag{1.2}$$

It is clear that (1.2) is satisfied if $F(\cdot)$ is monotone; i.e.,

$$\langle F(x) - F(y), x - y \rangle \geq 0 \quad \text{for all} \quad x, y \in \Re^n.$$

More generally, (1.2) also holds if $F(\cdot)$ is pseudomonotone (as defined in [11]); i.e., for all $x, y \in \Re^n$

$$\langle F(y), x - y \rangle \geq 0 \quad \Longrightarrow \quad \langle F(x), x - y \rangle \geq 0.$$

Moreover, it is not difficult to construct examples where (1.2) is satisfied but $F(\cdot)$ is not monotone or pseudomonotone everywhere. Typically, condition (1.2) holds under some kind of generalized monotonicity assumptions on $F(\cdot)$, some of which are not difficult to check (see [26, 25]).

*Received by the editors February 28, 1997; accepted for publication (in revised form) June 3, 1998; published electronically March 2, 1999.

http://www.siam.org/journals/sicon/37-3/31747.html

†Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil (solodov@impa.br, benar@impa.br). The research of the first author was supported by CNPq grant 300734/95-6 and by PRONEX–Optimization. The research of the second author was supported by CNPq grant 301200/93-9(RN) and by PRONEX–Optimization.

In the case when $F(\cdot)$ is strongly monotone and/or the feasible set C has some special structure (e.g., C is the nonnegative orthant or, more generally, a box), there exist many efficient methods that can be used to solve those special cases of $\mathrm{VI}(F, C)$ (see [4, 5, 14, 18, 19, 20, 22, 23, 34, 36, 2, 16, 29, 15, 28, 27, 31]). In some of those methods, $F(\cdot)$ is further assumed to be differentiable, or Lipschitz continuous, or affine. Sometimes it is also assumed that the method starts close enough to the solution set (i.e., only local convergence is guaranteed). In the general case when $F(\cdot)$ and C do not possess any special structure, relatively few methods are applicable. In that case, projection-type algorithms are of particular relevance (we refer the reader to [32] for a more detailed discussion). The oldest algorithm of this class is the extragradient method proposed in [13] and later refined and extended in [10, 12, 17, 33, 9]. Some new projection-type algorithms that appear to be more efficient than the extragradient method were recently introduced in [32] (see also references therein).

In this paper, we are mainly concerned with the general case when the projection operator

$$P_C[x] := \arg\min_{y \in C} \|y - x\|$$

is computationally expensive (i.e., one has to solve an optimization problem to find a projection). Furthermore, we make no assumptions on the problem other than continuity of $F(\cdot)$ and condition (1.2). In this setting, one of the important tasks in devising efficient algorithms is to minimize the number of projection operations performed at each iteration. We note that in the case when $F(\cdot)$ is not Lipschitz continuous or the Lipschitz constant is not known, the extragradient method, as described in [12, 17, 10, 33], requires a linesearch procedure to compute the stepsize, with a new projection needed for each trial point. The same holds for the modified projection-type method in [32]. Clearly, this can be very computationally expensive. A novel idea to get around this inefficiency was proposed in [9] for the extragradient method. Here we will use this idea to devise a new projection algorithm that has even better properties, both theoretically and in our computational experience.

The algorithm proposed here allows a nice geometric interpretation, which is given in Figure 1.1 for its simplest version. Suppose we have x^i, a current approximation to the solution of $\mathrm{VI}(F, C)$. First, we compute the point $P_C[x^i - F(x^i)]$. Next, we search the line segment between x^i and $P_C[x^i - F(x^i)]$ for a point z^i such that the hyperplane $\partial H_i := \{x \in \Re^n \mid \langle F(z^i), x - z^i \rangle = 0\}$ strictly separates x^i from any solution x^* of the problem. A computationally inexpensive Armijo-type procedure is used to find such z^i. Once the hyperplane is constructed, the next iterate x^{i+1} is computed by projecting x^i onto the intersection of the feasible set C with the halfspace $H_i := \{x \in \Re^n \mid \langle F(z^i), x - z^i \rangle \leq 0\}$, which contains the solution set S. It can be seen that x^{i+1} thus computed is closer to any solution $x^* \in S$ than x^i. At each iteration, our algorithm uses one projection onto the set C (to construct the separating hyperplane H_i), and one projection onto the intersection $C \cap H_i$, which gives the next iterate.

Before proceeding, we emphasize the differences between the method of [9] and our Algorithms 2.1 and 2.2. First, the second projection step in our method is onto the intersection $C \cap H_i$. In [9], x^i is projected first onto the separating hyperplane ∂H_i (this point is denoted by $\bar{x}^i$ in Figure 1.1) and then onto C (in Figure 1.1, the resulting point is denoted by $P_C[\bar{x}^i]$). It can be verified that our iterate x^{i+1} is closer to the solution set S than the iterate computed by the method of [9]. We also avoid the extra work of computing the point $\bar{x}^i$. (Even though it can be carried out

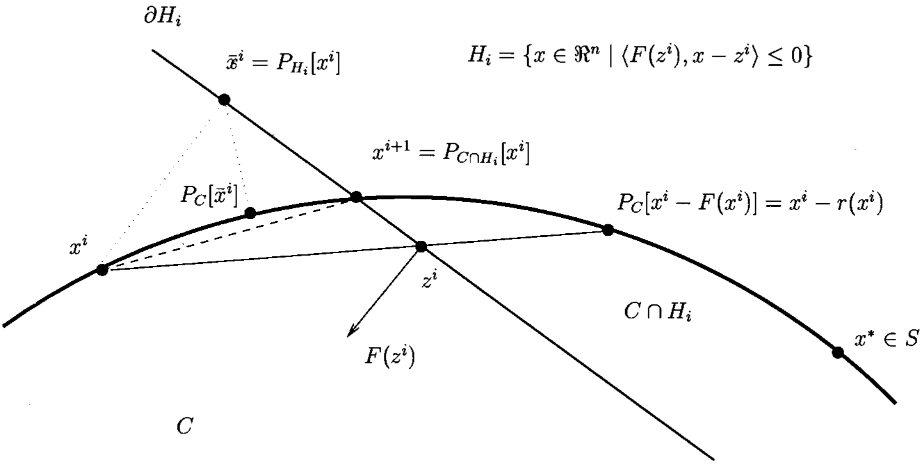

∂H_i
$\bar{x}^i = P_{H_i}[x^i]$
$H_i = \{x \in \Re^n \mid \langle F(z^i), x - z^i \rangle \leq 0\}$
$x^{i+1} = P_{C\cap H_i}[x^i]$
$P_C[\bar{x}^i]$
$P_C[x^i - F(x^i)] = x^i - r(x^i)$
x^i
z^i
$C \cap H_i$
$x^* \in S$
$F(z^i)$
C

FIG. 1.1. *The new projection method.*

explicitly, this is still some extra work.) Furthermore, the search direction in our method is not the same as in [9] for the following reasons. The search directions we use here are $P_C[x^i - \mu_i F(x^i)] - x^i$, where the stepsizes μ_i are chosen so that they are of the same order as the stepsizes η_i generating separating hyperplanes H_i (see Algorithm 2.2). The important point is that we allow *both* of them to go to zero if needed (but at the same rate). Coordination of the two stepsizes proves to be very significant in our computational experience (see section 3). We point out that [9] does not permit the stepsizes to go to zero in the first projection step even if the stepsizes generating the hyperplanes go to zero, and the proof there does not handle this case. The above-mentioned modifications seem to make a drastic difference in the numerical performance when our algorithm is compared to that of [9]. Our preliminary computational experience with the new algorithm is quite encouraging and is reported in section 3. However, we emphasize that comprehensive numerical study is not the primary focus of this paper.

Finally, our convergence results are stated under the assumption (1.2), which is considerably weaker than monotonicity of $F(\cdot)$ used in [9].

2. The algorithm and its convergence.

We first note that solutions of VI(F, C) coincide with zeros of the following projected residual function:

$$r(x) := x - P_C[x - F(x)];$$

i.e., $x \in S$ if and only if $r(x) = 0$.

We now formally state our algorithm described in section 1.

ALGORITHM 2.1. *Choose $x^0 \in C$ and two parameters $\gamma \in (0,1)$ and $\sigma \in (0,1)$.*

Having x^i, compute $r(x^i)$. If $r(x^i) = 0$, stop. Otherwise, compute

$$z^i = x^i - \eta_i r(x^i),$$

where $\eta_i = \gamma^{k_i}$, with k_i being the smallest nonnegative integer k satisfying

$$\langle F\left(x^i - \gamma^k r(x^i)\right), r(x^i)\rangle \geq \sigma \|r(x^i)\|^2. \tag{2.1}$$

Compute

$$x^{i+1} = P_{C \cap H_i}[x^i],$$

where

$$H_i = \{x \in \Re^n \mid \langle F(z^i), x - z^i \rangle \le 0\}.$$

The following well-known properties of the projection operator will be used below.

LEMMA 2.1. (see [35]). *Let B be any nonempty closed convex set in $\Re^n$. For any $x, y \in \Re^n$ and any $z \in B$ the following properties hold.*

1. $\langle x - P_B[x], z - P_B[x] \rangle \le 0$.
2. $\|P_B[x] - P_B[y]\|^2 \le \|x - y\|^2 - \|P_B[x] - x + y - P_B[y]\|^2$.

We start with a preliminary result. For now, we assume that the linesearch procedure in Algorithm 2.1 is well defined. This fact will be formally established in Theorem 2.1.

LEMMA 2.2. *Suppose that the linesearch procedure* (2.1) *of Algorithm* 2.1 *is well defined. Then it holds that*

$$x^{i+1} = P_{C \cap H_i}[\bar{x}^i],$$

where

$$\bar{x}^i = P_{H_i}[x^i].$$

Proof. Assuming that the point z^i is well defined, by (2.1) we have that

$$\langle F(z^i), x^i - z^i \rangle > 0.$$

It immediately follows that $x^i \notin H_i$. Also, by (1.2), $\langle F(z^i), x^* - z^i \rangle \le 0$ for any $x^* \in S$ because $z^i = (1 - \eta_i)x^i + \eta_i P_C[x^i - F(x^i)] \in C$ by the convexity of C. Therefore, $x^* \in H_i$. Since also $x^* \in C$, it follows that $C \cap H_i \ne \emptyset$. Because $C \cap H_i$ is a closed convex set that is nonempty, $x^{i+1} = P_{C \cap H_i}[x^i]$ is well defined.

It can be further verified that

$$\begin{aligned} \bar{x}^i = P_{H_i}[x^i] &= x^i - \frac{\langle F(z^i), x^i - z^i \rangle}{\|F(z^i)\|^2} F(z^i) \\ &= x^i - \frac{\eta_i \langle F(z^i), r(x^i) \rangle}{\|F(z^i)\|^2} F(z^i). \end{aligned}$$

Take any $y \in C \cap H_i$. Since $x^i \in C$ but $x^i \notin H_i$, there exist $\beta \in [0, 1]$ such that $\tilde{x} = \beta x^i + (1 - \beta) y \in C \cap \partial H_i$, where $\partial H_i := \{x \in \Re^n \mid \langle F(z^i), x - z^i \rangle = 0\}$. We have

$$\begin{aligned} \|y - \bar{x}^i\|^2 &\ge (1 - \beta)\|y - \bar{x}^i\|^2 \\ &= \|\tilde{x} - \beta x^i - (1 - \beta)\bar{x}^i\|^2 \\ &= \|\tilde{x} - \bar{x}^i\|^2 + \beta^2 \|x^i - \bar{x}^i\|^2 - 2\beta \langle \tilde{x} - \bar{x}^i, x^i - \bar{x}^i \rangle \\ &\ge \|\tilde{x} - \bar{x}^i\|^2, \end{aligned} \tag{2.2}$$

where the last inequality follows from Lemma 2.1 applied with $B = H_i$, $x = x^i$, and $z = \tilde{x} \in H_i$. Furthermore, we have

$$\begin{aligned} \|\tilde{x} - \bar{x}^i\|^2 &= \|\tilde{x} - x^i\|^2 - \|x^i - \bar{x}^i\|^2 \\ &\ge \|x^{i+1} - x^i\|^2 - \|x^i - \bar{x}^i\|^2 \\ &= \|x^{i+1} - \bar{x}^i\|^2, \end{aligned} \tag{2.3}$$

where the first equality is by $\bar{x}^i = P_{\partial H_i}[x^i]$, $\tilde{x} \in \partial H_i$, and Pythagoras's theorem; the inequality is by the fact that $\tilde{x} \in C \cap H_i$ and $x^{i+1} = P_{C\cap H_i}[x^i]$; and the last equality is again by Pythagoras's theorem. Combining (2.2) and (2.3), we obtain

$$\|y - \bar{x}^i\| \geq \|x^{i+1} - \bar{x}^i\| \quad \text{for all } y \in C \cap H_i.$$

Hence, $x^{i+1} = P_{C\cap H_i}[\bar{x}^i]$. □

We next prove our main convergence result.

THEOREM 2.1. *Let $F(\cdot)$ be continuous. Suppose the solution set S of VI(F, C) is nonempty and condition* (1.2) *is satisfied.*

Then any sequence $\{x^i\}$ generated by Algorithm 2.1 *converges to a solution of VI(F, C).*

Proof. We first show that the linesearch procedure in Algorithm 2.1 is well defined. If $r(x^i) = 0$, then the method terminates at a solution of the problem. Therefore, from now on, we assume that $\|r(x^i)\| > 0$. Also note that $x^i \in C$ for all i. Suppose that, for some i, (2.1) is not satisfied for any integer k, i.e., that

$$\langle F\left(x^i - \gamma^k r(x^i)\right), r(x^i)\rangle < \sigma\|r(x^i)\|^2 \quad \text{for all } k. \tag{2.4}$$

Applying Lemma 2.1 with $B = C,\ x = x^i - F(x^i),\ z = x^i \in C$, we obtain

$$\begin{aligned} 0 &\geq \langle x^i - F(x^i) - P_C[x^i - F(x^i)], x^i - P_C[x^i - F(x^i)]\rangle \\ &= \|r(x^i)\|^2 - \langle F(x^i), r(x^i)\rangle. \end{aligned}$$

Hence,

$$\langle F(x^i), r(x^i)\rangle \geq \|r(x^i)\|^2. \tag{2.5}$$

Since $x^i - \gamma^k r(x^i) \to x^i$ as $k \to \infty$, and $F(\cdot)$ is continuous, passing onto the limit as $k \to \infty$ in (2.4), we obtain

$$\langle F(x^i), r(x^i)\rangle \leq \sigma\|r(x^i)\|^2.$$

But the latter relation contradicts (2.5) because $\sigma < 1$ and $\|r(x^i)\| > 0$. Hence (2.1) is satisfied for some integer k_i.

Thus the linesearch step is well defined, and by Lemma 2.2 we know that the rest of the method is as well. In particular, $x^{i+1} = P_{C\cap H_i}[\bar{x}^i]$, where $\bar{x}^i = P_{H_i}[x^i]$. By Lemma 2.1 applied with $B = C \cap H_i,\ x = \bar{x}^i$, and $z = x^* \in S \subset C \cap H_i$, we obtain

$$\begin{aligned} 0 &\geq \langle \bar{x}^i - x^{i+1}, x^* - x^{i+1}\rangle \\ &= \|x^{i+1} - \bar{x}^i\|^2 + \langle \bar{x}^i - x^{i+1}, x^* - \bar{x}^i\rangle. \end{aligned}$$

Hence,

$$\langle x^* - \bar{x}^i, x^{i+1} - \bar{x}^i\rangle \geq \|x^{i+1} - \bar{x}^i\|^2.$$

Therefore,

$$\begin{aligned} \|x^{i+1} - x^*\|^2 &= \|\bar{x}^i - x^*\|^2 + \|x^{i+1} - \bar{x}^i\|^2 + 2\langle \bar{x}^i - x^*, x^{i+1} - \bar{x}^i\rangle \\ &\leq \|\bar{x}^i - x^*\|^2 - \|x^{i+1} - \bar{x}^i\|^2 \\ &= \|x^i - x^*\|^2 - \|x^{i+1} - \bar{x}^i\|^2 \end{aligned}$$

$$
\begin{aligned}
&+\left(\frac{\eta_i\langle F(z^i),r(x^i)\rangle}{\|F(z^i)\|}\right)^2-\frac{2\eta_i\langle F(z^i),r(x^i)\rangle}{\|F(z^i)\|^2}\langle F(z^i),x^i-x^*\rangle\\
&=\|x^i-x^*\|^2-\|x^{i+1}-\bar{x}^i\|^2-\left(\frac{\eta_i\langle F(z^i),r(x^i)\rangle}{\|F(z^i)\|}\right)^2\\
&\quad-\frac{2\eta_i\langle F(z^i),r(x^i)\rangle}{\|F(z^i)\|^2}\langle F(z^i),z^i-x^*\rangle\\
&\le\|x^i-x^*\|^2-\|x^{i+1}-\bar{x}^i\|^2-\left(\frac{\eta_i\sigma}{\|F(z^i)\|}\right)^2\|r(x^i)\|^4, \qquad (2.6)
\end{aligned}
$$

where the last inequality follows from (2.1) and (1.2).

Now (2.6) implies that the sequence $\{\|x^i-x^*\|\}$ is nonincreasing. Therefore, it converges. We further deduce that the sequence $\{x^i\}$ is bounded, and so is $\{z^i\}$. Thus there exists a constant $M>0$ such that $\|F(z^i)\|\le M$ for all i. Hence, from (2.6),

$$(2.7)\qquad \|x^{i+1}-x^*\|^2\le\|x^i-x^*\|^2-\|x^{i+1}-\bar{x}^i\|^2-(\sigma/M)^2\eta_i^2\|r(x^i)\|^4.$$

From convergence of $\{\|x^i-x^*\|\}$, it follows from (2.7) that

$$(2.8)\qquad \lim_{i\to\infty}\eta_i\|r(x^i)\|=0.$$

We consider the two possible cases. Suppose first that $\limsup_{i\to\infty}\eta_i>0$. For (2.8) to hold it must then be the case that $\liminf_{i\to\infty}\|r(x^i)\|=0$. Since $r(\cdot)$ is continuous and $\{x^i\}$ is bounded, there exists $\hat{x}$, an accumulation point of $\{x^i\}$, such that $r(\hat{x})=0$. It follows that $\hat{x}\in S$ and we can take $x^*=\hat{x}$ in the preceding arguments and, in particular, in (2.7). Thus the sequence $\{\|x^i-\hat{x}\|\}$ converges. Since $\hat{x}$ is an accumulation point of $\{x^i\}$, it easily follows that $\{\|x^i-\hat{x}\|\}$ converges to zero, i.e., that $\{x^i\}$ converges to $\hat{x}\in S$.

Suppose now that $\lim_{i\to\infty}\eta_i=0$. By the choice of η_i we know that (2.1) was not satisfied for k_i-1 (at least for i large enough, so that $\eta_i<1$); i.e.,

$$(2.9)\qquad \langle F\left(x^i-\gamma^{-1}\eta_i r(x^i)\right),r(x^i)\rangle<\sigma\|r(x^i)\|^2\quad\text{for all}\quad i\ge i_0.$$

Let $\hat{x}$ be any accumulation point of $\{x^i\}$ and $\{x^{i_j}\}$ be the corresponding subsequence converging to $\hat{x}$. Passing onto the limit in (2.9) along this subsequence, and using (2.5), we obtain

$$\sigma\|r(\hat{x})\|^2\ge\langle F(\hat{x}),r(\hat{x})\rangle\ge\|r(\hat{x})\|^2,$$

implying that $r(\hat{x})=0$, i.e., that $\hat{x}\in S$. Setting $x^*=\hat{x}$ in (2.7) and repeating the previous arguments, we conclude that the whole sequence $\{x^i\}$ converges to $\hat{x}\in S$. This completes the proof. □

We next propose a modification of Algorithm 2.1 that is motivated by our computational experience and appears to be more practical. For $\mu>0$, define

$$r(x,\mu):=x-P_C[x-\mu F(x)].$$

With this definition, we have $r(x,1)=r(x)$.

The idea is to use, for the first projection step at the current iteration, the stepsize that is not too different from the stepsize computed at the previous iteration (a similar technique was also used in [32, Algorithm 3.2]). Note that Algorithm 2.2 has a certain coordination between the stepsizes μ_i in the first projection step and η_i in the step computing the separating hyperplane. For example, both of them can go to zero if needed. This is in contrast with the method of [9], where the stepsize in the first projection step can never go to zero, even when η_i does. We found this coordination mechanism important in our computational results, reported in section 3. Note also that both stepsizes can increase from one iteration to the next.

ALGORITHM 2.2. *Choose* $x^0 \in C$, $\eta_{-1} > 0$, *and three parameters* $\gamma \in (0,1)$, $\sigma \in (0,1)$, *and* $\theta > 1$.

Having x^i, *compute* $r(x^i, \mu_i)$, *where* $\mu_i := \min\{\theta\eta_{i-1}, 1\}$. *If* $r(x^i, \mu_i) = 0$, *stop. Otherwise, compute*

$$z^i = x^i - \eta_i r(x^i, \mu_i),$$

where $\eta_i = \gamma^{k_i}\mu_i$ *with* k_i *being the smallest nonnegative integer* k *satisfying*

$$\langle F\left(x^i - \gamma^k \mu_i r(x^i, \mu_i)\right), r(x^i, \mu_i)\rangle \geq \frac{\sigma}{\mu_i}\|r(x^i, \mu_i)\|^2.$$

Compute

$$x^{i+1} = P_{C\cap H_i}[x^i],$$

where

$$H_i = \{x \in \Re^n \mid \langle F(z^i), x - z^i\rangle \leq 0\}.$$

THEOREM 2.2. *Let* $F(\cdot)$ *be continuous. Suppose that the solution set* S *of* VI(F, C) *is nonempty and* (1.2) *is satisfied.*

Then any sequence $\{x^i\}$ *generated by Algorithm* 2.2 *converges to a solution of* VI(F, C).

Proof. The proof of convergence uses the same ideas as the proof for Algorithm 2.1, so we supply only a sketch.

As with (2.5), it can be established that

$$\langle F(x^i), r(x^i, \mu_i)\rangle \geq \frac{1}{\mu_i}\|r(x^i, \mu_i)\|^2.$$

In particular, it follows that the linesearch procedure is well defined.

The proof then follows the pattern of the proof of Theorem 2.1, with (2.7) replaced by

$$\|x^{i+1} - x^*\|^2 \leq \|x^i - x^*\|^2 - \|x^{i+1} - \bar{x}^i\|^2 - (\sigma\eta_i/M)^2\mu_i^{-2}\|r(x^i, \mu_i)\|^4.$$

We next use the fact (see [6, Lemma 1]) that

$$\|r(x^i, \mu_i)\| \geq \min\{1, \mu_i\}\|r(x^i)\|.$$

It follows that

$$\|x^{i+1} - x^*\|^2 \leq \|x^i - x^*\|^2 - \|x^{i+1} - \bar{x}^i\|^2 - (\sigma\eta_i/M)^2\mu_i^2\|r(x^i)\|^4.$$

Taking into account that $\eta_i = \gamma^{k_i}\mu_i \leq \mu_i$, we further obtain

$$\|x^{i+1} - x^*\|^2 \leq \|x^i - x^*\|^2 - \|x^{i+1} - \bar{x}^i\|^2 - (\sigma/M)^2\eta_i^4\|r(x^i)\|^4,$$

and the rest of the convergence proof is identical to that of Theorem 2.1. □

3. Computational experience. To give some insight into the behavior of the new projection algorithm (Algorithm 2.2), we implemented it in MATLAB to solve linearly constrained variational inequality problems (using the quadratic-program solver qp.m from the MATLAB optimization toolbox to perform the projection). For a benchmark, we compared the performance of this implementation with analogous implementations of two versions of the extragradient method (as described in [17] and [33]) and with the modified projection algorithm as given in [32, Algorithm 3.2]. We also implemented the algorithm of [9] and tested it on the same problems as the other four methods. We do not report here the full results for the method of [9], mainly because they were rather poor. In particular, like the extragradient method, the method of [9] failed on the first two problems, and was *by far* the worst among all the methods on the remaining test problems. Thus we found it to be not useful for a benchmark comparison (unfortunately, no computational experience was reported in [9]). By contrast, our algorithm seems to perform better than the alternatives in most cases.

The choice of linearly constrained variational inequalities for our experiments is not incidental. It is clear that the new method should be especially effective when feasible sets are "no simpler" than general polyhedra (so that an optimization problem has to be solved to find a projection). In that case, adding one more linear constraint to perform a projection onto $C \cap H_i$ does not increase the cost compared to projecting onto the feasible set C. Actually, if the constraints are nonlinear, projecting onto $C \cap H_i$ can sometimes turn easier than onto C. On the other hand, when C has some special structure (for example, C is a box), adding a linear constraint would require solving an optimization problem, while projecting onto C can be carried out explicitly. In that case, the extragradient methods may be more attractive than our new method. However, in the case when C is a box (or the nonnegative orthant), there are many other efficient methods available (as discussed in section 1), so we will not focus on this case.

Though our experience is limited in scope, it suggests that the new projection method is a valuable alternative to the extragradient [17, 33] and modified projection [32] methods. We describe the test details below.

All MATLAB codes were run on the Sun UltraSPARCstation 1 under MATLAB version 5.0.0.4064. Our test problems are the same as those used in [32] to test the modified projection method in the nonlinear case. The first test problem, used first by Mathiesen [21], and later in [24, 36], has

$$F(x_1, x_2, x_3) = \begin{bmatrix} .9(5x_2 + 3x_3)/x_1 \\ .1(5x_2 + 3x_3)/x_2 - 5 \\ -3 \end{bmatrix},$$

$$C = \left\{(x_1, x_2, x_3) \in \Re^3_+ \mid x_1 + x_2 + x_3 = 1, x_1 - x_2 - x_3 \le 0\right\}.$$

For the other test problems, the feasible set is the simplex

$$C = \{x \in \Re^n_+ \mid x_1 + \cdots + x_n = n\},$$

and $F(\cdot)$ and n are specified as follows. For the third to fifth problems, $F(\cdot)$ is the function from, respectively, the Kojima–Shindo Nonlinear Complementarity Problem (NCP) (with $n = 4$) and the Nash–Cournot NCP (with $n = 5$ and $n = 10$) [24, pp. 321–322]. For the sixth problem, $n = 20$ and $F(\cdot)$ is affine, i.e.,

$$F(x) = Mx + q,$$

with the matrix M randomly generated as suggested in [8]:

$$M = AA^\top + B + D,$$

where every entry of the $n \times n$ matrix A and of the $n \times n$ skew-symmetric matrix B is uniformly generated from $(-5, 5)$, and every diagonal entry of the $n \times n$ diagonal B is uniformly generated from $(0, 0.3)$ (so M is positive definite), with every entry of q uniformly generated from $(-500, 0)$. For the last problem, we took the F from the sixth problem and added to its ith component the linear-quadratic term $\max\{0, x_i\}^2$ for $i = 1, \dots, \lfloor n/2 \rfloor$.

In the implementation of our Algorithm 2.2, we choose $\sigma = .3$, $\eta_{-1} = 1$, $\gamma = .5$, and $\theta = 4$. Implementations of the modified projection method and the extragradient method of [17] are the same as those reported in [32]. In particular, the parameters for both algorithms were tuned to optimize the performance. In the implementation of the other version of the extragradient method [33, Algorithm C], we set parameters as follows: $\eta = .1$, $\alpha = .3$, and $\gamma = 1.7$. On the Mathiesen problem, we used the same x^0 as in [36]; on the other problems, we used $x^0 = (1, \dots, 1)$. (The $F(\cdot)$ from the Mathiesen problem and from the Nash–Cournot NCP are defined on the positive orthant only.) The test results are summarized in Tables 3.1 and 3.2. In most cases, Algorithm 2.2 requires fewer iterations, function evaluations, and projections, and takes less CPU time than the other methods considered. Also note that our method solves problems, such as the Kojima–Shindo problem, even when $F(\cdot)$ is not monotone. We caution, however, that this study is very preliminary.

To investigate the effect of different starting points, we tested the methods on the problem Nash5, using as starting points the mesh points of the uniform triangulation of the simplex

$$\left\{ x \in \Re^5 \mid x \geq 0, \ \sum_{j=1}^{5} x_j = 5 \right\}.$$

The triangulation is obtained by cutting the simplex by equally spaced hyperplanes parallel to the faces of the simplex, with four cuts per face (the first cut is the face of the simplex, and the last is the opposing vertex). We have chosen the problem Nash5 because its function is defined on the positive orthant only, so that the *boundary effect* can also be studied. The modified projection method of [32] had trouble when starting at points with many zero components (such as, e.g., the point $(5, 0, 0, 0, 0)$). This is not very surprising, since the modified projection is actually an *infeasible* method; i.e., the iterates generated need not belong to the feasible set. Therefore, in principle, we may need to evaluate F at points outside of the nonnegative orthant, which gives trouble for problems like Nash5.However, the modified projection algorithm managed to solve the problem in most cases, with the average of 95 iterations (note that this is quite a bit more than the 74 iterations reported in Table 3.1 as needed for the starting point $x^0 = (1, 1, 1, 1, 1)$). In general, the more positive components x^0 has, the fewer iterations the modified projection method needs to solve the problem. The extragradient method of Marcotte [17] managed to solve the problem for all starting points, with the average of 62 iterations (which is again considerably higher than the 43 iterations needed when starting at the unit vector). Finally, Algorithm 2.2 managed to solve the problem from all the starting points, except from a few points for which the second and fourth components are *both* zero. On average, 32 iterations were required for convergence, which is fewer than for the modified projection or the

TABLE 3.1

Results for Algorithm 2.2 *and the modified projection method on linearly constrained variational inequality problems.*

		Algorithm 2.2[1]		**Modified Projection [32]** [2]	
Name	n	**iter.**(nf/np)[3]	CPU	**iter.**(nf/np)[3]	CPU
Mathiesen	3	14(53/28)	0.5	30(68/38)	0.9
		14(55/28)	0.5	25(56/31)	0.7
KojimaSh	4	7(16/14)	0.3	38(84/46)	0.7
Nash5	5	24(100/48)	1.5	74(155/81)	1.9
Nash10	10	34(140/68)	3	93(192/99)	2.7
HPHard	20	379(1520/758)	196	692(1391/699)	128
qHPHard	20	317(1272/634)	154	562(1131/569)	98

[1] Algorithm 2.2 with $\sigma = .3$, $\eta_{-1} = 1$, $\gamma = .5$, and $\theta = 4$.

[2] Modified projection method as described in [32, Algorithm 3.2] and parameters set as reported in that reference ($P = I$, $\alpha_{-1} = 1$, $\theta = 1.5$, $\rho = .1$, and $\beta = .3$).

[3] For all methods, the termination criterion is $\|r(x)\| \leq 10^{-4}$. (nf denotes the total number of times $F(\cdot)$ is evaluated and np denotes the total number of times a projection is performed.) CPU denotes time (in seconds) obtained using the intrinsic MATLAB function etime and with the codes run on a Sun UltraSPARCstation 1; does not include time to read problem data. On the Mathiesen problem, we ran each method twice with $x^0 = (.1, .8, .1)$ and $x^0 = (.4, .3, .3)$ respectively; on the other problems, we used $x^0 = (1, \ldots, 1)$.

TABLE 3.2

Results for the two versions of extragradient method on linearly constrained variational inequality problems.

		Extragradient [17][4]		**Extragradient [33]**[5]	
Name	n	**iter.**(nf/np)[3]	CPU	**iter.**(nf/np)[3]	CPU
Mathiesen	3	—	—	—	—
		—	—	—	—
KojimaSh	4	16(36/36)	0.5	78(157/79)	2.5
Nash5	5	43(89/89)	1.8	92(184/93)	2.8
Nash10	10	84(172/172)	3.4	103(191/172)	5.5
HPHard	20	532(1067/1067)	163	1003(2607/1607)	562
qHPHard	20	461(926/925)	162	892(2536/1536)	503

[4] The extragradient method as described in [17], with $\beta = .7$ and initial $\alpha = 1$. Dash—indicates that the method did not converge.

[5] The extragradient method as described in [33, Algorithm C], with $\eta = .1$, $\alpha = .3$, $\gamma = 1.7$. Dash—indicates that the method did not converge.

extragradient methods. However, the extragradient method appeared to be somewhat more robust for this problem.

4. Concluding remarks. A new projection algorithm for solving variational inequality problems was proposed. Under minimal assumptions of continuity of the underlying function and generalized monotonicity (for example, pseudomonotonicity), it was established that the iterates converge to a solution of the problem. The new method has some clear theoretical advantages over most of existing projection methods for general variational inequality problems with no special structure. Preliminary computational experience is also encouraging.

Some of the projection ideas presented here also proved to be useful in devising *truly* globally convergent (i.e., the *whole* sequence of iterates is globally convergent to a solution without any regularity assumptions) and locally superlinearly convergent inexact Newton methods for solving systems of monotone equations [30] and monotone NCPs [31].

Acknowledgments. We are grateful to the two anonymous referees and the editor for constructive suggestions that led to improvements in the paper. In particular, the observation that pseudomonotonicity is sufficient to establish convergence of the method is due to a referee.

REFERENCES

[1] A. A. Auslender, *Optimisation Méthodes Numériques*, Masson, Paris, 1976.

[2] J. F. Bonnans, *Local analysis of Newton-type methods for variational inequalities and nonlinear programming*, Appl. Math. Optim., 29 (1994), pp. 161–186.

[3] R. W. Cottle, F. Giannessi, and J.-L. Lions, *Variational Inequalities and Complementarity Problems: Theory and Applications*, Wiley, New York, 1980.

[4] S. C. Dafermos, *An iterative scheme for variational inequalities*, Math. Programming, 26 (1983), pp. 40–47.

[5] M. Fukushima, *Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems*, Math. Programming, 53 (1992), pp. 99–110.

[6] E. M. Gafni and D. P. Bertsekas, *Two-metric projection methods for constrained optimization*, SIAM J. Control Optim., 22 (1984), pp. 936–964.

[7] R. Glowinski, J.-L. Lions, and R. Trémolières, *Numerical Analysis of Variational Inequalities*, North-Holland, Amsterdam, 1981.

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

[9] A. N. Iusem and B. F. Svaiter, *A variant of Korpelevich's method for variational inequalities with a new search strategy*, Optimization, 42 (1997), pp. 309–321.

[10] A. N. Iusem, *An iterative algorithm for the variational inequality problem*, Comput. Appl. Math., 13 (1994), pp. 103–114.

[11] S. Karamardian, *Complementarity problems over cones with monotone and pseudomonotone maps*, J. Optim. Theory Appl., 18 (1976), pp. 445–455.

[12] E. N. Khobotov, *A modification of the extragradient method for the solution of variational inequalities and some optimization problems*, USSR Comput. Math. Math. Phys., 27 (1987), pp. 1462–1473.

[13] G. M. Korpelevich, *The extragradient method for finding saddle points and other problems*, Matecon, 12 (1976), pp. 747–756.

[14] T. L. Magnanti and G. Perakis, *On the Convergence of Classical Variational Inequality Algorithms*, Working Paper, Operations Research Center, MIT, Cambridge, MA, May 1993.

[15] O. L Mangasarian and M. V. Solodov, *A linearly convergent derivative-free descent method for strongly monotone complementarity problems*, Comput. Optim. Appl., to appear.

[16] O. L Mangasarian and M. V. Solodov, *Nonlinear complementarity as unconstrained and constrained minimization*, Math. Programming, 62 (1993), pp. 277–297.

[17] P. Marcotte, *Application of Khobotov's algorithm to variational inequalities and network equilibrium problems*, Inform. Systems Oper. Res., 29 (1991), pp. 258–270.

[18] P. Marcotte and J.-P. Dussault, *A note on a globally convergent Newton method for solving monotone variational inequalities*, Oper. Res. Lett., 6 (1987), pp. 35–42.

[19] P. Marcotte and J.-P. Dussault, *A sequential linear programming algorithm for solving monotone variational inequalities*, SIAM J. Control Optim., 27 (1989), pp. 1260–1278.

[20] P. Marcotte and J.-H. Wu, *On the convergence of projection methods: Application to the decomposition of affine variational inequalities*, J. Optim. Theory Appl., 85 (1995), pp. 347–362.

[21] L. Mathiesen, *An algorithm based on a sequence of linear complementarity problems applied to a Walrasian equilibrium model: An example*, Math. Programming, 37 (1987), pp. 1–18.

[22] J.-S. Pang, *Asymmetric variational inequality problems over product sets: Applications and iterative methods*, Math. Programming, 31 (1985), pp. 206–219.

[23] J.-S. Pang and D. Chan, *Iterative methods for variational and complementarity problems*, Math. Programming, 24 (1982), pp. 284–313.

[24] J.-S. Pang and S. A. Gabriel, *NE/SQP: A robust algorithm for the nonlinear complementarity problem*, Math. Programming, 60 (1993), pp. 295–337.

[25] R. Pini and C. Singh, *A survey of recent* (1985–1995) *advances in generalized convexity with applications to duality theory and optimality conditions*, Optimization, 39 (1997), pp. 311–

360.

[26] S. Schaible, S. Karamardian, and J.-P. Crouzeix, *Characterizations of generalized monotone maps*, J. Optim. Theory Appl., 76 (1993), pp. 399–413.

[27] M. V. Solodov, *Implicit Lagrangian*, in Encyclopedia of Optimization, C. Floudas and P. Pardalos, eds., Kluwer Academic Publishers, Norwell, MA, 1999.

[28] M. V. Solodov, *Some optimization reformulations of the extended linear complementarity problem*, Comput. Optim. Appl., to appear.

[29] M. V. Solodov, *Stationary points of bound constrained reformulations of complementarity problems*, J. Optim. Theory Appl., 94 (1997), pp. 449–467.

[30] M. V. Solodov and B. F. Svaiter, *A globally convergent inexact Newton method for systems of monotone equations*, in Reformulation—Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Norwell, MA, 1998, pp. 355–369.

[31] M. V. Solodov and B. F. Svaiter, *A truly globally convergent Newton-type method for the monotone nonlinear complementarity problem*, SIAM J. Optim., submitted.

[32] M. V. Solodov and P. Tseng, *Modified projection-type methods for monotone variational inequalities*, SIAM J. Control Optim., 34 (1996), pp. 1814–1830.

[33] D. Sun, *A new step-size skill for solving a class of nonlinear projection equations*, J. Comput. Math., 13 (1995), pp. 357–368.

[34] P. Tseng, *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities*, SIAM J. Control Optim., 29 (1991), pp. 119–138.

[35] E. H. Zarantonello, *Projections on convex sets in Hilbert space and spectral theory*, in Contributions to Nonlinear Functional Analysis, E. Zarantonello, ed., Academic Press, New York, 1971, pp. 237–424.

[36] L. Zhao and S. Dafermos, *General economic equilibrium and variational inequalities*, Oper. Res. Lett., 10 (1991), pp. 369–376.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 3, pp. 777–803

CONTINUOUS DEPENDENCE WITH RESPECT TO THE INPUT OF TRAJECTORIES OF CONTROL-AFFINE SYSTEMS[*]

WENSHENG LIU[†] AND HÉCTOR J. SUSSMANN[†]

Abstract. We study the continuous dependence on the input of trajectories of control-affine systems belonging to the class $C^0(m)$ of all systems Σ of the form

$$\Sigma : \qquad \dot{x} = f_0(x) + \sum_{i=1}^{m} u_i(t) f_i(x),$$

where $f_0, \ldots, f_m$ are continuous vector fields on some open subset of $\mathbb{R}^n$ and the control functions belong to $L^1([0,T],\mathbb{R}^m)$. We give a simple necessary and sufficient condition for a control sequence $\{u^j\}_{j=1}^{\infty}$ to "$\mathcal{T}^0$-converge" to a control u^∞, i.e., to be such that, for every system Σ in $C^0(m)$, the trajectories generated by the u^j converge as $j \to \infty$ to the trajectories generated by u^∞. We also characterize $\mathcal{T}^k$-convergence (the concept of control convergence that arises when we use, instead of $C^0(m)$, the class $C^k(m)$ of systems Σ where the f_i are of class C^k) for $k \geq 1$ in the scalar input case, and we explain how the analogous characterization for the multi-input case fails to be true, unless one restricts oneself to the class $C^k_{comm}(m)$ of systems for which the vector fields $f_1, \ldots, f_m$ commute. As a preliminary, we define a "topology of trajectory convergence" (or "T-convergence") on the set of all time-varying vector fields $\Omega \times I \ni (x,t) \mapsto f(x,t) \in \mathbb{R}^n$, where Ω is an open subset of $\mathbb{R}^n$ and I is an interval, and we study some of its properties. This enables us to make the definition of $\mathcal{T}^k$-convergence precise for sequences and, more generally, for nets, by saying that a net $\{u^\alpha\}_{\alpha\in A}$ in $L^1([0,T],\mathbb{R}^m)$ $\mathcal{T}^k$-converges to a limit u^∞ if for every system Σ in $C^k(m)$ the time-varying vector fields $(x,t) \mapsto f_0(x) + \sum_{i=1}^m u_i^\alpha(t) f_i(x)$ $\mathcal{T}^k$-converge to $(x,t) \mapsto f_0(x) + \sum_{i=1}^m u_i^\infty(t) f_i(x)$.

Key words. control-affine systems, continuous dependence

AMS subject classifications. 34A10, 34H05, 94C10, 94C15

PII. S0363012996237378

1. Introduction.

In the control theory literature, the property of continuous dependence of the trajectories with respect to the control has received wide attention (cf. [1], [2], [3], [5], [7], [9], [12], [19], [20]) because of its fundamental role in proofs of closedness of reachable sets, existence of optimal controls, and continuity or lower semicontinuity of value functions. It is clear that continuous dependence always holds for trivial reasons if the space of input functions is given a sufficiently strong topology, but continuous dependence results are useful only when the topology involved is reasonably weak. (For example, many existence theorems for optimal controls depend on the fact that the space of inputs is compact, or has many compact subsets.) Thus it is of interest to determine concepts of convergence of inputs that lead to continuous dependence of trajectories and are as weak as possible.

In this paper we characterize the weakest possible concept of convergence of a sequence of input functions such that the trajectories depend continuously on the input for *all* control-affine systems of the form

$$\dot{x} = f_0(x) + \sum_{i=1}^{m} u_i f_i(x)\,, \quad x \in \Omega\,, \tag{1.1}$$

[*] Received by the editors November 18, 1996; accepted for publication (in revised form) April 8, 1998; published electronically March 2, 1999. This research was supported in part by NSF grant DMS95-00798 and AFOSR grant 0923.

http://www.siam.org/journals/sicon/37-3/23737.html

[†] Department of Mathematics, Rutgers University, Piscataway, NJ 08854-8019 (wliu@hilbert.rutgers.edu, sussmann@math.rutgers.edu).

with a right-hand side continuous with respect to the state variable x, and we prove some results on the corresponding problem for more restricted classes of systems. These questions are in fact special cases of the following more general situation: we are given

(a) a set $\mathcal{U}$ of input functions $t \mapsto u(t)$ defined on an interval I,
(b) a class $\mathcal{C}$ of control systems Σ of the form $\dot{x} = g(x, u)$, all of which admit the members of $\mathcal{U}$ as inputs.

We then seek to characterize the *weakest* topology $\mathcal{T}$ on $\mathcal{U}$ with the property that, if $\mathcal{U}$ is equipped with $\mathcal{T}$, then

(CD) the trajectories of Σ depend continuously on the input $u \in \mathcal{U}$ for all systems Σ in $\mathcal{C}$.

Let us write $\mathcal{T}(\mathcal{U}, \mathcal{C})$ to denote this topology, and assume temporarily that the definition of $\mathcal{T}(\mathcal{U}, \mathcal{C})$ has been justified, by

(J1) assigning a mathematically precise meaning to (CD), and
(J2) proving the existence of a weakest topology on $\mathcal{U}$ for which (CD) holds.

Then the problems to be studied here are those of characterizing the topology $\mathcal{T}(\mathcal{U}, \mathcal{C})$ in two cases, namely, (i) when $\mathcal{C}$ is the class $C^0(m)$ of all systems of the form (1.1), with $f_0, \ldots, f_m$ continuous vector fields on some open subset Ω of some Euclidean space $\mathbb{R}^n$ and (ii) when $\mathcal{C}$ is an interesting subclass of $C^0(m)$, such as the set $C^1(m)$ of all systems (1.1) with C^1 vector fields. The input space $\mathcal{U}$ will be $L^1([0,T], \mathbb{R}^m)$, the set of all Lebesgue-integrable functions $t \mapsto u(t) = (u_1(t), \ldots, u_m(t))$ on some fixed interval $[0, T]$. Convergence with respect to $\mathcal{T}(L^1([0,T], \mathbb{R}^m), C^0(m))$ will be called "$\mathcal{T}^0$-convergence." For case (i), we will provide a complete and rather simple characterization of sequential $\mathcal{T}^0$-convergence by showing that a sequence $\{u^j\}$ $\mathcal{T}^0$-converges to a limit u^∞ iff the indefinite integrals $t \mapsto \int_0^t u^j(s)ds$ converge uniformly on $[0, T]$ to $t \mapsto \int_0^t u^\infty(s)ds$ and in addition $\sup_j \|u^j\|_{L^1} < \infty$. (Naturally, in order to provide a complete characterization of the topology $\mathcal{T}(L^1([0,T], \mathbb{R}^m), C^0(m))$, we should not limit our analysis to sequences and should seek instead to characterize $\mathcal{T}^0$-convergence of nets. This, however, appears to be a much more difficult question, as explained below, on which we will only be able to present some partial results.) For case (ii), we provide an even simpler characterization of $\mathcal{T}(L^1([0,T], \mathbb{R}^1), C^1(1))$-convergence of general nets $\{u^\alpha\}_{\alpha \in A}$ to a limit u^∞, by showing that it is equivalent to uniform convergence on $[0, T]$ of the indefinite integrals $t \mapsto \int_0^t u^\alpha(s)ds$ to the limit $t \mapsto \int_0^t u^\infty(s)ds$. This result is, however, valid only for the single-input case. To explain why, we analyze the m-input situation and prove, in sections 5 and 6, two results (Theorems 5.2 and 6.1), according to which (a) a characterization similar to that for $m = 1$ is true for general m on the class $C^1_{comm}(m)$ of systems (1.1) with C^1 vector fields such that all the Lie brackets $[f_j, f_k]$ for $j, k \in \{1, \ldots, m\}$ vanish, and (b) the result becomes false for any class $\mathcal{C}$ that contains at least one system in $C^1(m) \backslash C^1_{comm}(m)$. This shows that Lie brackets are intimately related to input convergence and ought to play a decisive role in any effort to achieve a better understanding of $\mathcal{T}(L^1([0,T], \mathbb{R}^m), C^1(m))$-convergence for $m > 1$.

So far, we have assumed that steps (J1) and (J2) have been carried out. We explain how this is done in section 2, where we define, on the set $TVVF(\Omega, I)$ of all time-varying vector fields $f : \Omega \times I \mapsto \mathbb{R}^n$ (where Ω is an open subset of $\mathbb{R}^n$ and I is a subinterval of $\mathbb{R}$), a topology $\mathcal{T}_T(\Omega, I)$—called the *topology of trajectory* *convergence*, or *T-convergence*—and argue that this topology captures the concept of "continuous dependence with respect to f of the solutions of the ordinary differential equation $\dot{x} = f(x,t)$." This topology takes into account the possibility of explosions and nonuniqueness of solutions—a fact of paramount importance for us, since, typically, the time-varying vector fields that arise from the systems (1.1) are continuous only with respect to x—and we will show in section 2 that it has all the right properties, at least when restricted to the set $TVVF_{Car,LIB}(\Omega, I)$ of those f's that satisfy the Carathéodory condition (i.e., are such that f is measurable in t for each x and continuous in x for each t) and are "locally integrably bounded," i.e., that satisfy, for every compact subset K of Ω, an integral bound

$$||f(x,t)|| \leq \varphi_K(t) \ \text{ for } \ (x,t) \in K \times I,$$

with a locally integrable function $\varphi_K : I \mapsto \mathbb{R}$. Moreover, the topology induced by $\mathcal{T}_T(\Omega, I)$ on the subset $TVVF_{Car,LIB}(\Omega, I)$ of $TVVF(\Omega, I)$ is the weakest topology on $TVVF_{Car,LIB}(\Omega, I)$ that gives rise to joint continuous dependence of the solutions on f and the initial condition.

Having defined $\mathcal{T}_T(\Omega, I)$, it is clear how to interpret (CD) and how to define $\mathcal{T}(\mathcal{U}, \mathcal{C})$: if $\Sigma \in \mathcal{C}$ is of the form $\dot{x} = g^\Sigma(x,u)$, $x \in \Omega^\Sigma$, then each input $t \mapsto u(t)$ belonging to $\mathcal{U}$ gives rise to a time-varying vector field $f^\Sigma(u) \in TVVF(\Omega^\Sigma, I)$, defined by

$$f^\Sigma(u) \overset{\text{def}}{=} \text{the map } (x,t) \mapsto g^\Sigma(x, u(t)) \,. \tag{1.2}$$

Then f^Σ is the *input-to-vector-field map* associated with Σ. If $\mathcal{T}$ is a topology on $\mathcal{U}$, then the statement that "the trajectories of Σ depend continuously on the input with respect to $\mathcal{T}$" can be translated as "f^Σ is continuous as a map from $\mathcal{U}$, equipped with $\mathcal{T}$, to $TVVF(\Omega^\Sigma, I)$, equipped with $\mathcal{T}_T(\Omega^\Sigma, I)$." Then the formal definition of $\mathcal{T}(\mathcal{U}, \mathcal{C})$ is simply the following:

> $\mathcal{T}(\mathcal{U}, \mathcal{C})$ *is the weakest topology on* $\mathcal{U}$ *that renders all the input-to-vector-field maps* $f^\Sigma : \mathcal{U} \mapsto TVVF(\Omega^\Sigma, I)$ *continuous for all* $\Sigma \in \mathcal{C}$, *when each* $TVVF(\Omega^\Sigma, I)$ *is given the topology* $\mathcal{T}_T(\Omega^\Sigma, I)$ *of trajectory convergence.*

To analyze in more detail the special case when $\mathcal{U} = L^1([0,T], \mathbb{R}^m)$ and $\mathcal{C}$ is the class $C^0(m)$, let us agree to use f_i^Σ to denote the vector fields f_i corresponding to a given system $\Sigma \in C^0(m)$ of the form (1.1), and let us go on using Ω^Σ to denote the state space of Σ. Write $\mathcal{U}_T^m \overset{\text{def}}{=} L^1([0,T], \mathbb{R}^m)$. As before, use "$\mathcal{T}^0$-convergence" for "convergence with respect to $\mathcal{T}(\mathcal{U}_T^m, C^0(m))$." Then our problem is to determine necessary and sufficient conditions for a sequence $\{u^j\}_{j=1}^\infty$ of functions belonging to the input space $\mathcal{U}_T^m$ to $\mathcal{T}^0$-converge to an input $u^\infty \in \mathcal{U}_T^m$, i.e., for the following property to hold:

(TC) For every possible choice of the system $\Sigma \in C^0(m)$, the time-varying vector fields $f^\Sigma(u^j)$ T-converge to $f^\Sigma(u^\infty)$.

As will be explained in section 2, one may substitute for (TC) either one of the following conditions, both of which turn out in fact to be equivalent to (TC):

(TC′) For every possible choice of the system $\Sigma \in C^0(m)$, if (i) ξ^j are maximal trajectories of (1.1) corresponding to the u^j and satisfying the initial conditions $\xi^j(0) = \bar{x}^j \in \Omega^\Sigma$, (ii) $\bar{x}^j \to \bar{x} \in \Omega^\Sigma$, and (iii) for the limiting initial value problem

$$\dot{x} = f^{\Sigma}(u^{\infty})(x,t) \overset{\mathrm{def}}{=} f_0^{\Sigma}(x) + \sum_{i=1}^{m} u_i^{\infty}(t) f_i^{\Sigma}(x)\,, \;\; x(0) = \bar{x}, \tag{1.3}$$

there is uniqueness of solutions as well as global existence on $[0,T]$, then the ξ^j are defined on $[0,T]$ for j large enough and converge uniformly on $[0,T]$ to the unique maximal solution ξ^{∞} of (1.3).

(TC$''$) For every possible choice of the system $\Sigma \in C^0(m)$, if (i) $\bar{x} \in \Omega^{\Sigma}$, (ii) ξ^j are maximal trajectories of (1.1) corresponding to the u^j and satisfying $\xi^j(0) = \bar{x}$, and (iii) the initial value problem (1.3) has uniqueness of solutions as well as global existence on $[0,T]$, then the ξ^j are defined on $[0,T]$ for j large enough and converge uniformly on $[0,T]$ to the unique maximal solution ξ^{∞} of (1.3).

Notice that (TC$'$) asserts "joint continuous dependence on the input and the initial condition," whereas (TC$''$) makes the weaker assertion of "continuous dependence on the input for each fixed initial condition," so the implication (TC$'$)$\Rightarrow$(TC$''$) is trivial, but the converse (TC$''$)$\Rightarrow$(TC$'$) is not at all obvious. (Example A.3 in the appendix shows that for a general sequence $\{f^j\}$ of time-varying vector fields it need not be true that convergence for each fixed initial condition of the trajectories of f^j to those of an $f \in TVVF(\Omega, I)$ implies convergence of trajectories for convergent initial conditions.)

Our characterization of sequential $\mathcal{T}^0$-convergence will be formulated in terms of a type of convergence somewhat weaker than other concepts of weak convergence previously considered in the literature. Precisely, let us say that a net $\{u^{\alpha}\}_{\alpha \in A} \subseteq \mathcal{U}_T^m$ based on a directed set $(A, \preceq_A)$ *converges in the integral sense* (or, for short, *I-converges*) to a $u^{\infty} \in \mathcal{U}_T^m$, if the following holds:

(IC) The indefinite integrals $U^{\alpha}(t) = \int_0^t u^{\alpha}(s)ds$ converge uniformly with respect to $t \in [0,T]$ to $U^{\infty}(t) = \int_0^t u^{\infty}(s)ds$.

Also, let us say that $\{u^{\alpha}\}_{\alpha \in A}$ satisfies the *uniform boundedness* condition if the following is true:

(UB) There exist a constant $C > 0$ and an $\alpha_0 \in A$ such that $\|u^{\alpha}\|_{L^1} \leq C$ whenever $\alpha_0 \preceq_A \alpha$.

Our main result (Theorem 4.1 below) then says that a sequence $\{u^j\}_{j=1}^{\infty}$ in $\mathcal{U}_T^m$ $\mathcal{T}^0$-converges to a function $u^{\infty} \in \mathcal{U}_T^m$ iff $\{u^j\}_{j=1}^{\infty}$ satisfies (UB) and I-converges to u^{∞}. An alternative formulation is in terms of *uniform weak convergence*. We say that a net $\{u^{\alpha}\}_{\alpha \in A}$ *UW-converges* to a limit u^{∞} if the following holds:

(UWC) For every continuous function $\varphi : [0,T] \mapsto \mathbb{R}$, if we define $\mathbb{R}^m$-valued functions U_{φ}^{α} by $U_{\varphi}^{\alpha}(t) = \int_0^t u^{\alpha}(s)\varphi(s)ds$ for $\alpha \in A \cup \{\infty\}$, then the U_{φ}^{α} converge to U_{φ}^{∞} uniformly on $[0,T]$.

For sequences, it is easily seen (cf. section 4) that the conjunction (IC) $\wedge$ (UB) is equivalent to (UWC), so Theorem 4.1 says in fact that a sequence $\{u^j\}_{j=1}^{\infty}$ $\mathcal{T}^0$-converges to a limit u^{∞} iff it UW-converges to u^{∞}.

For more general nets, the condition (IC) $\wedge$ (UB) still implies $\mathcal{T}^0$-convergence, which in turn implies UW-convergence, but the reverse implications no longer hold. (This is shown in the appendix by means of two examples.) Thus, although Theorem 4.1 provides a complete characterization of sequential $\mathcal{T}^0$-convergence in $\mathcal{U}_T^m$, it does not give a full characterization of the topology itself. To get such a characterization, one would have to understand $\mathcal{T}^0$-convergence of nets, and we have not yet been able to find a satisfactory characterization of $\mathcal{T}^0$-convergence of nets, comparable in simplicity to our results for sequences.

For sequences, the condition (IC) $\wedge$ (UB) or, equivalently, (UWC) is strictly weaker than weak convergence in $L^1([0,T],\mathbb{R}^m)$, as explained in section 7. In addition, the characterization of sequential $\mathcal{T}^0$-convergence in $\mathcal{U}_T^m$ depends crucially on the fact that we are working with the collection of *all* systems in $C^0(m)$. We could equally well have defined a concept of $\mathcal{T}^k$-convergence for any integer $k \geq 0$ by just taking $\mathcal{C}$ to be the class $C^k(m)$ of all systems of the form (1.1) with vector fields f_i of class C^k. But then it is no longer obvious that condition (UB) is still necessary for sequential $\mathcal{T}^k$-convergence for $k \geq 1$. (The proof that (UB) is necessary for sequential $\mathcal{T}^0$-convergence in $\mathcal{U}_T^m$ depends on using the uniform boundedness theorem and requires that we admit systems with an arbitrary continuous right-hand side. On the other hand, the necessity of (IC) depends only on the fact that the systems $\dot{x} = u_i$, $i = 1, \ldots, m$, are in our class, so (IC) is still necessary for trajectory convergence for practically any reasonable class of systems.) In fact, we show in section 5 that a sequence $\{u^j\}_{j=1}^{\infty} \subseteq \mathcal{U}_T^1$ may $\mathcal{T}^k$-converge to a $u^\infty \in \mathcal{U}_T^1$ even if (UB) fails.

To our knowledge, the problem of characterizing sequential $\mathcal{T}^k$-convergence in $\mathcal{U}_T^m$ for $k \geq 1$ is still open for $m > 1$ and appears to be quite difficult. The case $m = 1$ is special, as explained in section 5, where we give a complete characterization (in Theorem 5.1) of the topology $\mathcal{T}(\mathcal{U}_T^1, C^1(1))$ by showing that a net $\{u^\alpha\}_{\alpha\in A} \subseteq \mathcal{U}_T^1$ $\mathcal{T}(\mathcal{U}_T^1, C^1(1))$-converges to a limit u^∞ iff the u^α I-converge to u^∞. The profound reason why this characterization is possible for $m = 1$ but not for $m > 1$ is provided by Theorems 5.2 and 6.1: Theorem 5.2 shows that the condition of Theorem 5.1 is necessary and sufficient, for general m, for $\mathcal{T}(\mathcal{U}_T^m, \mathcal{C})$-convergence of the u^α to u^∞ if we take $\mathcal{C} = C^1_{comm}(m)$, the class of systems (1.1) where the f_i are of class C^1 and $f_1, \ldots, f_m$ *commute* (i.e., the Lie brackets $[f_i, f_j]$ vanish for $1 \leq i, j \leq m$), whereas Theorem 6.1 shows that for any class $\mathcal{C}$ such that $C^1_{comm}(m) \subseteq \mathcal{C} \subseteq C^1(m)$ and $\mathcal{C} \neq C^1_{comm}(m)$, the condition no longer suffices for $\mathcal{T}(\mathcal{U}_T^m, \mathcal{C})$-convergence, because there exists a sequence $\{u^j\}_{j=1}^{\infty}$ such that

(1) $\int_0^t u^j(s)ds \to 0$ uniformly;
(2) for every $\Sigma \in \mathcal{C}$ there is a time-varying vector field $h^\Sigma \in TVVF(\Omega^\Sigma, [0,T])$ such that the $\{f^\Sigma(u^j)\}$ T-converge to h^Σ;
(3) it is not true that $h^\Sigma = f^\Sigma(0)$ for all $\Sigma \in \mathcal{C}$.

(What really happens is that the sequence $\{u^j\}$ $\mathcal{T}(\mathcal{U}_T^m, \mathcal{C})$-converges to a “generalized input” that is no longer an ordinary input. A more detailed discussion of these issues is provided in section 6.)

The paper is organized as follows. In section 2 we present the definition and basic properties of the topology of T-convergence. We then give, in section 3, our precise definition of $\mathcal{T}^0$-convergence. The main theorem (Theorem 4.1) of the paper is stated and proved in section 4. In section 5 we discuss how the situation changes when $\mathcal{T}^k$-convergence for $k \geq 1$ is considered instead of $\mathcal{T}^0$-convergence. In section 6 we briefly discuss some of the phenomena that can cause a sequence $\{f^\Sigma(u^j)\}$ to T-converge to something other than its limit $f^\Sigma(u^\infty)$, in particular the occurrence of Lie bracket terms. Section 7 discusses how our continuous dependence conditions relate to those of Buttazzo–Conti, Kurzweil–Vorel, and Neustadt and shows that our $\mathcal{T}^0$-convergence conditions for sequences in $L^1([0,T],\mathbb{R}^m)$ are weaker than weak convergence in L^1. Finally, we include an appendix that presents three examples, showing (i) that UW-convergence of nets does not imply $\mathcal{T}^0$-convergence, (ii) that $\mathcal{T}^0$-convergence of nets does not imply the uniform boundedness condition (UB), and (iii) that, for general time-varying vector fields, "continuous dependence of trajectories for every fixed initial condition" does not imply "joint continuous dependence of trajectories on the vector field and the initial condition."

2. The topology of trajectory convergence. We begin by giving a precise meaning to the "T-convergence" condition that appears in (TC).

In what follows, the word "interval" means "connected, nonempty subset of $\mathbb{R}$." An interval is *nontrivial* if it is not reduced to a single point.

If Ω is a metric space, we use d_Ω to denote the corresponding distance function. A *curve* in Ω is a continuous map ξ from an interval $I \subseteq \mathbb{R}$ to Ω. If I is compact, then ξ will be called an *arc* in Ω. We use $ARC(\Omega)$ to denote the set of all arcs in Ω. For every $-\infty < a \leq b < \infty$, we let $C^0([a,b],\Omega)$ denote the set of all arcs in Ω whose domain is $[a,b]$. Then

$$ARC(\Omega) = \bigcup_{-\infty<a\leq b<\infty} C^0([a,b],\Omega)\,. \tag{2.1}$$

For each $\xi \in ARC(\Omega)$, we let $a(\xi)$, $b(\xi)$ be the unique a,b such that ξ belongs to $C^0([a,b],\Omega)$. If $\xi \in ARC(\Omega)$, let $\tilde{\xi}$ be the continuous curve $\tilde{\xi} : \mathbb{R} \mapsto \Omega$ obtained by extending ξ to the whole real line by just letting $\tilde{\xi}$ be constant on $]-\infty, a(\xi)]$ and on $[b(\xi),+\infty[$. We topologize $ARC(\Omega)$ by means of the metric $d_{ARC} : ARC(\Omega) \times ARC(\Omega) \mapsto \mathbb{R}$ given by

$$d_{ARC}(\xi,\eta) = |a(\xi)-a(\eta)| + |b(\xi)-b(\eta)| + \sup\{d_\Omega(\tilde{\xi}(t),\tilde{\eta}(t)) \,:\, t \in \mathbb{R}\}\,.$$

Then a sequence $\{\xi^j\}$ of arcs, possibly defined on different time intervals $[a^j,b^j]$, converges to an arc $\xi : [a,b] \mapsto \Omega$ if (i) $a^j \to a$, (ii) $b^j \to b$, and (iii) $\xi^j \to \xi$ uniformly, i.e., $\xi^j(t^j) \to \xi(t)$ whenever $t^j \in [a^j,b^j]$, $t^j \to t$.

If $\xi : J \mapsto \Omega$ is a curve, the *graph* of ξ is the set $G(\xi) = \{(\xi(t),t) : t \in J\}$.

If $\xi_i : [a_i,b_i] \mapsto \Omega$ for $i = 1,2$ are arcs in Ω, then the *concatenation* $\xi_1\#\xi_2$ is defined iff $b_1 = a_2$ and $\xi_1(b_1) = \xi_2(a_2)$, and is the arc whose graph is $G(\xi_1) \cup G(\xi_2)$. Concatenation is an associative partially defined binary operation on $ARC(\Omega)$.

A subset $\mathcal{A} \subseteq ARC(\Omega)$ is *bounded* if there is a compact subset K of $\Omega \times \mathbb{R}$ such that $G(\xi) \subseteq K$ for all $\xi \in \mathcal{A}$. We call $\mathcal{A}$ *equicontinuous* if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $d_\Omega(\xi(t),\xi(s)) < \varepsilon$ whenever $\xi \in \mathcal{A}$, $t,s \in [a(\xi),b(\xi)]$, and $|t-s| \leq \delta$. It is then clear that a subset $\mathcal{A}$ of $ARC(\Omega)$ is relatively compact iff $\mathcal{A}$ is bounded and equicontinuous.

An *arc system in* Ω *with time domain* I is a subset $\mathcal{A}$ of $ARC(\Omega)$ such that

(AS.1) $G(\xi) \subseteq \Omega \times I$ for every $\xi \in \mathcal{A}$,

(AS.2) $\mathcal{A}$ contains every arc in Ω whose domain is of the form $\{t\}$ for $t \in I$,

(AS.3) $\mathcal{A}$ is closed under concatenations (that is, if ξ_1, ξ_2 are in $\mathcal{A}$ and $\xi_1\#\xi_2$ is defined, then $\xi_1\#\xi_2 \in \mathcal{A}$),

(AS.4) $\mathcal{A}$ is closed under restrictions (that is, if $\xi \in \mathcal{A}$ and $a(\xi) \leq \alpha \leq \beta \leq b(\xi)$, then the restriction $\xi\lceil[\alpha,\beta]$ of ξ to $[\alpha,\beta]$ is in $\mathcal{A}$).

We use $AS(\Omega,I)$ to denote the set of all arc systems in Ω with time domain I. If $\mathcal{A} \in AS(\Omega,I)$ and $K \subseteq \Omega \times I$, we write

$$\mathcal{A}(K) \overset{\text{def}}{=} \{\xi \in \mathcal{A} : G(\xi) \subseteq K\}\,. \tag{2.2}$$

If $\mathcal{A} \in AS(\Omega, I)$, then it follows easily that if (i) the arcs $\xi_1, \ldots, \xi_m$ belong to $\mathcal{A}$, (ii) $[a(\xi_1), b(\xi_1)] \cup \cdots \cup [a(\xi_m), b(\xi_m)]$ is an interval, (iii) $\xi_i \lceil I_i \cap I_j = \xi_j \lceil I_i \cap I_j$ for all $i, j \in \{1, \ldots, m\}$, where we write $I_\ell = [a(\xi_\ell), b(\xi_\ell)]$ for $\ell \in \{1, \ldots, m\}$, and (iv) we let $\xi_1 \vee \cdots \vee \xi_m$ be the arc whose graph is $G(\xi_1) \cup \cdots \cup G(\xi_m)$, then $\xi_1 \vee \cdots \vee \xi_m$ is also in $\mathcal{A}$.

If $\mathcal{A} \in AS(\Omega, I)$, we define $\mathcal{A}^\#$ to be the set of all curves $\xi : J \mapsto \Omega$ such that J is a subinterval of I and $\xi \lceil L \in \mathcal{A}$ for every compact subinterval L of J. The members of $\mathcal{A}^\#$ will be called *$\mathcal{A}$-curves*. An $\mathcal{A}$-curve $\xi \in \mathcal{A}^\#$ is *maximal* if there does not exist $\eta \in \mathcal{A}^\#$ such that $G(\xi)$ is a proper subset of $G(\eta)$. We use $\mathcal{A}^\#_{max}$ to denote the set of all maximal elements of $\mathcal{A}^\#$. A trivial application of Zorn's lemma shows that if $\mathcal{A} \in AS(\Omega, I)$, then every $\xi \in \mathcal{A}^\#$ can be extended to an $\eta \in \mathcal{A}^\#_{max}$. This implies, in particular, that through every point of $\Omega \times I$ there passes the graph of a $\xi \in \mathcal{A}^\#_{max}$.

If $\mathcal{A} \in AS(\Omega, I)$, we say that $\mathcal{A}$ has the *compactness property* if for every compact subset $K \subseteq \Omega \times I$ the set $\mathcal{A}(K)$ is compact in $ARC(\Omega)$. We say that $\mathcal{A}$ has the *local existence property* if for every $(x, t) \in \Omega \times I$ there exists $\xi \in \mathcal{A}$ such that t belongs to the interior of $[a(\xi), b(\xi)]$ relative to I and $\xi(t) = x$. We use $AS_{comp,le}(\Omega, I)$ to denote the set of all $\mathcal{A} \in AS(\Omega, I)$ that have the compactness property and the local existence property.

PROPOSITION 2.1. *Let Ω be a locally compact metric space, and let $I \subseteq \mathbb{R}$ be an interval. Assume that $\mathcal{A} \in AS_{comp,le}(\Omega, I)$. Let $\xi : J \mapsto \Omega$ be a curve in $\mathcal{A}^\#$. Then $\xi \in \mathcal{A}^\#_{max}$ iff* (i) *J is relatively open in I and* (ii) *for every compact subset K of $\Omega \times I$ the set $\{t \in J : (\xi(t), t) \in K\}$ is compact.*

Proof. Let $t_+ = \sup J$, $t_- = \inf J$.

Suppose first that (i) and (ii) hold, but ξ is not maximal. Let $\eta \in \mathcal{A}^\#$ be an extension of ξ to a strictly larger interval J'. Pick $s \in J' \backslash J$. Then either $s \geq t_+$ or $s \leq t_-$. Assume that $s \geq t_+$. (The other case is similar.) Then $t_+ \in I$, because $t_+ \leq s \in I$ and $t_+ \geq t$ for some $t \in J \subseteq I$. Therefore $t_+ \notin J$, because J is relatively open in I by (i). Pick $t \in J$. Let K_0 be the set $\{\eta(\tau) : t \leq \tau \leq t_+\}$. Let $K = K_0 \times [t, t_+]$. Then K is compact, so the set $L = \{\tau \in J : (\xi(\tau), \tau) \in K\}$ is compact by (ii). But this is a contradiction, because $[t, t_+[\subseteq L$ but $t_+ \notin L$.

Now assume that ξ is maximal. Suppose that J is not relatively open in I. Then either $t_+ \in J$ and $t_+ < \sup I$ or $t_- \in J$ and $t_- > \inf I$. Assume that $t_+ \in J$ and $t_+ < \sup I$, the other case being similar. By the local existence property, there is an $\eta \in \mathcal{A}$ for which $\eta(t_+) = \xi(t_+)$ and whose domain Q contains t in its interior relative to I. By the restriction property, we may assume that the domain of η is of the form $[t_+, t_+ + \varepsilon]$ for some $\varepsilon > 0$. But then $G(\xi) \cup G(\eta)$ is the graph of an arc in $\mathcal{A}^\#$, contradicting the maximality of ξ. So J is relatively open in I, as stated.

Let $K \subseteq \Omega \times I$ be compact, and assume that $L = \{t \in J : (\xi(t), t) \in K\}$ is not compact. Then in particular $L \neq \emptyset$. Let $\pi : \Omega \times I \mapsto I$ be the projection map $(x, t) \mapsto t$. Let $b = \sup L$, $a = \inf L$. Then $a \in I$ and $b \in I$, because $L \subseteq \pi(K) \subseteq I$, and $\pi(K)$ is compact. Also, $a < b$, since if $a = b$, then L would be equal to $\{a\}$, which is compact. Since $L \subseteq [a, b]$ and L is not compact, there exists a sequence $\{t_j\}$ in L that converges to a limit $\bar{t} \notin L$. On the other hand, the open interval $]a, b[$ is contained in J. If $a < \bar{t} < b$, it would follow that $\xi(t_j) \to \xi(\bar{t})$, so $(\xi(\bar{t}), \bar{t})$ would be in K, and $\bar{t}$ would belong to L. Therefore either $\bar{t} = a$ or $\bar{t} = b$. Assume $\bar{t} = b$, the case when $\bar{t} = a$ being similar. Since $(\xi(t_j), t_j) \in K$, which is compact, we may assume, after passing to a subsequence, that $\bar{x} = \lim_{j \to \infty} \xi(t_j)$ exists. Clearly, $(\bar{x}, \bar{t}) \in K$. Let K_0', be a compact neighborhood of $\bar{x}$ in Ω, and let J_0' be a compact subinterval of I such that $\bar{t}$ belongs to the interior of J_0' relative to I. Let $K' = K_0' \times J_0'$. Let

$\varepsilon > 0$ be such that $x \in \Omega$, $t \in I$, $||x - \bar{x}|| \leq \varepsilon$, $|\bar{t} - t| \leq \varepsilon$ implies $(x,t) \in K'$. Using the compactness property for $\mathcal{A}$, we can infer that $\mathcal{A}(K')$ is equicontinuous, so there exists a $\delta > 0$ such that $\eta \in \mathcal{A}$, $G(\eta) \subseteq K'$ implies $||\eta(t) - \eta(s)|| \leq \frac{\varepsilon}{3}$ whenever $|t-s| \leq \delta$. Pick j so large that $b - t_j < \delta$ and $||\xi(t_j) - \bar{x}|| < \frac{\varepsilon}{3}$. Then $\xi([t_j, b[) \subseteq K_0'$. (Otherwise, let $\tau = \inf\{t \in [t_j, b[: \xi(t) \notin K_0'\}$. Then $t_j < \tau < b$, $\xi(\tau) \in K_0'$, and $\xi\lceil[t_j,\tau] \in \mathcal{A}(K')$. Since $\tau - t_j < \delta$, we have $||\xi(\tau) - \xi(t_j)|| < \frac{\varepsilon}{3}$, so $||\xi(\tau) - \bar{x}|| < \frac{2\varepsilon}{3}$, and then $\xi(\tau)$ is an interior point of K_0'. But then $\xi(t) \in K_0'$ for $t > \tau$, $t - \tau$ small enough, contradicting the definition of τ.) Using equicontinuity again, for the set of arcs $\{\xi\lceil[t_j,t] : t_j \leq t < b\} \subseteq \mathcal{A}(K')$, we see that the limit of $\xi(t)$ as $t \uparrow b$ must exist. Therefore $\lim_{t\uparrow b} \xi(t) = \bar{x}$. Then the arcs $\xi\lceil[t_j,t]$ converge to η as $t \uparrow b$, where $\eta : [t_j, b] \mapsto \Omega$ is such that $\eta\lceil[t_j,b[= \xi\lceil[t_j,b[$ and $\eta(b) = \bar{x}$. By the compactness property, $\eta \in \mathcal{A}$. Therefore, if we extend ξ to $J \cup \{b\}$ by letting $\xi(b) = \bar{x}$, we see that the extension belongs to $\mathcal{A}^{\#}$. Since ξ is maximal, we conclude that $b \in J$ and $\xi(b) = \bar{x}$. Since $(\bar{x}, b) = \lim_{j\to\infty}(\xi(t_j), t_j)$, we see that $(\bar{x}, b) \in K$, so $b \in L$, contradicting the fact that $b = \bar{t} \notin L$. □

If X is a set, we use 2^X to denote the set of all subsets of X. If X is a topological space, we can topologize 2^X as follows. For each open subset U of X, let

$$V(U) = \{S \in 2^X \, : \, S \subseteq U\}\,. \tag{2.3}$$

We then let $\mathcal{T}_{USC}(X)$ denote the set of all subsets S of 2^X such that $S = \cup_\alpha V(U_\alpha)$ for some family $\{U_\alpha\}$ of open subsets of X. Then $\mathcal{T}_{USC}(X)$ is a topology on 2^X. We will call it the *USC topology* on 2^X, because of its relation to upper-semicontinuity of set-valued maps. (If Y is a topological space, a set-valued map $F : Y \mapsto 2^X$ is upper semicontinuous iff it is continuous as a map from Y to $(2^X, \mathcal{T}_{USC}(X))$.) The name *upper semifinite topology* has also been used in the literature; cf., e.g., [17].

If Ω is a metric space and I is an interval, then for every subset K of $\Omega \times I$ we have a set-valued map $T_{\Omega,I,K} : AS(\Omega, I) \mapsto 2^{ARC(\Omega)}$ given by $T_{\Omega,I,K}(\mathcal{A}) = \mathcal{A}(K)$. As K ranges over all the compact subsets of $\Omega \times I$, we have a collection of maps $T_{\Omega,I,K}$ from $AS(\Omega, I)$ to $2^{ARC(\Omega)}$. Therefore there exists a weakest topology on $AS(\Omega, I)$ that renders all the maps $T_{\Omega,I,K}$ continuous, where $2^{ARC(\Omega)}$ is equipped with the USC topology. This topology will be called the *topology of trajectory convergence* (or simply *T-convergence*) and will be denoted by $\mathcal{T}_T(\Omega, I)$. If $\{\mathcal{A}^\alpha\}_{\alpha\in A}$ is a net in $AS(\Omega, I)$ and $\mathcal{A} \in AS(\Omega, I)$, we say that $\{\mathcal{A}^\alpha\}_{\alpha\in A}$ *T-converges* to $\mathcal{A}$ (and write $\mathcal{A}^\alpha \xrightarrow{\text{T}} \mathcal{A}$) if the $\mathcal{A}^\alpha$ converge to $\mathcal{A}$ in $\mathcal{T}_T(\Omega, I)$.

Let Ω be a metric space, let $\xi : I \mapsto \Omega$ be a curve, and assume that $\{\xi^\alpha\}_{\alpha\in A}$ is a net of curves in Ω, with $\mathrm{Dom}(\xi^\alpha) = I^\alpha$. We say that the ξ^α *converge to* ξ *on compact sets* if for every compact subset $J \subseteq I$, there exists an $\alpha^*(J)$ such that (i) $J \subseteq I^\alpha$ for $\alpha \succeq_A \alpha^*(J)$ and (ii) the $\xi^\alpha\lceil J$ converge uniformly to $\xi\lceil J$.

We say that a net $\{\mathcal{A}^\alpha\}_{\alpha\in A}$ in $AS(\Omega, I)$ *converges to* $\mathcal{A} \in AS(\Omega, I)$ *in the maximal curve convergence sense* if, whenever $\{(x^\alpha, t^\alpha)\}_{\alpha\in A}$ is a net of points in $\Omega \times I$ that converges to a limit $(\bar{x}, \bar{t}) \in \Omega \times I$, and $\{\xi^\alpha\}_{\alpha\in A}$ is a net such that for each $\alpha \in A$ $\xi^\alpha \in (\mathcal{A}^\alpha)^{\#}_{max}$ and $\xi^\alpha(t^\alpha) = x^\alpha$, it follows that there exist a $\xi^\infty \in \mathcal{A}^{\#}_{max}$, with domain $I^\infty \subseteq I$, such that $\xi^\infty(\bar{t}) = \bar{x}$, and a subnet $\{\xi^{\mu(\beta)}\}_{\beta\in B}$ of $\{\xi^\alpha\}_{\alpha\in A}$ that converges to ξ^∞ on compact sets.

Proposition 2.2. *Let Ω be a locally compact metric space, and let $I \subseteq \mathbb{R}$ be an interval. Suppose that $\{\mathcal{A}^\alpha\}_{\alpha\in A}$ is a net in $AS_{comp,le}(\Omega, I)$, and $\mathcal{A} \in AS_{comp,le}(\Omega, I)$. Then the following conditions are equivalent:*

(I) *the $\mathcal{A}^\alpha$ T-converge to $\mathcal{A}$;*

(II) *every subnet of* $\{\mathcal{A}^\alpha\}_{\alpha\in A}$ *converges to* $\mathcal{A}$ *in the maximal curve convergence sense.*

Proof. The implication (II)⇒(I) is trivial. Indeed, suppose (II) holds, and let K be a compact subset of $\Omega\times I$. Since $\mathcal{A}(K)$ is compact, a fundamental system of neighborhoods of $\mathcal{A}(K)$ in $2^{ARC(\Omega)}$ is given by the sets

$$V_\varepsilon = \{\xi \in ARC(\Omega) : d_{ARC}(\xi, \mathcal{A}(K)) < \varepsilon\}$$

for $\varepsilon > 0$. So it suffices to fix ε and prove that $\mathcal{A}^\alpha(K)\subseteq V_\varepsilon$ for sufficiently large α. If this was not so, there would exist a subnet $\{\mathcal{A}^{\mu(\beta)}\}_{\beta\in B}$ of $\{\mathcal{A}^\alpha\}_{\alpha\in A}$ such that there are $\xi^\beta\in\mathcal{A}^{\mu(\beta)}(K)$ for which $d_{ARC}(\xi^\beta,\mathcal{A}(K))\geq\varepsilon$. Extend each ξ^β to a $\tilde\xi^\beta\in(\mathcal{A}^{\mu(\beta)})^\#_{max}$. Pick $t^\beta\in\mathrm{Dom}(\xi^\beta)$ in an arbitrary fashion, and let $x^\beta=\xi^\beta(t^\beta)$. Then $(x^\beta,t^\beta)\in K$, so by passing to a subnet if necessary we may assume that (x^β,t^β) converges to a limit $(\bar x,\bar t)$. Using (II) we may assume, after passing to a subnet, that there exists a $\tilde\xi\in\mathcal{A}^\#_{max}$ such that $\tilde\xi(\bar t)=\bar x$ and $\tilde\xi^\beta\to\tilde\xi$ on compact sets. After passing once again to a subnet, we may assume that $a(\xi^\beta)$ and $b(\xi^\beta)$ have limits a,b, both belonging to I. Let $J=\mathrm{Dom}(\tilde\xi)$. We will prove that $a\in J$ and $b\in J$. Since both proofs are similar, we will show only that $b\in J$. Suppose that $b\notin J$. Let $c=\sup J$, so $c\leq b$. Since $b\in I$, and J is relatively open in I, we have $c\notin J$. This implies in particular that $c>\bar t$, since $\bar t\in J$. Pick $c'\in[\bar t,c[$. Then $[\bar t,c']\subseteq J$, so $[\bar t,c']\subseteq\mathrm{Dom}(\tilde\xi^\beta)$ for large β, and $\tilde\xi^\beta\lceil[\bar t,c']\to\tilde\xi\lceil[\bar t,c']$ uniformly. Since $b(\xi^\beta)\to b>c'$, we have $b(\xi^\beta)>c'$ for large β. Since $a(\xi^\beta)\leq t^\beta\to\bar t<c'$, we have $a(\xi^\beta)<c'$ for large β. Therefore $c'\in\mathrm{Dom}(\xi^\beta)$ for large β, so $\tilde\xi(c')=\lim\xi^\beta(c')$. Therefore $(\tilde\xi(c'),c')\in K$. Since this is true for $c'<c$ but arbitrarily close to c, and the set $L=\{t\in J:(\tilde\xi(t),t)\in K\}$ is compact by Proposition 2.1, we conclude that $c\in L$, so in particular $c\in J$, which is a contradiction.

Thus a and b belong to J. Therefore $[a,b]\subseteq\mathrm{Dom}(\tilde\xi^\beta)$ for large β, and $\tilde\xi^\beta\lceil[a,b]$ converges to $\tilde\xi\lceil[a,b]$. Since $a(\xi^\beta)\to a$, and $b(\xi^\beta)\to b$, given any $s\in[a,b]$ we can find $s^\beta\in\mathrm{Dom}(\xi^\beta)$ such that $a^\beta\leq s^\beta\leq b^\beta$ and $s^\beta\to s$. Therefore $\xi^\beta(s^\beta)\to\tilde\xi(s)$. So $(\tilde\xi(s),s)\in K$. This shows that the restriction ξ of $\tilde\xi$ to $[a,b]$ is in $\mathcal{A}(K)$. Therefore $d_{ARC}(\xi^\beta,\xi)\geq\varepsilon$. On the other hand, $\xi^\beta\to\xi$ in $ARC(\Omega)$, so $d_{ARC}(\xi^\beta,\xi)\to0$. This contradiction proves that (II)⇒(I), as stated.

We now prove that (I)⇒(II). Assume that the net $\{\mathcal{A}^\alpha\}_{\alpha\in A}$ T-converges to $\mathcal{A}$. We want to show that every subnet of $\{\mathcal{A}^\alpha\}_{\alpha\in A}$ converges to $\mathcal{A}$ in the maximal curve convergence sense. Since every subnet of $\{\mathcal{A}^\alpha\}_{\alpha\in A}$ T-converges to $\mathcal{A}$, because the notion of T-convergence arises from a topology, it suffices to show that $\{\mathcal{A}^\alpha\}_{\alpha\in A}$ converges to $\mathcal{A}$ in the maximal curve convergence sense.

Let $\{\xi^\alpha\}_{\alpha\in A}$ be a net such that $\xi^\alpha\in(\mathcal{A}^\alpha)^\#_{max}$, and let I^α be the domain of ξ^α. Let $t^\alpha\in I^\alpha$ be such that $t^\alpha\to\bar t\in I$, and $x^\alpha\overset{\mathrm{def}}{=}\xi^\alpha(t^\alpha)\to\bar x\in\Omega$. Pick a compact neighborhood K_0 of $\bar x$ in Ω and a compact interval J_0 which is a relative neighborhood of $\bar t$ in I. By passing to a subnet of $\{\xi^\alpha\}$, we may assume that $x^\alpha\in K_0$ and $t^\alpha\in J_0$ for all α. Let $\mathbf{Q}$ be the set of all products $Q=K\times J$, where K is a compact subset of Ω such that $K_0\subseteq K$, and J is a compact subinterval of I such that $J_0\subseteq J$. For each $Q=K\times J\in\mathbf{Q}$, let $L^\alpha(Q)=\{t\in I^\alpha:(\xi^\alpha(t),t)\in Q\}$. Then $L^\alpha(Q)$ is compact. Let $S^\alpha(Q)$ be the connected component of $L^\alpha(Q)$ that contains t^α. Then $S^\alpha(Q)$ is a compact subinterval of I. Let $\xi^{\alpha,Q}$ be the restriction of ξ^α to $S^\alpha(Q)$. Then $\xi^{\alpha,Q}\in\mathcal{A}^\alpha(Q)$. Since $\mathcal{A}(Q)$ is compact, we can pick $\zeta^{\alpha,Q}\in\mathcal{A}(Q)$ such that $d_{ARC}(\xi^{\alpha,Q},\zeta^{\alpha,Q})=d_{ARC}(\xi^{\alpha,Q},\mathcal{A}(Q))$. Since $\mathcal{A}^\alpha\xrightarrow{\mathrm{T}}\mathcal{A}$, the distance

$d_{ARC}(\xi^{\alpha,Q},\mathcal{A}(Q))$ goes to 0. So $d_{ARC}(\xi^{\alpha,Q},\zeta^{\alpha,Q})\to 0$. Let

$$U=\prod_{Q\in\mathbf{Q}}\mathcal{A}(Q)\,. \tag{2.4}$$

Then U is a product of compact topological spaces. By Tikhonov's theorem, U is compact with respect to the product topology. Let $Z^\alpha=\{\zeta^{\alpha,Q}\}_{Q\in\mathbf{Q}}$. Then each Z^α is a member of U. Therefore there is a subnet $\{Z^{\mu(\beta)}\}_{\beta\in B}$ of $\{Z^\alpha\}_{\alpha\in A}$ that converges to a limit $Z=\{\zeta^Q\}_{Q\in\mathbf{Q}}\in U$. Then $\zeta^{\mu(\beta),Q}\to\zeta^Q$ for every $Q\in\mathbf{Q}$. Since $d_{ARC}(\xi^{\mu(\beta),Q},\zeta^{\mu(\beta),Q})\to 0$, we have shown that $\xi^{\mu(\beta),Q}\to\zeta^Q$ for each Q.

If $Q,Q'\in\mathbf{Q}$ and $Q\subseteq Q'$, then it is easy to see that $L^\alpha(Q)\subseteq L^\alpha(Q')$ for all $\alpha\in A$. Therefore $S^\alpha(Q)\subseteq S^\alpha(Q')$, and then $G(\xi^{\alpha,Q})\subseteq G(\xi^{\alpha,Q'})$. In particular, $G(\xi^{\mu(\beta),Q})\subseteq G(\xi^{\mu(\beta),Q'})$ for all β and then, taking limits, it follows that $G(\zeta^Q)\subseteq G(\zeta^{Q'})$. Since for any two members Q, Q' of $\mathbf{Q}$ there is a $Q''\in\mathbf{Q}$ such that $Q\cup Q'\subseteq Q''$, it follows that the arcs ζ^Q match, in the sense that the union of the graphs $G(\zeta^Q)$ is the graph of a map ζ. If $S(Q)=\mathrm{Dom}(\zeta^Q)$, then $\bar t\in S(Q)$, because $t^{\mu(\beta)}\in\mathrm{Dom}(\xi^{\mu(\beta),Q})$ for each β, and $t^{\mu(\beta)}\to\bar t$. Then $S=\cup_{Q\in\mathbf{Q}}S(Q)$ is an interval, and it is clear that $S=\mathrm{Dom}(\zeta)$. Obviously, $\zeta(\bar t)=\bar x$. We now have to show that $\zeta\in\mathcal{A}^{\#}_{max}$ and $\xi^{\mu(\beta)}\to\zeta$ on compact sets.

Write

$$S^\alpha(Q)=[a^\alpha(Q),b^\alpha(Q)]\,,\quad S(Q)=[a(Q),b(Q)]\,, \tag{2.5}$$

so $a(Q)=\lim_\beta a^{\mu(\beta)}(Q)$ and $b(Q)=\lim_\beta b^{\mu(\beta)}(Q)$.

Next, we make the following observation:

(*) Given any $\alpha\in A$ and any $Q=K\times J\in\mathbf{Q}$, the point $\xi^\alpha(b^\alpha(Q))$ belongs to $K\backslash\mathrm{Int}(K)$ unless $b^\alpha(Q)=\sup I$.

Indeed, if $b^\alpha(Q)<\sup I$ and $\xi^\alpha(b^\alpha(Q))\in\mathrm{Int}(K)$, then $\xi^\alpha(t)$ would be defined for $t=b^\alpha(Q)+h$ if $h>0$ is small enough, because the domain of ξ^α is relatively open in I. By continuity, $(\xi^\alpha(t),t)$ would be in Q for $t=b^\alpha(Q)+h$, $h>0$, h small, contradicting the definition of $b^\alpha(Q)$. Thus (*) holds, as stated.

Let $\sigma=\sup S$. We distinguish three cases, namely,

(i) $\sigma\notin I$,
(ii) $\sigma\in I\backslash S$,
(iii) $\sigma\in S$.

In cases (i) and (ii), if $b\in S$, then $b<\sigma$, so $b<b(Q)$ for some $Q\in\mathbf{Q}$, since $\sigma=\sup\{b(Q):Q\in\mathbf{Q}\}$. Therefore $b<b^{\mu(\beta)}(Q)$ for large enough β. In case (iii), we must have $\sigma=b(\bar Q)$ for a $\bar Q\in\mathbf{Q}$. Let $Q'=K'\times J'$ be such that $\zeta(\sigma)\in\mathrm{Int}_\Omega(K')$ and pick a $Q=K\times J\in\mathbf{Q}$ such that $\bar Q\cup Q'\subseteq Q$. Then $b(Q)=\sigma$, and $\zeta^Q(b(Q))=\zeta(b(Q))\in\mathrm{Int}(K)$. Thus $\xi^{\mu(\beta),Q}\in\mathrm{Int}(K)$ for large β. Therefore (*) implies that $b^{\mu(\beta)}(Q)=\sigma$ for large β.

Summarizing the above, we have shown the following:

(**) If $b\in S$, then there is a $Q\in\mathbf{Q}$ such that $b\le b^{\mu(\beta)}(Q)$ for all sufficiently large β.

Naturally, a similar reasoning shows that if $a\in S$, then there is a Q such that $a^{\mu(\beta)}(Q)\le a$ for all sufficiently large β. Therefore, if $S'=[a,b]$ is any compact subinterval of S, there exists a Q such that $[a,b]\subseteq S^{\mu(\beta)}(Q)$ for all large enough β. It follows that the restriction $\zeta\lceil[a,b]$ is the arc $\zeta^Q\lceil[a,b]$. Since $\zeta^Q\in\mathcal{A}$, we see that $\zeta\lceil[a,b]\in\mathcal{A}$.

This is true for every S'. So we have shown that $\zeta \in \mathcal{A}^{\#}$. Moreover, it also follows that, for any S', there is a Q such that the inclusion $S' \subseteq S^{\mu(\beta)}(Q) \subseteq \mathrm{Dom}(\xi^{\mu(\beta)})$ holds for large enough β. This shows that $\xi^{\mu(\beta)} \to \zeta$ on compact sets.

To conclude, we have to show that $\zeta \in \mathcal{A}^{\#}_{max}$. Assume that this is not true, and let $\tilde{\zeta} : \tilde{S} \mapsto \Omega$ be a curve, defined on a subinterval $\tilde{S}$ of I such that $S \subseteq \tilde{S}$ but $S \neq \tilde{S}$, for which $\tilde{\zeta}\lceil S = \zeta$. Pick $s \in \tilde{S} \backslash S$. Then either $s \geq \sup S$ or $s \leq \inf S$. We assume that $s \geq \sup S$, the other case being similar. As before, we write $\sigma = \sup S$. Then $\sigma \in I$, because $\sigma \leq s \in I$ and $\sigma \geq t$ for some $t \in S$. Pick a $Q_1 = K_1 \times J_1 \in \mathbf{Q}$ such that $\tilde{\zeta}(\sigma) \in \mathrm{Int}(K_1)$. Pick a $\sigma' \in S$ such that $\sigma' \leq \sigma$ and $\zeta(t) \in \mathrm{Int}(K_1)$ whenever $\sigma' \leq t$ and $t \in S$. Using (**), pick a $Q_2 = K_2 \times J_2$ such that $\sigma' \leq b^{\mu(\beta)}(Q_2)$ for all sufficiently large β. Pick $Q = K \times J \in \mathbf{Q}$ such that $Q_1 \cup Q_2 \subseteq Q$. Then $\xi^{\mu(\beta)}(b^{\mu(\beta)}(Q)) \to \zeta(b(Q)) \in \mathrm{Int}(K)$. Thus $\xi^{\mu(\beta)}(b^{\mu(\beta)}(Q))$ belongs to $\mathrm{Int}(K)$ for large β, and then it follows from (*) that $b^{\mu(\beta)}(Q) = \sup I$ for large β. But then $b(Q) = \sup I$, and this implies that $\sigma = b(Q) = \sup I$, and $\sigma \in S$. Since $s \in I$, and $s \geq \sigma$, we must have $s = \sigma$, contradicting the fact that $s \notin S$. □

Remark 2.1. If the space Ω is σ-compact (i.e., a countable union of compact subsets), then instead of considering the product U defined in (2.4), we can use the product $U' = \prod_{j=1}^{\infty} \mathcal{A}(Q_j)$, where $\{Q_j\}$ is a sequence of compact subsets of $\Omega \times I$ whose relative interiors cover $\Omega \times I$. Then U' is metrizable. Therefore the statement and proof of Proposition 2.2 can be repeated using sequences instead of nets, and subsequences instead of subnets. □

We now specialize to the arc systems defined by vector fields. Let Ω be an open subset of $\mathbb{R}^n$, and let I be a subinterval of $\mathbb{R}$. A *time-varying vector field* on Ω with time domain I is a map $f : \Omega \times I \mapsto \mathbb{R}^n$. We use $TVVF(\Omega, I)$ to denote the set of all such maps. An $f \in TVVF(\Omega, I)$ satisfies the *Carathéodory condition* if it is continuous in x for each fixed $t \in I$ and Lebesgue measurable in t for each fixed $x \in \Omega$. We call f *locally integrably bounded* (LIB) if f is bounded in norm by a locally integrable function of t, as long as x stays in a compact subset of Ω. We use $TVVF_{Car,LIB}(\Omega, I)$ to denote the set of all $f \in TVVF(\Omega, I)$ that satisfy the Carathéodory condition, and are LIB.

If $f \in TVVF(\Omega, I)$, a *trajectory* (or *integral curve*) of f is an absolutely continuous function $\xi : J \mapsto \Omega$ defined on some nonempty subinterval $J \subseteq I$ such that $\dot{\xi}(t) = f(\xi(t), t)$ for almost all $t \in J$. A *maximal trajectory* of f is a trajectory $\xi : J \mapsto \Omega$ that cannot be extended to a trajectory of f defined on a strictly larger interval. We write $\mathrm{Traj}(f)$ to denote the set of all trajectories of f, and $\mathrm{Traj}_c(f)$ to denote the set of all $\xi \in \mathrm{Traj}(f)$ that are arcs, i.e., such that $\mathrm{Dom}(\xi)$ is a compact interval. Also, we write $\mathrm{Traj}(f)_{max}$ to denote the set of all maximal trajectories of f. Then it is clear that $\mathrm{Traj}_c(f)$ is an arc system, and $\mathrm{Traj}_c(f)^{\#}$ is precisely $\mathrm{Traj}(f)$. Therefore $\mathrm{Traj}(f)_{max}$ is exactly the set $\mathrm{Traj}_c(f)^{\#}_{max}$.

Thus $f \mapsto \mathrm{Traj}_c(f)$ is a map from $TVVF(\Omega, I)$ to $AS(\Omega, I)$. Using this map, we can pull back the topology of trajectory convergence on $AS(\Omega, I)$ and define a *topology of trajectory convergence*, or *T-convergence*, on $TVVF(\Omega, I)$. We can also pull back properties of arc systems and apply them to vector fields. (For example, we will say that f has the local existence property if the arc system $\mathrm{Traj}_c(f)$ does.) We call f *trajectory compact* if $\mathrm{Traj}_c(f)$ has the compactness property. We use $TVVF_{comp,le}(\Omega, I)$ to denote the set of those $f \in TVVF(\Omega, I)$ that are trajectory compact and have the local existence property. We say that f has the *uniqueness property* if for every initial condition $(\bar{x}, \bar{t})$ in $\Omega \times I$ the maximal trajectory of f whose graph contains $(\bar{x}, \bar{t})$ is unique.

The Carathéodory existence theorem (cf. [6]) implies that every vector field $f \in TVVF_{Car,LIB}(\Omega, I)$ has the local existence property. Moreover, a simple application of the Ascoli–Arzelà theorem and the Lebesgue dominated convergence theorem shows that every time-varying vector field $f \in TVVF_{Car,LIB}(\Omega, I)$ is trajectory compact. Thus

$$TVVF_{Car,LIB}(\Omega, I) \subseteq TVVF_{comp,le}(\Omega, I)\,. \tag{2.6}$$

If S is a subset of $\Omega \times I$, we write $\mathrm{Traj}_c(f, S)$ to denote the set of all arcs $\xi \in \mathrm{Traj}_c(f)$ such that $G(\xi) \subseteq S$. We then have the following propositions.

PROPOSITION 2.3. *Let f be in $TVVF(\Omega, I)$ and $\{f^\alpha\}_{\alpha\in A}$ be a net in $TVVF(\Omega, I)$. Then $\{f^\alpha\}_{\alpha\in A}$ T-converges to f iff for every compact subset K of $\Omega \times I$ and every open subset V of $ARC(\Omega)$ containing $\mathrm{Traj}_c(f, K)$ there exists an $\alpha^*(V) \in A$ such that $\mathrm{Traj}_c(f^\alpha, K) \subseteq V$ whenever $\alpha^*(V) \preceq \alpha$.*

PROPOSITION 2.4. *If $f \in TVVF_{comp,le}(\Omega, I)$, then a net $\{f^\alpha\}_{\alpha\in A}$ in $TVVF(\Omega, I)$ T-converges to f iff*

$$\lim_{\alpha} \sup\{d_{ARC}(\xi, \mathrm{Traj}_c(f, K)) : \xi \in \mathrm{Traj}_c(f^\alpha, K)\} = 0 \tag{2.7}$$

for every compact subset K of $\Omega \times I$.

PROPOSITION 2.5. *Let $\mathbf{f} = \{f^\alpha\}_{\alpha\in A}$ be a net in $TVVF_{comp,le}(\Omega, I)$, and suppose that $f \in TVVF_{comp,le}(\Omega, I)$. Then $\mathbf{f}$ T-converges to f iff, whenever $\{f^{\mu(\beta)}\}_{\beta\in B}$ is a subnet of $\mathbf{f}$, $(x^\beta, t^\beta) \in \Omega \times I$ converge to $(\bar{x}, \bar{t}) \in \Omega \times I$, and $\boldsymbol{\xi} = \{\xi^\beta\}_{\beta\in B}$ is a family such that each ξ^β is a maximal trajectory of $f^{\mu(\beta)}$ for which $\xi^{\mu(\beta)}(t^\beta) = x^\beta$, it follows that there exist a maximal trajectory ξ of f such that $\xi(\bar{t}) = \bar{x}$ and a subnet $\{\xi^{\nu(\gamma)}\}_{\gamma\in C}$ of $\boldsymbol{\xi}$ that converges to ξ on compact sets.*

PROPOSITION 2.6. *If $f \in TVVF_{comp,le}(\Omega, I)$, then a sequence $\mathbf{f} = \{f^j\}_{j=1}^\infty$ of members of $TVVF_{comp,le}(\Omega, I)$ T-converges to f iff, whenever J is an infinite set of positive integers, $(x^j, t^j) \in \Omega \times I$ converge to $(\bar{x}, \bar{t}) \in \Omega \times I$ as $j \to \infty$ via values in J, and $\boldsymbol{\xi} = \{\xi^j\}_{j\in J}$ is a family such that each ξ^j is a maximal trajectory of f^j for which $\xi^j(t^j) = x^j$, it follows that there exist a maximal trajectory ξ of f such that $\xi(\bar{t}) = \bar{x}$ and an infinite subset J' of J such that $\{\xi^j\}_{j\in J'}$ converges to ξ on compact sets.*

PROPOSITION 2.7. *Suppose that $f \in TVVF_{comp,le}(\Omega, I)$ and that a net $\mathbf{f} = \{f^\alpha\}_{\alpha\in A}$ of members of $TVVF_{comp,le}(\Omega, I)$ T-converges to f; assume that $(x^\alpha, t^\alpha) \in \Omega \times I$ converge to $(\bar{x}, \bar{t}) \in \Omega \times I$, $\boldsymbol{\xi} = \{\xi^\alpha\}_{\alpha\in A}$ is a family such that each ξ^α is a maximal trajectory of f^α for which $\xi^\alpha(t^\alpha) = x^\alpha$, and the maximal trajectory ξ of f for which $\xi(\bar{t}) = \bar{x}$ is unique; then the net $\boldsymbol{\xi}$ converges to ξ on compact sets.*

PROPOSITION 2.8. *Suppose that $f \in TVVF_{comp,le}(\Omega, I)$ has the uniqueness property, and let $\mathbf{f} = \{f^\alpha\}_{\alpha\in A}$ be a net of members of $TVVF_{comp,le}(\Omega, I)$; then $\mathbf{f}$ T-converges to f iff, whenever $(x^\alpha, t^\alpha) \in \Omega \times I$ converge to $(\bar{x}, \bar{t}) \in \Omega \times I$, $\boldsymbol{\xi} = \{\xi^\alpha\}_{\alpha\in A}$ is a family such that each ξ^α is a maximal trajectory of f^α for which $\xi^\alpha(t^\alpha) = x^\alpha$, and ξ is the maximal trajectory of f for which $\xi(\bar{t}) = \bar{x}$, it follows that the net $\boldsymbol{\xi}$ converges to ξ on compact sets.*

All these properties are very easy to prove. Proposition 2.3 is just a restatement of the definition of T-convergence. Then Proposition 2.4 immediately follows from Proposition 2.3 by observing that if $\mathrm{Traj}_c(f, K)$ is compact, then the sets $\{\xi \in ARC(\Omega) : d_{ARC}(\xi, \mathrm{Traj}_c(f, K)) < \varepsilon\}$ form a fundamental system of neighborhoods of $\mathrm{Traj}_c(f)$. Proposition 2.5 follows from Proposition 2.2, and Proposition 2.6 follows from Remark 2.1. Finally, Propositions 2.7 and 2.8 follow trivially from Proposition 2.5. $\square$

3. $\mathcal{T}^0$-convergence for control-affine systems. We are now ready to give the precise definition of the topology of $\mathcal{T}^0$-convergence on $L^1([0,T],\mathbb{R}^m)$.

Let m be a positive integer. A $C^0(m)$ *system* is an $(m+2)$-tuple

$$\Sigma = (\Omega, f_0, \dots, f_m)\,, \tag{3.1}$$

where Ω is a nonempty open subset of $\mathbb{R}^n$ for some integer $n \geq 1$, and $f_0,\dots,f_m$ are continuous vector fields on Ω. The expressions f_i^Σ, Ω^Σ will be used to denote the vector fields f_i and the set Ω corresponding to a given system $\Sigma \in C^0(m)$.

If $T > 0$, and $u \in L^1([0,T],\mathbb{R}^m)$, then for each $\Sigma = (\Omega^\Sigma, f_0^\Sigma, \dots, f_m^\Sigma)$, we can consider the ordinary differential equation

$$\dot{x}(t) = f_0^\Sigma(x(t)) + \sum_{k=1}^m u_k(t) f_k^\Sigma(x(t))\,. \tag{3.2}$$

The right-hand side of (3.2) is a time-varying vector field on Ω^Σ with time domain $[0,T]$, which clearly belongs to $TVVF_{Car,LIB}(\Omega^\Sigma,[0,T])$. Thus we can associate with Σ the map

$$\Gamma_\Sigma\,:\, L^1([0,T],\mathbb{R}^m) \mapsto TVVF_{Car,LIB}(\Omega^\Sigma,[0,T]) \tag{3.3}$$

defined by

$$\Gamma_\Sigma(u)(x,t) = f^\Sigma(u)(x,t) \overset{\text{def}}{=} f_0^\Sigma(x) + u_1(t) f_1^\Sigma(x) + \cdots + u_m(t) f_m^\Sigma(x)\,. \tag{3.4}$$

For each $\Sigma \in C^0(m)$ and $T > 0$, the set $TVVF_{Car,LIB}(\Omega^\Sigma,[0,T])$ is a topological space with the topology of trajectory convergence. Therefore, for any subclass $\mathcal{C}$ of $C^0(m)$ we can consider the topology $T(\mathcal{C})$ on $L^1([0,T],\mathbb{R}^m)$ induced by the family of mappings $\{\Gamma_\Sigma : \Sigma \in \mathcal{C}\}$, i.e., the weakest topology on $L^1([0,T],\mathbb{R}^m)$ that makes all the maps Γ_Σ continuous, for all $\Sigma \in \mathcal{C}$. In particular, the *topology of* $\mathcal{T}^0$*-convergence* on $L^1([0,T],\mathbb{R}^m)$ is, by definition, the $T(C^0(m))$-topology. We say that a net $\{u^\alpha\}_{\alpha\in A}$ $\mathcal{T}^0$*-converges* to a u^∞ in $L^1([0,T],\mathbb{R}^m)$ if it converges to u^∞ in this topology. Equivalently, a net $\{u^\alpha\}_{\alpha\in A}$ $\mathcal{T}^0$-converges to a limit u^∞ iff

(TCN) for every possible choice of the system $\Sigma \in C^0(m)$, the time-varying vector $(x,t) \mapsto f^\Sigma(u^\alpha)(x,t)$ T-converge to $(x,t) \mapsto f^\Sigma(u^\infty)(x,t)$.

(This is exactly the version for nets of condition (TC) of the introduction.)

If (TCN) holds, then Proposition 2.7 implies that in particular the following conditions hold as well:

(TCN′) For every possible choice of $\Sigma \in C^0(m)$, if ξ^α are maximal trajectories of (1.1) determined by the u^α, $x^\alpha = \xi^\alpha(0)$, and $x^\alpha \to \bar{x} \in \Omega^\Sigma$, then the ξ^α are defined on $[0,T]$ for α large enough and converge uniformly on $[0,T]$ to the maximal trajectory ξ^∞ for u^∞, provided that for the limiting initial value problem $\dot{x} = f_0^\Sigma(x) + \sum_{i=1}^m u_i^\infty(t) f_i^\Sigma(x)$, $x(0) = \bar{x}$, there is uniqueness of solutions as well as global existence on $[0,T]$.

(TCN″) For every possible choice of the system $\Sigma \in C^0(m)$ and the initial condition $x(0) = \bar{x} \in \Omega^\Sigma$, if ξ^α are maximal trajectories of (1.1) determined by the u^α and $\bar{x}$, then the ξ^α are defined on $[0,T]$ for α large enough and converge uniformly on $[0,T]$ to the maximal trajectory ξ^∞ for u^∞, provided that for the limiting initial value problem $\dot{x} = f_0^\Sigma(x) + \sum_{i=1}^m u_i^\infty(t) f_i^\Sigma(x)$, $x(0) = \bar{x}$, there is uniqueness of solutions as well as global existence on $[0,T]$.

4. A characterization of sequential $\mathcal{T}^0$-convergence.

THEOREM 4.1. *Let* $\mathbf{u} = \{u^j\}_{j=1}^{\infty}$ *be a sequence of controls in* $L^1([0,T],\mathbb{R}^m)$*, and let* $u^\infty \in L^1([0,T],\mathbb{R}^m)$*. Then the following conditions are equivalent:*

(i) $\mathbf{u}$ *satisfies condition* (UB) *and I-converges to* u^∞*,*
(ii) $\mathbf{u}$ $\mathcal{T}^0$*-converges to* u^∞*,*
(iii) (TCN′) *holds,*
(iv) (TCN″) *holds,*
(v) $\mathbf{u}$ *UW-converges to* u^∞*.*

If $\mathbf{u} = \{u^\alpha\}_{\alpha\in A}$ *is a net in* $L^1([0,T],\mathbb{R}^m)$*, and* u^∞ *belongs to* $L^1([0,T],\mathbb{R}^m)$*, then* (i)$\Longrightarrow$(ii)$\Longrightarrow$(iii)$\Longrightarrow$(iv)$\Longrightarrow$(v)*, but in general* (v)$\not\Longrightarrow$(iv) *and* (ii)$\not\Longrightarrow$(i)*.*

Remark 4.1. At present, we do not know if the implications (iv)$\Longrightarrow$(iii)$\Longrightarrow$(ii) hold for general nets, i.e., whether (ii), (iii), and (iv) of Theorem 4.1 are equivalent for nets.

THEOREM 4.2. *Let* $\mathbf{u} = \{u^\alpha\}_{\alpha\in A}$ *be a net of inputs belonging to* $L^1([0,T],\mathbb{R}^m)$*, and let* $u^\infty \in L^1([0,T],\mathbb{R}^m)$ *be such that* (i) *of Theorem* 4.1 *holds. Let* Ω *be open in* $\mathbb{R}^n$*, and let* $h^\alpha : \Omega\times[0,T] \mapsto \mathbb{R}^n$ *be time-varying vector fields given by*

$$h^\alpha(x,t) = \dot\theta^\alpha(t) + f_0(x) + \sum_{i=1}^m u_i^\alpha(t) f_i(x)\,,$$

where the $f_i : \Omega \mapsto \mathbb{R}^n$ *are continuous, the* $\theta^\alpha : [0,T] \mapsto \mathbb{R}^n$ *are absolutely continuous, and* $\theta^\alpha \to 0$ *uniformly. Then the net* $\{h^\alpha\}_{\alpha\in A}$ *T-converges to the vector field* h^∞ *given by* $h^\infty(x,t) = f_0(x) + \sum_{i=1}^m u_i^\infty(t) f_i(x)$*.*

Proof of Theorems 4.1 *and* 4.2. We first prove that (i) implies (v). We have to show that (UB) $\wedge$ (IC) implies (UWC). Pick a function $\varphi\in C^0([0,T],\mathbb{R})$ and a $\delta>0$, and find a piecewise constant function $\psi=\sum_{k=1}^N \psi_k\chi_{J_k}$ (where the J_k are subintervals of $[0,T]$ and χ_J denotes the indicator function of J) such that $|\varphi(t)-\psi(t)|<\delta$ for all $t \in [0,T]$. If (UB) holds, then there exists α^* and a $C > 0$ such that $\|u_i^\alpha\|_{L^1} \le C$ for all $i = 1,\ldots,m$, and $\alpha \succeq_A \alpha^*$ including $\alpha = \infty$. But then $|\int_0^t u_i^\alpha(s)\varphi(s)ds - \int_0^t u_i^\alpha(s)\psi(s)ds| \le C\delta$ for all i, t, and $\alpha \succeq_A \alpha^*$. In view of (IC), $\int_0^t u_i^\alpha(s)\psi(s)ds \to \int_0^t u_i^\infty(s)\psi(s)ds$ uniformly with respect to t. So $|\int_0^t u_i^\alpha(s)\psi(s)ds - \int_0^t u_i^\infty(s)\psi(s)ds| \le \delta$ for all t, if α is large enough. But then $|\int_0^t u_i^\alpha(s)\varphi(s)ds - \int_0^t u_i^\infty(s)\varphi(s)ds| \le (2C+1)\delta$ for all t if α is sufficiently large. Since δ is arbitrary, (UWC) follows.

Next, we prove Theorem 4.2, which contains as a particular case the implication (i)$\Longrightarrow$(ii) of Theorem 4.1. Let $\mathbf{u} = \{u^\alpha\}_{\alpha\in A}$, and let $\Omega, f_0, \ldots, f_m$, and the θ^α be as in the statement of Theorem 4.2. Let $\Sigma = (\Omega, f_0, \ldots, f_m)$ be a $C^0(m)$ system. We have to show that the h^α T-converge to h^∞. Let K be a compact subset of $\Omega\times[0,T]$. We will show that the sets $\mathrm{Traj}_c(h^\alpha,K)$ converge to $\mathrm{Traj}_c(h^\infty,K)$. For this we first prove the following:

(CTR) If $\{\xi^\alpha\}_{\alpha\in A}$ is a net in $ARC(\Omega)$ such that $\xi^\alpha \in \mathrm{Traj}_c(h^\alpha,K)$, then $\{\xi^\alpha\}_{\alpha\in A}$ has a subnet $\{\xi^{\nu(\beta)}\}_{\beta\in B}$ that converges to an $\eta \in \mathrm{Traj}_c(h^\infty,K)$.

Clearly, if we show that $\{\xi^\alpha\}_{\alpha\in A}$ has a subnet that converges in $ARC(\Omega)$, then (CTR) would follow. Indeed, let $\{\xi^{\nu(\beta)}\}_{\beta\in B}$ be a subnet of $\{\xi^\alpha\}_{\alpha\in A}$ that converges in $ARC(\Omega)$. Let η be the limit. Let $I^\alpha = \mathrm{Dom}(\xi^\alpha)$ and $I = \mathrm{Dom}(\eta)$. If I just contains one point, then η is in $\mathrm{Traj}_c(h^\infty,K)$ by definition. Otherwise, for β large enough, $I^{\nu(\beta)} \cap I \neq \emptyset$. Write $I^\alpha = [a^\alpha,b^\alpha]$, and let $I = [a,b]$. Then $a^{\nu(\beta)} \to a$ and $\xi^{\nu(\beta)}(a^{\nu(\beta)}) \to \eta(a)$.

Let

$$\sigma^\alpha(t) = \xi^\alpha(a^\alpha) + \theta^\alpha(t) - \theta^\alpha(a^\alpha)\,. \tag{4.1}$$

We then have, for all $t \in I^{\nu(\beta)} \cap I$,

$$\begin{aligned} \xi^{\nu(\beta)}(t) = \sigma^{\nu(\beta)}(t) &+ \int_{a^{\nu(\beta)}}^t f_0\Big(\xi^{\nu(\beta)}(s)\Big)\,ds + \sum_{k=1}^m \int_{a^{\nu(\beta)}}^t u_k^{\nu(\beta)}(s) f_k\left(\eta(s)\right)\,ds \\ &+ \sum_{k=1}^m \int_{a^{\nu(\beta)}}^t u_k^{\nu(\beta)}(s) \Big\{ f_k\Big(\xi^{\nu(\beta)}(s)\Big) - f_k\left(\eta(s)\right)\Big\}\,ds\,. \end{aligned} \tag{4.2}$$

Then (4.2) implies (since $f_k\left(\xi^{\nu(\beta)}(s)\right) - f_k\left(\eta(s)\right) \to 0$ uniformly, the L^1 norms of the u^α are bounded by a fixed constant for large α, (UWC) holds, and $\theta^{\nu(\beta)}(t) - \theta^{\nu(\beta)}(a^\beta) \to 0$ uniformly) that, for $t \in I$,

$$\eta(t) = \eta(a) + \int_a^t f_0\left(\eta(s)\right)\,ds + \sum_{k=1}^m \int_a^t u_k^\infty(s) f_k\left(\eta(s)\right)\,ds\,.$$

Thus η is a trajectory of h^∞. Clearly the graph $G(\eta)$ of η is contained in K. Therefore $\eta \in \mathrm{Traj}_c(h^\infty, K)$ and then (CTR) follows.

We now show that the net $\{\xi^\alpha\}_{\alpha\in A}$ has a subnet that converges in $ARC(\Omega)$. Again let $I^\alpha = [a^\alpha, b^\alpha] = \mathrm{Dom}(\xi^\alpha)$. Then for $t \in I^\alpha$, we have

$$\begin{aligned} \xi^\alpha(t) &= \theta^\alpha(t) - \theta^\alpha(a^\alpha) + \xi^\alpha(a^\alpha) + \int_{a^\alpha}^t f_0\Big(\xi^\alpha(s)\Big)ds + \sum_{k=1}^m \int_{a^\alpha}^t u_k^\alpha(s) f_k\Big(\xi^\alpha(s)\Big)ds \\ &= \theta^\alpha(t) - \theta^\alpha(a^\alpha) + \xi^\alpha(a^\alpha) + \int_{a^\alpha}^t f_0\Big(\xi^\alpha(s)\Big)ds + \sum_{k=1}^m \int_{a^\alpha}^t u_k^\infty(s) f_k\Big(\xi^\alpha(s)\Big)ds \\ &\quad + \sum_{k=1}^m \int_{a^\alpha}^t \Big(u_k^\alpha(s) - u_k^\infty(s)\Big) f_k\Big(\xi^\alpha(s)\Big)ds\,. \end{aligned}$$

Let

$$\zeta^\alpha(t) = \xi^\alpha(a^\alpha) + \int_{a^\alpha}^t f_0\Big(\xi^\alpha(s)\Big)ds + \sum_{k=1}^m \int_{a^\alpha}^t u_k^\infty(s) f_k\Big(\xi^\alpha(s)\Big)ds\,.$$

It is obvious that the set $\{\zeta^\alpha : \alpha \in A\}$ is equicontinuous and uniformly bounded. The $\theta^\alpha(t) - \theta^\alpha(a^\alpha)$ converge to 0 uniformly by our assumption. Thus in order to show that $\{\xi^\alpha\}_{\alpha\in A}$ has a convergent subnet, we need only show that for each $k \in \{1, \ldots, m\}$, the integrals $\int_{a^\alpha}^t \left(u_k^\alpha(s) - u_k^\infty(s)\right) f_k\left(\xi^\alpha(s)\right) ds$ converge to 0 uniformly with respect to t.

Fix an integer $k \in \{1, \ldots, m\}$. Write $\hat{u}_k^\alpha = u_k^\alpha - u_k^\infty$. Then (UB) implies that there exist an $\alpha^* \in A$ and a constant $C > 0$ such that $\|\hat{u}^\alpha\|_{L^1} \le C$ for $\alpha \succeq_A \alpha^*$. Let

$$\hat{\xi}^\alpha(t) = \xi^\alpha(t) - \theta^\alpha(t)\,. \tag{4.3}$$

Then $\hat{\xi}^\alpha(t) - \xi^\alpha(t)$ goes to 0 uniformly, so $f_k\big(\hat{\xi}^\alpha(t)\big) - f_k\left(\xi^\alpha(t)\right)$ goes to 0 uniformly for each k. Since the L^1 norms of the functions $\hat{u}_k^\alpha$ are bounded, we can conclude that the integrals

$$\int_{a^\alpha}^t \Big(u_k^\alpha(s) - u_k^\infty(s)\Big)\Big(f_k\Big(\hat{\xi}^\alpha(s)\Big) - f_k\Big(\xi^\alpha(s)\Big)\Big)ds$$

converge to 0 uniformly with respect to t. So it suffices to show that

$$\int_{a^\alpha}^{t}\Big(u_k^\alpha(s)-u_k^\infty(s)\Big)f_k\Big(\hat{\xi}^\alpha(s)\Big)ds\to 0\ \text{uniformly}\,. \tag{4.4}$$

Let $\mathcal{K}$ be a compact subset of Ω such that $K\subseteq\mathcal{K}\times[0,T]$. For any given $\varepsilon>0$, let $g_k:\Omega\mapsto\mathbb{R}^n$ be vector fields of class C^∞ with compact support such that

$$\sup_{x\in\mathcal{K}}\|f_k(x)-g_k(x)\|<\frac{\varepsilon}{C}\,. \tag{4.5}$$

Then

$$\begin{aligned}\int_{a^\alpha}^{t}\hat{u}_k^\alpha(s)f_k\Big(\hat{\xi}^\alpha(s)\Big)ds&=\int_{a^\alpha}^{t}\hat{u}_k^\alpha(s)g_k\Big(\hat{\xi}^\alpha(s)\Big)\,ds\\&\quad+\int_{a^\alpha}^{t}\hat{u}_k^\alpha(s)\Big(f_k\Big(\hat{\xi}^\alpha(s)\Big)-g_k\Big(\hat{\xi}^\alpha(s)\Big)\Big)\,ds\,.\end{aligned} \tag{4.6}$$

The sup norm of the second term on the right-hand side of (4.6) is less than ε. Via integration by parts, we can rewrite the first term as

$$\hat{U}_k^\alpha(t)\cdot g_k\Big(\hat{\xi}^\alpha(t)\Big)-\int_{a^\alpha}^{t}\hat{U}_k^\alpha(s)\frac{d}{ds}\Big(g_k(\hat{\xi}^\alpha(s))\Big)\,ds\,, \tag{4.7}$$

where

$$\hat{U}_k^\alpha(t)=\int_{a^\alpha}^{t}\hat{u}_k^\alpha(s)\,ds\,. \tag{4.8}$$

Since $\hat{U}_k^\alpha(t)\to 0$ uniformly, the g_k are uniformly bounded, and the functions $t\mapsto\frac{d}{dt}\big(g_k(\hat{\xi}^\alpha(t))\big)$ are uniformly bounded in L^1 for large α, because

$$\frac{d}{dt}\Big(g_k(\hat{\xi}^\alpha(t))\Big)=\frac{\partial g_k}{\partial x}(\hat{\xi}^\alpha(t))\left(f_0(\xi^\alpha(t))+\sum_{\ell=1}^{m}u_\ell^\alpha(t)f_\ell(\xi^\alpha(t))\right)\,, \tag{4.9}$$

we can conclude that the integrals $\int_{a^\alpha}^{t}\hat{u}_k^\alpha(s)g_k\big(\hat{\xi}^\alpha(s)\big)\,ds$ go to zero uniformly. This implies that

$$\limsup_{\alpha}\left\|\int_{a^\alpha}^{t}\hat{u}_k^\alpha(s)f_k\Big(\hat{\xi}^\alpha(s)\Big)ds\right\|\le\varepsilon\,. \tag{4.10}$$

Since ε is arbitrary, we have shown that $\int_{a^\alpha}^{t}\hat{u}_k^\alpha(s)f_k\big(\hat{\xi}^\alpha(s)\big)\,ds\to 0$ uniformly. As explained before, this shows that $\{\xi^\alpha\}_{\alpha\in A}$ has a convergent subnet in $ARC(\Omega)$.

The above argument clearly applies to any subnet of $\{h^\alpha\}$. Thus we have actually proved the following:

(CTR1) For every compact subset K of $\Omega\times I$, if we are given a subnet $\{h^{\nu(\beta)}\}_{\beta\in B}$ of $\{h^\alpha\}_{\alpha\in A}$ and a family $\{\xi^\beta\}_{\beta\in B}\subseteq ARC(\Omega)$ with the property that $\xi^\beta\in\mathrm{Traj}_c(h^{\nu(\beta)},K)$, then the net $\{\xi^\beta\}_{\beta\in B}$ has a subnet $\{\xi^{\mu(\gamma)}\}_{\gamma\in C}$ that converges to an $\eta\in\mathrm{Traj}_c(h^\infty,K)$.

It follows easily from (CTR1) that the h^α T-converge to h^∞. Indeed, (CTR1) clearly implies that

$$\sup\{d_{ARC}(\xi, \mathrm{Traj}_c(h^\infty, K)) : \xi \in \mathrm{Traj}_c(h^\alpha, K)\} \to 0$$

for every K, and then the T-convergence of h^α to h^∞ follows from Proposition 2.4. This completes the proof of Theorem 4.2. In particular, we have established that (i) of Theorem 4.1 implies (ii).

We now continue with the proof of Theorem 4.1. The implication (ii)$\Longrightarrow$(iii) follows from Proposition 2.7, and (iii) implies (iv) trivially. We now prove that (iv) implies (v). To see this, let $\varphi : [0,T] \mapsto \mathbb{R}$ be continuous. Consider the $C^0(m)$ system $\Sigma = (\mathbb{R}^{m+1}, f_0, \ldots, f_m)$ where—using $(x_0, x_1, \ldots, x_m)$ to denote the standard coordinates in $\mathbb{R}^{m+1}$ and $(e_0, e_1, \ldots, e_m)$ to denote the canonical basis—$f_0(x) \equiv e_0$, $f_i(x) \equiv \varphi(x_0)e_i$ for $i = 1, \ldots, m$. Then the differential equation associated with Σ and u^α is $\dot{x}_0 = 1$, $\dot{x}_i = \varphi(x_0)u_i^\alpha$ for $i = 1, \ldots, m$. Take the initial condition $\bar{x} = (0, \ldots, 0)$. Then the components ξ_i^α of the solution, for $i = 1, \ldots, m$, are precisely the components of the vector U_φ^α. Thus (UWC) holds.

Next we prove that (v) implies (i) for sequences. Let $\{u^j\}_{j=1}^\infty$ be a sequence in $L^1([0,T], \mathbb{R}^m)$, and let $u^\infty \in L^1([0,T], \mathbb{R}^m)$. Assume that (UWC) holds for $\{u^j\}$ and u^∞. Then, in particular, the integrals $\langle u_i^j, \varphi\rangle = \int_0^T u_i^j(t)\varphi(t)dt$ converge to $\langle u_i^\infty, \varphi\rangle = \int_0^T u_i^\infty(t)\varphi(t)dt$ for each i and for each continuous function $\varphi : [0,T] \mapsto \mathbb{R}$. So the u_i^j weak*-converge to u_i^∞ as linear functionals on the Banach space $C^0([0,T], \mathbb{R})$ of continuous real-valued functions on $[0,T]$. It is well known that, if ψ belongs to $L^1([0,T], \mathbb{R})$, then the norm of the linear functional $C^0([0,T], \mathbb{R}) \ni \rho \mapsto \int_0^T \rho(t)\psi(t)dt$ is precisely the L^1 norm of ψ. Therefore the L^1 norms of the u_i^j are bounded by a fixed constant, by the uniform boundedness theorem. Thus (UB) holds. Moreover, taking $\varphi \equiv 1$ we conclude that the u^j I-converge to u^∞.

The proof that in general (v) $\not\Longrightarrow$ (iv) and (ii)$\not\Longrightarrow$ (i) is given in the appendix, where we exhibit an example of a net that satisfies (v) but not (iv) and one of a net that satisfies (ii) but not (i). □

5. Other classes of systems. The equivalence for sequences of the five conditions of Theorem 4.1 depends very strongly on the fact that the class of systems under consideration is $C^0(m)$. One could equally well have considered T($\mathcal{C}$)-convergence in $L^1([0,T], \mathbb{R}^m)$ with respect to other classes $\mathcal{C}$ of systems. For example, we could have taken $\mathcal{C}$ to be the class $C^k(m)$ of all the $\Sigma \in C^0(m)$ such that $f_0^\Sigma, \ldots, f_m^\Sigma$ are of class C^k on Ω^Σ, for an arbitrary integer $k > 0$. Let us call the T($C^k(m)$)-convergence in $L^1([0,T], \mathbb{R}^m)$ simply *$\mathcal{T}^k$-convergence.*

It is then natural to ask whether explicit characterizations of $\mathcal{T}^k$-convergence of sequences exist. The necessary and sufficient conditions for $\mathcal{T}^0$-convergence are of course sufficient to ensure $\mathcal{T}^k$-convergence, but they are no longer necessary. For example, if a sequence $\{u^j\}$ $\mathcal{T}^k$-converges to u^∞ in $L^1([0,T], \mathbb{R}^m)$ for some $k \geq 1$, then we can no longer apply the uniform boundedness theorem to conclude that the L^1 norms of the u^j are bounded. The following theorem, which characterizes $\mathcal{T}^k$-convergence of nets in $L^1([0,T], \mathbb{R})$ for $k \geq 1$, implies in particular that it is indeed possible for a sequence $\{u^j\}_{j=1}^\infty \subseteq L^1([0,T], \mathbb{R})$ to $\mathcal{T}^1$-converge to a $u^\infty \in L^1([0,T], \mathbb{R})$ even if the $\|u^j\|_{L^1}$ are not bounded.

THEOREM 5.1. *Let $T > 0$, and let $k > 0$ be an integer. A net $\{u^\alpha\}_{\alpha \in A}$ in $L^1([0,T], \mathbb{R})$ $\mathcal{T}^k$-converges to a $u^\infty \in L^1([0,T], \mathbb{R})$ if and only if $\{u^\alpha\}$ I-converges to u^∞, i.e., iff the integrals $\int_0^t u^\alpha(s)ds$ converge to $\int_0^t u^\infty(s)ds$ uniformly on $[0,T]$.*

Before proving Theorem 5.1, it is natural to ask whether a similar characterization might be valid for $m > 1$. The answer is negative, as will be shown by an example in the next section. The reason for this is by now well understood. The obstructions for the validity of Theorem 5.1 when $m > 1$ are the Lie brackets of the vector fields f_i for $i > 0$. Precisely, call a system $\Sigma \in C^1(m)$ *commutative* if the Lie brackets $[f_i^\Sigma, f_j^\Sigma]$ vanish identically for all $i, j \in \{1, \dots, m\}$. (We do *not* require that the brackets involving f_0 vanish. Therefore for $m = 1$ *every* system is commutative.) For $k \geq 1$, let $C^k_{comm}(m)$ be the class of all commutative systems in $C^k(m)$. A net $\mathbf{u}$ in $L^1([0,T], \mathbb{R}^m)$ *$\mathcal{T}^k_{comm}$-converges* to $u \in L^1([0,T], \mathbb{R}^m)$ if it converges to u in the topology $\mathcal{T}(C^k_{comm}(m))$, i.e., if the time-varying vector fields $f^\Sigma(u^\alpha)$ T-converge to $f^\Sigma(u)$ for all $\Sigma \in C^k_{comm}(m)$.

THEOREM 5.2. *Let $T > 0$, and let m be a positive integer. For any integer $k \geq 1$, a net $\{u^\alpha\}_{\alpha \in A}$ in $L^1([0,T], \mathbb{R}^m)$ $\mathcal{T}^k_{comm}$-converges to a $u^\infty \in L^1([0,T], \mathbb{R}^m)$ iff $\{u^\alpha\}$ I-converges to u^∞, i.e., iff the integrals $\int_0^t u^\alpha(s)ds$ converge to $\int_0^t u^\infty(s)ds$ uniformly on $[0,T]$.*

Since for $m = 1$ every C^k system is commutative, Theorem 5.1 is a particular case of Theorem 5.2.

Proof of Theorem 5.2. First of all, we observe that $\mathcal{T}^k_{comm}$-convergence obviously implies I-convergence. To show the converse, we assume that $\mathbf{u} = \{u^\alpha\}_{\alpha \in A}$ I-converges to u^∞ and seek to prove that the net $\{f^\Sigma(u^\alpha)\}_{\alpha \in A}$ T-converges to $f^\Sigma(u^\infty)$ for every $C^1_{comm}(m)$ system Σ. We consider first the special case of systems

$$\Sigma : \quad \dot{x} = f_0(x) + \sum_{i=1}^m u_i b_i = f_0(x) + u \cdot B \,, \; x \in \Omega \,, \tag{5.1}$$

where f_0 is continuous (but not necessarily of class C^1) and the vector fields $f_1, \dots, f_m$ are constant. Let $U_i^\alpha(t) = \int_a^t u_i^\alpha(s)\, ds$ for $\alpha \in A \cup \{\infty\}$, so $U^\alpha \to U^\infty$ uniformly on $[0,T]$. Let $\xi^\alpha : [a^\alpha, b^\alpha] \mapsto \Omega$ be trajectories of Σ for the u^α, such that $G(\xi^\alpha) \subseteq K$ for some compact subset K of $\Omega \times [0,T]$. We want to find a subnet of $\{\xi^\alpha\}_{\alpha \in A}$ that converges to a $\xi^\infty \in \mathrm{Traj}_c(f^\Sigma(u^\infty), K)$. By passing to a subnet, we may assume that a^α, b^α have limits a, b, and $x^\alpha = \xi^\alpha(a^\alpha)$ converges to a limit $\bar{x}$. Then $(\bar{x}, a) \in K$. Since

$$\xi^\alpha(t) = x^\alpha + \int_{a^\alpha}^t f_0(\xi^\alpha(s))\, ds + (U^\alpha(t) - U^\alpha(a^\alpha)) \cdot B \,, \tag{5.2}$$

and the U^α converge uniformly to U^∞, it is clear that $\{\xi^\alpha\}$ is equicontinuous. Thus by passing to a subnet we may assume that ξ^α converges uniformly to a limit ξ^∞. Then a trivial argument allows us to pass to the limit in (5.2) and conclude that ξ^∞ is a trajectory for U^∞. The above argument clearly applies to any subnet of $\{\xi^\alpha\}_{\alpha \in A}$. Then Proposition 2.4 implies that the $f^\Sigma(u^\alpha)$ T-converge to $f^\Sigma(u^\infty)$ in this special case.

We now consider the case when Σ is such that the vector fields $f_1^\Sigma, \dots, f_m^\Sigma$ are linearly independent at each point. In this case, each point $x \in \Omega^\Sigma$ has an open neighborhood U_x such that there is a C^1 diffeomorphism Φ_x from U_x to some other open subset V_x of $\mathbb{R}^n$ that transforms the f_i^Σ, for $i = 1, \dots, m$, to constant vector fields $g_{i,x}$. Let $g_{0,x}$ be the vector field on V_x that corresponds to $f_0^\Sigma \lceil U_x$ under Φ_x, that is, $g_{0,x}(y) = D\Phi_x(z).f_0^\Sigma(z)$, where $z = \Phi_x^{-1}(y)$. (Notice that $g_{0,x}$ is continuous but not necessarily of class C^1.) Let Σ_x be the restriction of Σ to U_x and $\tilde{\Sigma}_x$ be the system $(V_x, g_{0,x}, g_{1,x}, \dots, g_{m,x})$. Then $\tilde{\Sigma}_x$ is a $C^0(m)$ system. Since the vector fields $g_{1,x}, \ldots, g_{m,x}$ are constant, we know that $f^{\tilde{\Sigma}_x}(u^\alpha) \xrightarrow{\mathrm{T}} f^{\tilde{\Sigma}_x}(u^\infty)$. Since T-convergence is preserved by C^1 diffeomorphisms, we conclude that $f^{\Sigma_x}(u^\alpha) \xrightarrow{\mathrm{T}} f^{\Sigma_x}$ (u^∞). This says that our desired conclusion holds locally, in the sense that every $x \in \Omega$ has a neighborhood U_x such that the conclusion holds for the restriction of Σ to U_x. From this, we have to infer that the conclusion also holds globally. It suffices to show that, if (a) $\{\xi^\alpha\}_{\alpha \in A}$ is a net of maximal trajectories of the $f^\Sigma(u^\alpha)$, (b) $\{(x^\alpha, t^\alpha)\}_{\alpha \in A}$ is a net of points in $\Omega \times [0,T]$ such that $\xi^\alpha(t^\alpha) = x^\alpha$ for each α, and (c) $(x^\alpha, t^\alpha) \to (x^\infty, t^\infty) \in \Omega \times [0,T]$, then $\{\xi^\alpha\}_{\alpha \in A}$ converges on compact sets to the maximal trajectory ξ^∞ of $f^\Sigma(u^\infty)$ such that $\xi(t^\infty) = x^\infty$ (using the fact that Σ is a C^1 system and therefore has uniqueness of trajectories). For $\alpha \in A \cup \{\infty\}$, let $I^\alpha = \mathrm{Dom}(\xi^\alpha)$, so I^α is relatively open in $[0,T]$ and $t^\alpha \in I^\alpha$. Call a compact subinterval J of I^∞ "good" if there exists an $\alpha^*(J) \in A$ such that $J \subseteq I^\alpha$ whenever $\alpha \succeq_A \alpha^*(J)$, and the net $\{\xi^\alpha\}_{\alpha \in A; \alpha \succeq_A \alpha^*(J)}$ converges to ξ^∞ uniformly on J. We have to prove that every compact subinterval of I^∞ is good. It clearly suffices to consider intervals J of the form $[a, t^\infty]$ or $[t^\infty, b]$. The proof is similar in both cases, so we will only consider the second case and assume that $J = [t^\infty, b]$ for some $b \in [0,T]$. Pick an integer $N > 0$ such that $\frac{1}{N}$ is a Lebesgue number of the covering $\{(\xi^\infty)^{-1}(U_x) : x \in \Omega\}$ of J. Let $b_k = t^\infty + \frac{k}{N}(b - t^\infty)$ for $k = 0, 1, \ldots, N$, so $b_0 = t^\infty$ and $b_N = b$. Let $J_k = [b_0, b_k]$, so $J = J_N$. Let $L_k = [b_{k-1}, b_k]$ for $k = 1, \ldots, N$. Pick $x_k \in \Omega$ such that $\xi^\infty(L^k) \subseteq U_{x_k}$. Write $\Omega_k = U_{x_k}$, $\Sigma_k = \Sigma_{x_k}$. We prove by induction on k that J_k is good. First, we show that J_1 is good. Fix an $\varepsilon > 0$. For $\alpha \in A \cup \{\infty\}$, let η^α be the maximal trajectory of $f^{\Sigma_1}(u^\alpha)$ that goes through x^α at time t^α. Since $f^{\Sigma_1}(u^\alpha) \xrightarrow{\mathrm{T}} f^{\Sigma_1}(u^\infty)$, Σ_1 has uniqueness of trajectories, and $J_1 \subseteq \mathrm{Dom}(\eta^\infty)$, there exists an α^* such that $J_1 \subseteq \mathrm{Dom}(\eta^\alpha)$ and $\|\eta^\alpha(t) - \eta^\infty(t)\| < \varepsilon$ on J_1 whenever $\alpha \succeq_A \alpha^*$. On the other hand, η^α is a trajectory of $f^\Sigma(u^\alpha)$, so $\eta^\alpha = \xi^\alpha \lceil \mathrm{Dom}(\eta^\alpha)$. Therefore $J_1 \subseteq \mathrm{Dom}(\xi^\alpha)$ and $\|\xi^\alpha(t) - \xi^\infty(t)\| < \varepsilon$ for all $t \in J_1$ whenever $\alpha \succeq_A \alpha^*$. This establishes that J_1 is good. Now suppose that $0 < k \leq N$ and J_{k-1} is good. Pick α_1^* such that $J_{k-1} \subseteq I^\alpha$ whenever $\alpha \succeq_A \alpha_1^*$ and $\{\xi^\alpha\}_{\alpha \in A; \alpha \succeq_A \alpha_1^*}$ converges to ξ^∞ uniformly on J_{k-1}. Then $\{\xi^\alpha(b_{k-1})\}_{\alpha \in A; \alpha \succeq_A \alpha_1^*}$ converges to $\xi^\infty(b_{k-1})$. Thus there exists $\alpha_2^* \in A$ such that $\alpha_2^* \succeq_A \alpha_1^*$ and $\xi^\alpha(b_{k-1}) \in \Omega_k$ for $\alpha \succeq_A \alpha_2^*$. For $\alpha \succeq_A \alpha_2^*$, or $\alpha = \infty$, let η^α be the maximal trajectory of $f^{\Sigma_k}(u^\alpha)$ that goes through $\xi^\alpha(b_{k-1})$ at time b_{k-1}. Since $\xi^\infty \lceil L_k$ is a trajectory of $f^{\Sigma_k}(u^\infty)$, we have $L_k \subseteq \mathrm{Dom}(\eta^\infty)$. Since $f^{\Sigma_k}(u^\alpha) \xrightarrow{\mathrm{T}} f^{\Sigma_k}(u)$, Σ_k has uniqueness of trajectories, and $L_k \subseteq \mathrm{Dom}(\eta^\infty)$, there exists an α^* such that $\alpha^* \succeq_A \alpha_2^*$, $L_k \subseteq \mathrm{Dom}(\eta^\alpha)$ whenever $\alpha \succeq_A \alpha^*$, and $\eta^\alpha \to \eta^\infty$ uniformly on L_k. On the other hand, η^α is a trajectory of $f^\Sigma(u^\alpha)$ and $\eta^\alpha(b_{k-1}) = \xi^\alpha(b_{k-1})$, so $\eta^\alpha = \xi^\alpha \lceil \mathrm{Dom}(\eta^\alpha)$. Therefore $L_k \subseteq \mathrm{Dom}(\xi^\alpha)$ whenever $\alpha \succeq_A \alpha^*$, and $\xi^\alpha \to \xi^\infty$ uniformly on L_k. Since $\alpha^* \succeq_A \alpha_1^*$, we have in fact shown that $J_k \subseteq \mathrm{Dom}(\xi^\alpha)$ whenever $\alpha \succeq_A \alpha^*$, and $\xi^\alpha \to \xi^\infty$ uniformly on J_k. So J_k is good. This completes the inductive proof that all the J_k are good, and in particular $J = J_N$ is good. So $\xi^\alpha \to \xi^\infty$ on compact sets. Therefore $f^\Sigma(u^\alpha) \xrightarrow{\mathrm{T}} f^\Sigma(u^\infty)$.

We now remove the restriction that $f_1^\Sigma, \ldots, f_m^\Sigma$ be linearly independent at each point. Let $\Sigma = (\Omega, f_0, f_1, \ldots, f_m)$ be an arbitrary system in $C^1_{comm}(m)$. Let $\hat{\Omega} = \Omega \times \mathbb{R}^m$, and let $\hat{\Sigma}$ be the system on $\hat{\Omega}$ given by

$$\dot{x} = f_0(x) + u_1 f_1(x) + \cdots + u_m f_m(x)\,, \quad \dot{y}_1 = u_1, \ldots, \dot{y}_m = u_m\,. \tag{5.3}$$

Then $\hat{\Sigma} \in C^1_{comm}(m)$, and it is clear that

$$\Big(f^{\hat{\Sigma}}(u^\alpha) \xrightarrow{\mathrm{T}} f^{\hat{\Sigma}}(u^\infty)\Big) \Rightarrow \Big(f^\Sigma(u^\alpha) \xrightarrow{\mathrm{T}} f^\Sigma(u^\infty)\Big)\,. \tag{5.4}$$

Moreover, the vector fields $f_1^{\hat\Sigma}, \dots, f_m^{\hat\Sigma}$ are linearly dependent at each point. Thus we conclude that $f^{\hat\Sigma}(u^\alpha) \xrightarrow{\mathrm{T}} f^{\hat\Sigma}(u^\infty)$, and therefore $f^{\Sigma}(u^\alpha) \xrightarrow{\mathrm{T}} f^{\Sigma}(u^\infty)$. □

6. Beyond I-convergence: The role of Lie brackets. Theorem 5.2 characterizes the topology of $\mathcal{T}^1_{comm}$-convergence on the input space $L^1([0,T],\mathbb{R}^m)$ by showing that $\mathcal{T}^1_{comm}$-convergence is exactly equivalent to I-convergence. It is natural to ask whether I-convergence actually implies T-convergence for at least some noncommutative systems. The answer to this question turns out to be negative.

THEOREM 6.1. *Let $\mathcal{C}$ be any subclass of $C^1(m)$ that contains at least one noncommutative system. Then there exists a sequence $\{u^j\}_{j=1}^\infty$ in $L^1([0,T],\mathbb{R}^m)$ such that* (i) *$\{u^j\}$ I-converges to a limit $u^\infty \in L^1([0,T],\mathbb{R}^m)$,* (ii) *for every $\Sigma \in \mathcal{C}$ the time-varying vector fields $f^\Sigma(u^j)$ T-converge to a limit h^Σ, but* (iii) *it is not true that $h^\Sigma = f^\Sigma(u^\infty)$ for all $\Sigma \in \mathcal{C}$.*

This result—which will follow from Proposition 6.2—says that noncommutativity is the fundamental obstruction for I-convergence of the inputs to imply convergence of trajectories. The best way to understand why this is so is to think of the space $L^1([0,T],\mathbb{R}^m)$ of ordinary inputs as embedded in a larger space $\mathcal{GI}(m)$ of *generalized m*-dimensional inputs, whose elements have components not only for the m input channels corresponding to the vector fields $f_1, \dots, f_m$ but also in "Lie bracket channels" such as $[f_1, f_2]$. An "ordinary input" $u \in L^1([0,T],\mathbb{R}^m)$ gives rise, for each system Σ, to a time-varying vector field $f^\Sigma(u)$. Thus an ordinary input can be thought of as a "functional" that to every system Σ in a given class $\mathcal{C}$ assigns a time-varying vector field. It is then natural to extend the class of ordinary inputs by allowing more general functionals. By analogy with other constructions of spaces of "generalized functions" (e.g., Schwartz distributions), we should also require that such a functional be approximable by ordinary inputs. So we first define, for a class $\mathcal{C} \subseteq C^0(m)$ and an interval $[a,b]$, the space $\mathcal{PGI}(\mathcal{C},a,b)$ of all maps v that assign to every $\Sigma = (\Omega, f_0, \dots, f_m) \in \mathcal{C}$ a time-varying vector field $v_\Sigma \in TVVF(\Omega,[a,b])$, and call the elements of $\mathcal{PGI}(\mathcal{C},a,b)$ "pseudogeneralized inputs." Then $\mathcal{PGI}(\mathcal{C},a,b)$ is just the product $\Pi_{(\Omega,f_0,\dots,f_m)\in\mathcal{C}} TVVF(\Omega,[a,b])$, and we can topologize it by means of the product topology. An ordinary input $u \in L^1([a,b],\mathbb{R}^m)$ can then be regarded as a member of $\mathcal{PGI}(\mathcal{C},a,b)$ by associating to it the map—also labeled u—$\Sigma \mapsto f^\Sigma(u)$. A "generalized input" on $[a,b]$ for the class $\mathcal{C}$ is then a member of $\mathcal{PGI}(\mathcal{C},a,b)$ which is a limit of ordinary inputs. We use $\mathcal{GI}(\mathcal{C},a,b)$ to denote the set of all generalized inputs, so by construction $\mathcal{GI}(\mathcal{C},a,b)$ is a topological space in which the space $L^1([a,b],\mathbb{R}^m)$ is densely embedded.

Particular examples of members of $\mathcal{PGI}(\mathcal{C},a,b)$ can be constructed by means of *formal Lie bracket expressions* such as

$$X_0 + v_1(t)X_1 + v_2(t)X_2 + v_3(t)[X_1,X_2] + v_4(t)[X_2,[X_1,X_2]]\,.$$

Here the indeterminates X_i should be thought of as "slots" where, for each particular system Σ with vector fields $f_0, \dots, f_m$, the f_i are to be plugged in, thus giving rise to a time-varying vector field. Precisely, we can define a *formal pseudogeneralized input on* $[a,b]$ to be an integrable function with values in the free Lie algebra $L(m+1)$ in $m+1$ indeterminates $X_0, \dots, X_m$. (The field of scalars is of course $\mathbb{R}$.) Use $\mathcal{FPGI}(m,a,b)$ to denote the set of all such objects. Then a $v \in \mathcal{FPGI}(m,a,b)$ gives rise to a true pseudogeneralized input—also labeled v—for any class $\mathcal{C} \subseteq C^0(m)$ of control-affine systems that has enough differentiability for all the Lie brackets occurring in the formal expression to make sense. For example, to an ordinary input $u = (u_1, \dots, u_m)$ we associate the $L(m+1)$-valued function $t \mapsto X_0+u_1(t)X_1+\cdots+u_m(t)X_m$, and in this case it is possible to plug in $f_0,\ldots,f_m$ for $X_0,\ldots,X_m$ for every system $\mathcal{C} \in C^0(m)$. On the other hand, if $m \geq 2$, then an expression v such as $X_0+v_1(t)X_1+v_2(t)X_2+v_3(t)[X_1,X_2]$ will give rise to a member of $\mathcal{PGI}(\mathcal{C},a,b)$ only if $\mathcal{C}$ consists of systems $(\Omega,f_0,\ldots,f_m)$ for which the Lie bracket $[f_1,f_2]$ is well defined. This will be the case, in particular, if $\mathcal{C} \subseteq C^1(m)$.

Thus, for example, the formal pseudogeneralized input $w = X_0+\frac{1}{2}[X_1,X_2]$ gives rise to a pseudogeneralized input $w \in \mathcal{PGI}(C^1(2),a,b)$. With this terminology, the following result exhibits an example of a sequence $\{u^j\} \subseteq L^1([0,T],\mathbb{R}^2)$ that I-converges to zero but converges in $\mathcal{PGI}(C^1(2),0,T)$ to w. This shows that $w \in \mathcal{GI}(C^1(2),a,b)$, i.e., that w is a true generalized input. It is now clear why, when we look only at classes $\mathcal{C}$ of systems for which $[f_1,f_2]=0$, as we did in Theorem 5.2, the sequence $\{u^j\}$ $\mathcal{T}(\mathcal{C})$-converges to 0, whereas, as soon as $\mathcal{C}$ contains one system for which $[f_1,f_2] \neq 0$, the $\mathcal{T}(\mathcal{C})$-limit of $\{u^j\}$ is no longer 0. The reason is that the "true" limit is w, and w contains "Lie bracket components" that are seen only when w is tested against a noncommutative system.

PROPOSITION 6.2. *Let $T>0$. For $j=1,2,\ldots$, let $u^j \in L^1([0,T],\mathbb{R}^2)$ be the input given by*

$$u^j(t) = (j^{\frac{1}{2}}\cos jt, j^{\frac{1}{2}}\sin jt)\,. \tag{6.1}$$

Let $\Sigma = (\Omega,f_0,f_1,f_2)$ be a system of class C^1. Let

$$f^j(x,t) = f^{\Sigma}(u^j)(x,t) = f_0(x)+j^{\frac{1}{2}}\cos jt\, f_1(x)+j^{\frac{1}{2}}\sin jt\, f_2(x) \tag{6.2}$$

and

$$f^{\infty}(x,t) = f_0(x)+\frac{1}{2}[f_1,f_2](x) \tag{6.3}$$

for $(x,t) \in \Omega \times [0,T]$. Then the f^j T-converge to f^{∞} in $TVVF_{LIB}(\Omega,[0,T])$.

Proof. Suppose K is a compact subset of the product $\Omega \times [0,T]$. We have to show that $\mathrm{Traj}_c(f^j,K)$ converges to $\mathrm{Traj}_c(f^{\infty},K)$. Let $\{\xi^j\} \subseteq ARC(\Omega)$ be a sequence of functions such that $\xi^j \in \mathrm{Traj}_c(f^j,K)$ for each j. We will first show that there exists a subsequence of $\{\xi^j\}$ that converges to a $\xi \in \mathrm{Traj}_c(f^{\infty},K)$ in $ARC(\Omega)$. Let $I^j=[a^j,b^j]=\mathrm{Dom}(\xi^j)$. On I^j we have

$$\dot{\xi}^j(t) = f_0(\xi^j(t))+j^{\frac{1}{2}}\cos jt\, f_1(\xi^j(t))+j^{\frac{1}{2}}\sin jt\, f_2(\xi^j(t))\,. \tag{6.4}$$

But $j^{\frac{1}{2}}\cos jt\, f_1(\xi^j(t))$ is equal to

$$\dot{\theta}^j_1(t) - j^{-\frac{1}{2}}\sin jt\, g_1(\xi^j(t)) - \frac{1}{2}\sin 2jt\, g_2(\xi^j(t)) - \sin^2 jt\, g_3(\xi^j(t))\,,$$

where $g_1 = Df_1 \cdot f_0$, $g_2 = Df_1 \cdot f_1$, $g_3 = Df_1 \cdot f_2$, Df_1 denotes the Jacobian matrix of f_1, and $\theta^j_1 = j^{-\frac{1}{2}}\sin jt\, f_1(\xi^j(t))$. Similarly, $j^{\frac{1}{2}}\sin jt\, f_2(\xi^j(t))$ is equal to

$$\dot{\theta}^j_2(t) + j^{-\frac{1}{2}}\cos jt\, g_4(\xi^j(t)) + \cos^2 jt\, g_5(\xi^j(t)) + \frac{1}{2}\sin 2jt\, g_6(\xi^j(t))\,,$$

where $g_4 = Df_2 \cdot f_0$, $g_5 = Df_2 \cdot f_1$, $g_6 = Df_2 \cdot f_2$, and $\theta^j_2 = -j^{-\frac{1}{2}}\cos jt\, f_2(\xi^j(t))$. Thus, if we let

$$\begin{aligned} h^j(x,t) = {} & f_0(x)+\dot{\theta}^j_1(t)+\dot{\theta}^j_2(t) \\ & -j^{-\frac{1}{2}}\sin jt\, g_1(x) - \frac{1}{2}\sin 2jt\, g_2(x) - \sin^2 jt\, g_3(x) \\ & +j^{-\frac{1}{2}}\cos jt\, g_4(x) + \cos^2 jt\, g_5(x) + \frac{1}{2}\sin 2jt\, g_6(x)\,, \end{aligned}$$

then $\xi^j \in \mathrm{Traj}_c(h^j, K)$.

Theorem 4.2 then implies that the sequence of time-varying vector fields h^j T-converges to the vector field h^∞ given by

$$h^\infty(x,t) = f_0(x) + \frac{1}{2}(g_5(x) - g_3(x)) = f_0(x) + \frac{1}{2}[f_1, f_2](x)\,.$$

(We apply the theorem to the control system $\dot{x} = f_0(x) + \sum_{i=1}^{6} v_i g_i(x)$, and the controls v^j given by

$$v^j(t) = \Big(- j^{-\frac{1}{2}} \sin jt, -\frac{1}{2}\sin 2jt, -\sin^2 jt, j^{-\frac{1}{2}}\cos jt, \cos^2 jt, \frac{1}{2}\sin 2jt\Big)\,.$$

It is clear that the controls v^j $\mathcal{T}^0$-converge to $(0, 0, -\frac{1}{2}, 0, \frac{1}{2})$. Since $\theta^j \to 0$ uniformly, Theorem 4.2 applies.)

It is clear that $h^\infty = f^\infty$. So $\{\xi^j\}$ has a subsequence that converges to a $\xi^\infty \in \mathrm{Traj}_c(f^\infty, K)$.

The above arguments clearly apply to any subsequence of $\{u^j\}$. Then from Proposition 2.4 we conclude that the f^j T-converge to f^∞. □

Proof of Theorem 6.1. Let $\Sigma \in \mathcal{C}$ have vector fields $f_0, f_1, \ldots, f_m$ such that $[f_i, f_\ell] \neq 0$ for some $i, \ell \in \{1, \ldots, m\}$. Define inputs u^j by letting $u_k^j(t) = 0$ if $k \notin \{i, \ell\}$, $u_i^j(t) = j^{\frac{1}{2}}\cos jt$, $u_\ell^j(t) = j^{\frac{1}{2}}\sin jt$. By Proposition 6.2, the time-varying vector fields $f^\Sigma(u^j)$ T-converge to $f_0 + \frac{1}{2}[f_i, f_\ell]$. □

To understand the relationship between ordinary and generalized inputs, it is useful to keep in mind the following simple analogy. Integrable functions on, say, the interval $[0,1]$, are members of the more general class of Borel measures on $[0,1]$, modulo an obvious identification. A measure μ has a decomposition $\mu = \mu_{ac} + \mu_{sing}$ into an absolutely continuous part and a singular part. (We include in μ_{sing} all the atoms of μ.) If μ is absolutely continuous, i.e., if $d\mu = f\,dx$, where $f \in L^1$, then of course μ has no singular part. A sequence $\{\mu^j\}$ of absolutely continuous measures will converge to an absolutely continuous measure μ provided that the μ^j weak*-converge to μ and the functions $f^j = \frac{d\mu^j}{dx}$ are "kept under control," e.g., if the sequence $\{f^j\}$ is equi-integrable. If, on the other hand, the μ^j weak*-converge to μ but $\{f^j\}$ is not equi-integrable, then the limit μ can have a singular part even though the μ^j do not. Similarly, the general picture for ordinary differential equations $\dot{x} = f_0(x) + \sum_i u_i(t) f_i(x)$ is that each right-hand side is in fact a member of a larger class of "generalized" objects, which contain Lie bracket terms as well as "ordinary" terms. An "ordinary" right-hand side has no Lie bracket terms, but a sequence $\{f^j\}$ of ordinary right-hand sides can converge to a "generalized" one, which contains Lie bracket terms in addition to the averaged limit, unless $\{f^j\}$ is "kept under control," e.g., by means of the boundedness conditions of our Theorem 4.1.

There is a more general theory that shows in a systematic way how iterated Lie brackets can occur as limits of the right-hand sides of *highly oscillatory* differential equations, such as the ones considered in Proposition 6.2. The results described in Theorem 4.1 can be thought of as the zeroth level of this more general theory, in which Lie bracket terms do not occur in the limit. High-level results, in which the structure of general limits involving Lie bracket terms is analyzed in detail, are described in [13], [14], [15], [16], [22], [23], and [24]. (Some special cases of these limiting processes have also been studied by Kurzweil and Jarnik in [10] and [11].)

7. Comparison with other conditions.

We recall that a sequence $\{u^j\}_{j=1}^\infty$ of functions in $L^1([0,T], \mathbb{R}^m)$ *converges weakly* to a limit $u^\infty \in L^1([0,T], \mathbb{R}^m)$ if

$\lim_{j\to\infty}\int_0^T \varphi(t)u^j(t)dt = \int_0^T \varphi(t)u^\infty(t)dt$ for all functions $\varphi \in L^\infty([0,T],\mathbb{R})$. It is well known (cf. Theorem 3.1 on page 53 of [5]) that $\{u^j\}_{j=1}^\infty$ converges weakly to u^∞ iff (i) the integrals $U^j(t) = \int_0^t u^j(s)ds$ converge to $U^\infty(t) = \int_0^t u^\infty(s)ds$ for each t, and (ii) the sequence $\{u^j\}$ is *equi-integrable* in the sense that for every $\varepsilon > 0$ there is a $\delta > 0$ such that $\int_E \|u^j(t)\|dt \leq \varepsilon$ for every j and every measurable subset E of $[0,T]$ whose measure is $\leq \delta$.

It is clear that equi-integrability implies that the sequence $\{U^j\}$ is equicontinuous. So (i) and (ii) actually imply (IC). Moreover, the equi-integrability condition also implies (UB). Thus one particular consequence of our main result is that *weak convergence implies $\mathcal{T}^0$-convergence.* On the other hand, the following example shows that a sequence can satisfy the conditions of Theorem 4.1 without being weakly convergent, so our necessary and sufficient condition for sequences is strictly weaker than weak convergence.

Example 7.1. Let $m = 1$, $u^j(t) = j^{\frac{1}{2}}\cos jt$ for $0 \leq t \leq j^{-\frac{1}{2}}$, and $u^j(t) = 0$ for $j^{-\frac{1}{2}} < t \leq T$. Then $\{u^j\}_{j=1}^\infty$ satisfies (IC) and (UB), with $u^\infty \equiv 0$. It follows from Theorem 4.1 that $\{u^j\}_{j=1}^\infty$ $\mathcal{T}^0$-converges to u^∞.

On the other hand, it is not hard to see that *the sequence $\{u^j\}_{j=1}^\infty$ is not weakly convergent in $L^1[0,T]$*. Indeed, the integrals $\int_0^{j^{-1/2}} |u^j(t)|dt$ converge to $\frac{2}{\pi}$ as $j \to \infty$, and then $\{u^j\}$ is not equi-integrable. □

The preceding example also makes it possible to show that our convergence condition is weaker than those of Buttazzo–Conti, Kurzweil–Vorel, and Neustadt; cf. [4], [9], [19]. Indeed, the Kurzweil–Vorel conditions amount to saying that

$$\int_0^t u^j(s)ds \to \int_0^t u^\infty(s)ds \quad \text{and} \quad \sup_j |u^j| \in L^1[0,T]\,.$$

(Our example violates the latter condition because if we let $v(t) = \sup_j |u^j(t)|$, then $v(t) \geq j^{1/2}|\cos jt|$ on the interval $[0, j^{-1/2}]$. Thus

$$\int_0^{j^{-\frac{1}{2}}} v(t)dt \geq j^{1/2}\int_0^{j^{-1/2}} |\cos jt|dt.$$

Since the latter integral goes to $\frac{2}{\pi}$ as $j \to \infty$, we see that $\int_0^{j^{-1/2}} v(t)dt$ is bounded away from zero as $j \to \infty$. Therefore v is not integrable.) The Neustadt theorem requires that $\int_0^t u^j(s)ds \to \int_0^t u^\infty(s)ds$ and $\sup_j \int_t^\tau |u^j(s)|ds \overset{\tau\to t+}{\longrightarrow} 0$ uniformly in t. Our example violates this if $t = 0$. The Buttazzo–Conti theorem, when specialized to control-affine systems (1.1), requires that the vector fields $f_0, \ldots, f_m$ be Lipschitz, although they assume only that the u^j and u^∞ satisfy (IC) and (UB).

Appendix A. We give three examples. The first two pertain, respectively, to the assertions made in Theorem 4.1 that in general (v) $\not\Longrightarrow$(iv) and (ii)$\not\Longrightarrow$ (i). The third one illustrates the difference, for general vector fields, between "continuous dependence of the trajectories for every fixed initial condition," and "joint continuous dependence of the trajectories on the right-hand side and the initial condition."

As shown in Theorem 4.1, uniform weak convergence is necessary for a net $\{u^\alpha\}_{\alpha\in A}$ to $\mathcal{T}^0$-converge to u^∞ in $L^1([0,T],\mathbb{R}^m)$ and is also sufficient if $\{u^\alpha\}_{\alpha\in A}$ is a sequence. The following example shows that (UWC) is no longer sufficient in general for a net $\{u^\alpha\}$ to $\mathcal{T}^0$-converge to u^∞, and therefore proves that condition (v) of Theorem 4.1 does not in general imply (iv).

Example A.1. Let $T = 2\pi$. For every integer $n \geq 1$, every choice $\varphi_1, \ldots, \varphi_n$ of n real-valued continuous functions on $[0, T]$, and every $\varepsilon > 0$, set

$$V(\varphi_1, \ldots, \varphi_n, \varepsilon) = \left\{ \psi \in L^1([0,T], \mathbb{R}) \colon \sup_{0 \leq t \leq T} \left| \int_0^t \varphi_i(s)\psi(s)ds \right| < \varepsilon, i = 1, \ldots, n \right\}.$$

Let $\mathcal{V}$ be the set of all sets $V(\varphi_1, \ldots, \varphi_n, \varepsilon)$, for all possible integers $n \geq 1$, all n-tuples $(\varphi_1, \ldots, \varphi_n)$ of continuous functions on $[0, T]$, and all positive numbers ε. We define a directed set $(I, \preceq_I)$ by letting

$$I = \{(V, N) \ : \ V \in \mathcal{V}, N \text{ is an integer} > 0\},$$

with the partial ordering $\preceq_I$ such that

$$(V_1, N_1) \preceq_I (V_2, N_2) \Longleftrightarrow (V_2 \subseteq V_1 \text{ and } N_2 \geq N_1).$$

Let $m = 2$. We define a net $\{u^\alpha\}_{\alpha \in I}$ in $L^1([0, T], \mathbb{R}^2)$ as follows. For each index $\alpha = (V, N) \in I$, we let $u_1^\alpha(t) = N\cos(j_\alpha t)$, $u_2^\alpha(t) = N \sin(j_\alpha t)$, where j_α is the smallest integer $j > 0$ such that the functions $N\cos(jt), N\sin(jt)$ belong to V. (It follows from the Riemann-Lebesgue lemma that, for each fixed V and N, the functions $N\cos(jt)$ and $N\sin(jt)$ belong to V when j is large enough. Thus j_α exists.) Now it is easy to see that the net $\{u^\alpha\}_{\alpha \in I}$ thus defined satisfies (UWC) with $u^\infty = (0, 0)$. (To see this, let $\varepsilon > 0$, and let φ be a continuous function on $[0, T]$. Let $\alpha_0 = (V(\varphi, \varepsilon), 1)$. Then, if $\alpha = (V, N) \succeq_I \alpha_0$, it follows that $V \subset V_0(\varphi, \varepsilon)$. By definition, $(u_1^\alpha, u_2^\alpha) \in V \subset V_0(\varphi, \varepsilon)$, which implies $\sup_{0 \leq t \leq T} | \int_0^t \varphi(s)u_k^\alpha(s)ds| < \varepsilon, k = 1, 2.$)

We will now show that the above net does not $\mathcal{T}^0$-converge to the limit $u^\infty{=}(0, 0)$ in $L^1([0, T], \mathbb{R}^2)$. Let $\Sigma{=}(\mathbb{R}^2, f_0, f_1, f_2)$ with $f_0 = (0, 0)$, $f_1 = (1, 0)$, $f_3 = (0, x^{\frac{1}{5}})$. It is clear that this system is in $C^0(2)$. Consider, for each input $u \in L^1([0, T], \mathbb{R}^2)$, the system of differential equations $\dot{x}_1 = u_1$, $\dot{x}_2 = x_1^{1/5} u_2$. This system has the global existence and uniqueness property for every initial condition and every $u \in L^1([0, T], \mathbb{R}^2)$. Fix the initial condition $x(0) = (0, 0)$. We prove that $\{u^\alpha\}$ does not $\mathcal{T}^0$-converge to u^∞ by showing that the trajectories x^α of the above differential equations that correspond to the u^α do not converge to $x^\infty \equiv (0, 0)^T$. First, we prove a lemma.

Lemma A.1. *Fix a $V = (\varphi_1, \ldots, \varphi_n, \varepsilon) \in \mathcal{V}$. Then there exists an $N_V > 0$ such that if $N > N_V$ and $\alpha = (V, N)$, then $j_\alpha \leq N^5$.*

Proof. Let $A = \max_{k=1,\ldots,n}\{\|\varphi_k\|\}$, where $\|\cdot\|$ denotes the sup norm on $[0, T]$. Let $N_V = \frac{4A\pi}{\varepsilon}(\frac{nAT}{\varepsilon} + n)^{\frac{1}{2}}$. We show that $j_\alpha \leq N^5$ if $\alpha = (V, N)$, $N > N_V$. Suppose there is an $N > N_V$ such that $j_\alpha > N^5$. Then by the definition of the j_α, for every $1 \leq j \leq N^5$, we have

$$\text{(A.1)} \qquad \sup_{0 \leq t \leq T, k=1,\ldots,n} \max\left(\left| \int_0^t \cos(js)\varphi_k(s)ds \right|, \left| \int_0^t \sin(js)\varphi_k(s)ds \right| \right) \geq \frac{\varepsilon}{N}.$$

Let $K = [\frac{2ANT}{\varepsilon}] + 1$, where $[\cdot]$ denotes the integer part function (i.e., $[r]$ is the largest integer ν such that $\nu \leq r$). Divide the interval $[0, T]$ into K equal parts. Let $t_\ell = \frac{T\ell}{K}, \ell = 0, \ldots, K$, be the partition points, and define $\varphi_{k,\ell}(t) = \varphi_k(t)\chi_{[0,t_\ell]}$, where $\chi_{[0,t_\ell]}$ is the indicator function of $[0, t_\ell]$. For any two functions φ, ψ in $L^2[0, T]$, let us use $\langle \varphi, \psi \rangle = \int_0^T \varphi(s)\psi(s)ds$ to denote the inner product of φ, ψ. Then $\{\frac{\cos jt}{\sqrt{\pi}} :$

$j = 1, 2, \ldots\} \cup \{\frac{\sin jt}{\sqrt{\pi}} : j = 1, 2, \ldots\}$ is an orthonormal set in $L^2[0, T]$. From Bessel's inequality we know that

$$\begin{aligned}&\sum_{\ell=1}^{K}\sum_{k=1}^{n}\sum_{j=1}^{N^5}\left\{\left|\left\langle\frac{\cos(jt)}{\sqrt{\pi}}, \varphi_{k,\ell}\right\rangle\right|^2 + \left|\left\langle\frac{\sin(jt)}{\sqrt{\pi}}, \varphi_{k,\ell}\right\rangle\right|^2\right\} \\ \text{(A.2)}\qquad &\leq \sum_{\ell=1}^{K}\sum_{k=1}^{n} \|\varphi_{k,\ell}\|^2_{L^2[0,T]} \leq \sum_{\ell=1}^{K}\sum_{k=1}^{n} \|\varphi_k\|^2_{L^2[0,T]} \leq 2\pi nKA^2\,.\end{aligned}$$

On the other hand, it follows from (A.1) that for each $j \in \{1, \ldots, N^5\}$ there exist a number $k(j) \in \{1, \ldots, n\}$ and a $t(j) \in [0, T]$ such that at least one of the numbers

$$\Big|\int_0^{t(j)} \cos(js)\varphi_{k(j)}(s)ds\Big|\,, \qquad \Big|\int_0^{t(j)} \sin(js)\varphi_{k(j)}(s)ds\Big|$$

is $\geq \frac{\varepsilon}{N}$. Then $t(j) \in [t_{\ell-1}, t_\ell]$ for some $\ell = \ell(j)$. It is clear that the integrals of $\cos(js)\varphi_{k(j)}(s)ds$ and $\sin(js)\varphi_{k(j)}(s)ds$ from $t(j)$ to $t_{\ell(j)}$ are bounded by $\frac{AT}{K}$. Since $K > \frac{2ANT}{\varepsilon}$, we have $\frac{AT}{K} < \frac{\varepsilon}{2N}$. Thus at least one of the numbers

$$\Big|\int_0^{t_{\ell(j)}} \cos(js)\varphi_{k(j)}(s)ds\Big|\,, \qquad \Big|\int_0^{t_{\ell(j)}} \sin(js)\varphi_{k(j)}(s)ds\Big|$$

is $\geq \frac{\varepsilon}{2N}$. Therefore

$$\left|\left\langle\frac{\cos(jt)}{\sqrt{\pi}}, \varphi_{k(j),\ell(j)}\right\rangle\right|^2 + \left|\left\langle\frac{\sin(jt)}{\sqrt{\pi}}, \varphi_{k(j),\ell(j)}\right\rangle\right|^2 \geq \frac{\varepsilon^2}{4\pi N^2}\,.$$

Thus we have

$$\text{(A.3)}\quad \sum_{\ell=1}^{K}\sum_{k=1}^{n}\sum_{j=1}^{N^5}\left\{\left|\left\langle\frac{\cos(jt)}{\sqrt{\pi}}, \varphi_{k,\ell}\right\rangle\right|^2 + \left|\left\langle\frac{\sin(jt)}{\sqrt{\pi}}, \varphi_{k,\ell}\right\rangle\right|^2\right\} \geq N^5\frac{\varepsilon^2}{4\pi N^2} = \frac{\varepsilon^2 N^3}{4\pi}\,.$$

If we compare this with (A.2), we see that

$$N^3 \leq \frac{8\pi^2 nKA^2}{\varepsilon^2}\,.$$

Since $K < \frac{2ANT}{\varepsilon} + 2$, and $N \geq 1$, we easily get the inequality

$$N^2 \leq \frac{16\pi^2 A^2}{\varepsilon^2}\left(\frac{nAT}{\varepsilon} + n\right),$$

so that $N \leq N_V$. This contradiction completes our proof. □

Using the lemma, it is easy to see that the x^α do not converge to x^∞. Indeed, for any $\alpha = (V, N) \in I$, the solution x^α is given by

$$\text{(A.4)}\qquad x_1^\alpha(t) = \frac{N}{j_\alpha}\sin(j_\alpha t)\,, \qquad x_2^\alpha(t) = \frac{N^{\frac{6}{5}}}{j_\alpha^{\frac{1}{5}}}\int_0^t (\sin j_\alpha s)^{\frac{6}{5}}ds\,.$$

Take $\varepsilon_0 = 1$. If the net $\{x^\alpha\}$ converged to x^∞, then there would exist an $\alpha_0 = (V, N_0) \in I$ such that $\|x^\alpha(t)\| \leq 1$ for all $\alpha \succeq_I \alpha_0$ and all $t \in [0, T]$. Let $\alpha = (V, N)$ with $N > \max\{N_0, N_V\}$. Then $\alpha \succeq_I \alpha_0$ and

$$x_2^\alpha(T) = \frac{N^{\frac{6}{5}}}{j_\alpha^{\frac{1}{5}}}\int_0^T \sin^{\frac{6}{5}} j_\alpha s\, ds\,.$$

Clearly, $\int_0^T \sin^{\frac{6}{5}} j_\alpha s\, ds$ is bounded away from 0. Since $j_\alpha \leq N^5$, we conclude that $x_2^\alpha(T) \to \infty$ as $N \to \infty$. This contradiction proves that $\{x^\alpha\}$ does not converge. □

Our second example shows that a net $\{u^\alpha\}_{\alpha\in A}$ in $L^1([0,T],\mathbb{R}^m)$ may $\mathcal{T}^0$-converge to a $u \in L^1([0,T],\mathbb{R}^m)$ while failing to satisfy (UB). This will prove the assertion that in general (ii)$\not\Longrightarrow$(i) in Theorem 4.1.

Example A.2. Take $m = 1, T = 1$. Let $\mathcal{F}_1$ be the collection of all pairs (Σ, K), where Σ is an element of $C^0(1)$ and K is a compact subset of $\Omega^\Sigma \times [0,1]$. Let $\mathcal{F}_2$ be the set of all finite subsets of $\mathcal{F}_1$. Let $\mathcal{F} = \mathcal{F}_2 \times \mathbb{N}$, where $\mathbb{N}$ is the set of positive integers. Define a partial ordering $\preceq_{\mathcal{F}}$ on $\mathcal{F}$ by letting $(F_1, n_1) \preceq_{\mathcal{F}} (F_2, n_2)$ if $F_1 \subseteq F_2$ and $n_1 \leq n_2$. Then $(\mathcal{F}, \preceq_{\mathcal{F}})$ is a directed set. We define a net $\{u^{(F,n)}\}_{(F,n)\in\mathcal{F}}$ in $L^1([0,1],\mathbb{R})$ as follows. For each $n, k \in \mathbb{N}$, let $u_n^k(t) = n \sin kt$. Then from Theorem 4.1 we know that, for each fixed n, the sequence $\{u_n^k\}_{k=1}^\infty$ $\mathcal{T}^0$-converges to $u = 0$ in $L^1([0,1],\mathbb{R})$. Therefore, for each $(\Sigma, K) \in \mathcal{F}_1$, the set $\mathrm{Traj}_c(f^\Sigma(u_n^k), K)$ converges, as $k \to \infty$, to $\mathrm{Traj}_c(f^\Sigma(u), K)$ in $2^{ARC(\Omega^\Sigma)}$ under the topology $\mathcal{T}_{USC}(ARC(\Omega^\Sigma))$. Let $k_{\Sigma,K,n}$ be an integer such that $\mathrm{Traj}_c(f^\Sigma(u_n^k), K) \subseteq U_{\frac{1}{n}}(f^\Sigma(u), K)$ if $k \geq k_{\Sigma,K,n}$, where $U_\varepsilon(f, K)$ is the ε-neighborhood of $\mathrm{Traj}_c(f, K)$, i.e.,

$$U_\varepsilon(f,K) = \{\xi \in ARC(\Omega) : d_{ARC}(\xi,\zeta) < \varepsilon \text{ for some } \zeta \in \mathrm{Traj}_c(f,K)\}\,.$$

For any $F \in \mathcal{F}$, we let $k_{F,n} = \max\{k_{\Sigma,K,n} : (\Sigma, K) \in F\}$, and then define $u^{(F,n)}(t) = u_n^{k_{F,n}}(t)$. Clearly $\{u^{(F,n)}\}_{(F,n)\in\mathcal{F}}$ does not satisfy (UB). However, it is easy to show that $\{u^{(F,n)}\}_{(F,n)\in\mathcal{F}}$ $\mathcal{T}^0$-converges to u in $L^1([0,1],\mathbb{R})$. To see this, let $\Sigma \in C^0(1)$ be a system and K be a compact subset of $\Omega^\Sigma \times [0,1]$. We need to show that $\mathrm{Traj}_c(f^\Sigma(u^{(F,n)}), K)$ converges to $\mathrm{Traj}_c(f^\Sigma(u), K)$. For any given $\varepsilon > 0$, let $n(\varepsilon)$ be an integer such that $\frac{1}{n(\varepsilon)} < \varepsilon$. Let $F_0 = \{(\Sigma, K)\}$. Then if $(F_0, n(\varepsilon)) \preceq_{\mathcal{F}} (F, n)$, by definition, $n \geq n(\varepsilon)$ and $(\Sigma, K) \in F$. Therefore,

$$\mathrm{Traj}_c(f^\Sigma(u^{(F,n)}), K) \subseteq U_{\frac{1}{n}}(f^\Sigma(u), K) \subseteq U_\varepsilon(f^\Sigma(u), K)\,. \qquad \square$$

Our third example shows that, for general vector fields, the fact that for every fixed initial condition $x(\bar{t}) = \bar{x}$ the trajectory of f^j converges to the trajectory of f does not imply that f^j T-converges to f.

Example A.3. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a smooth function such that $\varphi(s) = 0$ if $s \leq 0$ or $s \geq 2$, and $\varphi(s) > 0$ if $0 < s < 2$. Define $f^j : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^2$ by letting $f^j(x,y,t) = (0, j\varphi(jx))$, and let $f(x,y,t) \equiv 0$. Then f and the f^j are time-varying vector fields on $\mathbb{R}^2$. Given any initial condition $(\bar{x},\bar{y},\bar{t})$, the solution $\xi_{j,\bar{x},\bar{y},\bar{t}} : \mathbb{R} \to \mathbb{R}^2$ of the initial value problem $\dot{\xi}(t) = f^j(\xi(t), t)$, $\xi(\bar{t}) = (\bar{x},\bar{y})$, is given by $\xi_{j,\bar{x},\bar{y},\bar{t}}(t) = (\bar{x}, \bar{y} + j(t-\bar{t})\varphi(j\bar{x}))$. For each $\bar{x}$, the number $\varphi(j\bar{x})$ is equal to zero for sufficiently large j. So for each initial condition $(\bar{x},\bar{y},\bar{t})$ the curve $t \to \xi_{j,\bar{x},\bar{y},\bar{t}}(t)$ converges uniformly, as $j \to \infty$, to the constant curve $\xi_{\infty,\bar{x},\bar{y},\bar{t}} \equiv (\bar{x},\bar{y})$, and $\xi_{\infty,\bar{x},\bar{y},\bar{t}}$ is precisely the solution of the limiting initial value problem $\dot{\xi} = f(\xi, t)$, $\xi(\bar{t}) = (\bar{x},\bar{y})$. On the other hand, if we consider a j-dependent initial condition $(\bar{x}^j, \bar{y}^j, \bar{t}^j)$ given by $\bar{x}^j = \frac{1}{j}$, $\bar{y}^j = 0$, $\bar{t}^j = 0$, and let $t^j = \frac{1}{j}$, then $\xi_{j,\bar{x}^j,\bar{y}^j,\bar{t}^j}(t^j) = (\frac{1}{j}, \varphi(1))$, which does not converge to $\xi_{\infty,0,0,0}(0)$, even though $(\bar{x}^j,\bar{y}^j,\bar{t}^j) \to (0,0,0)$ and $t^j \to 0$. Thus the curves $\xi_{j,\bar{x}^j,\bar{y}^j,\bar{t}^j}$ do not converge uniformly to $\xi_{\infty,0,0,0}(0)$ on the compact interval $[0,1]$. (Alternatively, we could have taken $\bar{t}^j = -\frac{1}{j}$, and then $\xi_{j,\bar{x}^j,\bar{y}^j,\bar{t}^j}(0) = (\frac{1}{j}, \varphi(1))$, so $\xi_{j,\bar{x}^j,\bar{y}^j,\bar{t}^j}$ does not even converge pointwise to $\xi_{\infty,0,0,0}$.) □

REFERENCES

[1] H. Antosiewicz, *Continuous parameter dependence and the method of averaging*, in Analytic Methods in the Theory of Non-Linear Vibrations (Proc. Internat. Sympos. Non-Linear Vibrations, Vol. I, 1961), Izdat. Akad. Nauk Ukrain. SSR, Kiev, 1963, pp. 51–58.

[2] Z. Artstein, *The limiting equations of nonautonomous ordinary differential equations*, J. Differential Equations, 25 (1977), pp. 184–202.

[3] Z. Artstein, *Continuous dependence of solutions of Volterra integral equations*, SIAM J. Math. Anal., 6 (1975), pp. 446–456.

[4] G. Buttazzo and R. Conti, *Γ-convergence and optimal control problems*, J. Optim. Theory Appl., 38 (1982), pp. 385–407.

[5] L. D. Berkovitz, *Optimal Control Theory*, Springer-Verlag, New York, 1974.

[6] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[7] R. V. Gamkrelidze, *Principles of Optimal Control Theory*, Plenum Press, New York, 1978.

[8] P. Hartman, *Differential Equations*, John Wiley, New York, 1964.

[9] J. Kurzweil and Z. Vorel, *Continuous dependence of solutions of differential equations on a parameter*, Czechoslovak Math. J., 7 (1957), pp. 568–583.

[10] J. Kurzweil and J. Jarnik, *Limit process in ordinary differential equations*, J. Appl. Math. Phys., 38 (1987), pp. 241–256.

[11] J. Kurzweil and J. Jarnik, *Iterated Lie brackets in limit processes in ordinary differential equations*, Results Math., 14 (1988), pp. 125–137.

[12] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1968.

[13] W. Liu, *Averaging Theorems for Highly Oscillatory Ordinary Differential Equations and the Approximation of General Paths by Admissible Trajectories for Nonholonomic Systems*, Ph.D. thesis, Rutgers University, Piscataway, NJ, 1992.

[14] W. Liu, *Averaging theorems for highly oscillatory differential equations and iterated Lie brackets*, SIAM J. Control Optim., 35 (1997), pp. 1989–2020.

[15] W. Liu, *Averaging Theorems for Highly Oscillatory Differential Equations and Iterated Lie Brackets, the Hölder Continuous Input Case*, preprint.

[16] W. Liu, *An approximation algorithm for nonholonomic systems*, SIAM J. Control Optim., 35 (1997), pp. 1328–1365.

[17] E. Michael, *Topologies in spaces of subsets*, Trans. Amer. Math. Soc., 71 (1971), pp. 152–182.

[18] R. K. Miller, *Nonlinear Volterra Integral Equations*, W. A. Benjamin, Menlo Park, CA, 1971.

[19] L. W. Neustadt, *On the solutions of certain integral-like operator equations: Existence, uniqueness and dependence theorems*, Arch. Rational Mech. Anal., 38 (1970), pp. 132–160.

[20] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.

[21] G. Sansone and R. Conti, *Nonlinear Differential Equations*, Macmillan, New York, 1964.

[22] H. J. Sussmann and W. Liu, *Limits of highly oscillatory controls and approximation of general paths by admissible trajectories*, in Proc. 30th IEEE Conf. Decision and Control, Brighton, UK, 1991, IEEE Publications, Piscataway, NJ, 1991, pp. 437–442.

[23] H. J. Sussmann and W. Liu, *Lie bracket extensions and averaging: The single-bracket case*, in Nonholonomic Motion Planning, Z. X. Li, J. F. Canny, eds., Kluwer Academic Publishers, Norwell, MA, 1993, pp. 109–148.

[24] H. J. Sussmann and W. Liu, *Motion planning and approximate tracking for controllable systems without drift*, in Proc. 25th Annual Conf. on Inform. Sciences and Systems, Johns Hopkins University, Baltimore, MD, 1991, pp. 547–551.

SIAM J. Control Optim.
Vol. 37, No. 3, pp. 804–824

STOCHASTIC SHORTEST PATH GAMES*

STEPHEN D. PATEK[†] AND DIMITRI P. BERTSEKAS[‡]

Abstract. We consider dynamic, two-player, zero-sum games where the "minimizing" player seeks to drive an underlying finite-state dynamic system to a special terminal state along a least expected cost path. The "maximizer" seeks to interfere with the minimizer's progress so as to maximize the expected total cost. We consider, for the first time, undiscounted finite-state problems, with compact action spaces, and transition costs that are not strictly positive. We admit that there are policies for the minimizer which permit the maximizer to prolong the game indefinitely. Under assumptions which generalize deterministic shortest path problems, we establish (i) the existence of a real-valued equilibrium cost vector achievable with stationary policies for the opposing players and (ii) the convergence of value iteration and policy iteration to the unique solution of Bellman's equation.

Key words. game theory, stochastic games, optimization, dynamic programming, stochastic shortest paths

AMS subject classifications. 90D15, 93E05, 49L20

PII. S0363012996299557

1. Introduction.

This paper develops basic theory relating to stochastic shortest path games. These are two-player, zero-sum games where the minimizing player seeks to drive an underlying finite-state dynamic system to a special terminal state along a least expected cost path. The maximizer seeks to interfere with the minimizer's progress so as to maximize the expected total cost. In actual play, the players implement actions simultaneously at each stage with full knowledge of the state of the system but *without* knowledge of each other's current decision.

Games of this type have been studied for some time. The field was initiated by Shapley in his classical paper [7]. In Shapley's games, two players are successively faced with matrix games (in mixed strategies) where both the immediate cost and the transition probabilities to new matrix games are influenced by the stagewise decisions of the players. In this formulation, the state of the system is the matrix game currently being played. It is assumed that this set of states is finite and that there is a nonzero minimal probability that, at any stage, the game will transition to a terminal state, ending the sequence of rewards and payoffs. This formulation is closely related to infinite-horizon games with *discounted* additive cost. The analysis of such games is straightforward, the main results being (i) the existence and characterization of a unique real-valued equilibrium cost vector achievable in stationary randomized policies and (ii) the convergence of value iteration and policy iteration to the equilibrium cost.

Since Shapley's work, game theorists have actively studied extensions to the discounted-cost model. In [4], Kushner and Chamberlain consider undiscounted, pursuit/evasion, stochastic games where there is a terminal state corresponding to the evader being "caught." The state space is assumed to be finite (with $n + 1$ elements,

*Received by the editors February 21, 1996; accepted for publication (in revised form) November 26, 1997; published electronically March 2, 1999. This research was supported by National Science Foundation grant 9300494-DMI and an Office of Naval Research fellowship.

http://www.siam.org/journals/sicon/37-3/29955.html

[†]Department of Systems Engineering, School of Engineering and Applied Science, University of Virginia, Olsson Hall, Charlottesville, VA 22903-2442 (sdp5f@virginia.edu).

[‡]Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, Room 35-210, Cambridge, MA 02139 (dimitrib@mit.edu).

one of which is the terminal state). Making some regularity assumptions on the transition probabilities and cost functions, they consider pure strategies over compact action spaces. In addition, they assume either of the following.

1. The n-stage probability transition matrix $[P(\mu,\nu)]^n$ (from nonterminal states to nonterminal states) is a "uniform contraction" in the stationary policy pairs (μ,ν) of the two players. (That is, for some $\epsilon > 0$, $[P(\mu,\nu)]^n$ has row-sums less than $1-\epsilon$ for all stationary policy pairs (μ,ν).)
2. The transition costs (to the pursuer) are uniformly bounded below by $\delta > 0$, and there exists a stationary policy $\tilde{\mu}$ for the pursuer that makes $[P(\tilde{\mu},\nu)]^n$ a uniform contraction under all stationary policies for the evader.

They show that there exists an equilibrium cost vector for the game which can be found through value iteration. In [10], van der Wal considers a special case of Kushner and Chamberlain's games. Under more restrictive assumptions about the pursuer's ability to catch the evader, he gives error bounds for the updates in value iteration.

In [3], Kumar and Shiau give a detailed analysis of stochastic games with very mild assumptions about the state space and control constraint sets. For the case of nonnegative additive cost (with no discounting), they establish the existence of an extended real equilibrium cost vector in non-Markov randomized policies (where for both players the best mixed action can depend on all of the past states and controls, as well as the current state). They show that the minimizing player can achieve the equilibrium using a stationary Markov randomized policy and that, in case the state space is finite, the maximizing player can play ϵ-optimally using stationary randomized policies.

Other researchers have studied so-called nonterminating stochastic games (also sometimes called "undiscounted" stochastic games), where the costs are not discounted but are averaged instead. Such *average-cost* games have a rich mathematical structure which has been extensively covered in the literature [13, 5].

In this paper, we consider undiscounted additive cost games without averaging. We admit that there are policies for the minimizer which allow the maximizer to prolong the game indefinitely at infinite cost to the minimizer. We do not assume nonnegativity of cost, as in [4] and [3]. We make alternative assumptions which guarantee that, at least under optimal policies, the terminal state is reached with probability one. Our results imply the results of Shapley [7], as well as those of Kushner and Chamberlain [4]. Because of our assumptions relating to termination, we are able to derive stronger conclusions than those made by Kumar and Shiau [3] for the case of a finite state space. Note that because we do not assume nonnegativity of the costs, the analysis is much more complicated than the corresponding analysis of Kushner and Chamberlain [4]. Our formal assumptions generalize (to the case of two players) those for stochastic shortest path problems [2]. Because of this, we refer to our class of games as "stochastic shortest path games." Our games are characterized by either (i) inevitable termination (under all policies) or (ii) an incentive for the minimizer to drive the system to termination in a finite expected number of stages. We shall see that the results of [2] are essential in developing our present theory.

In section 2 we give a precise mathematical formulation for stochastic shortest path games. In section 3, we relate our general formulation to Shapley's original games [7]. We develop our main results in section 4. This is where we show that stochastic shortest path games have an equilibrium solution which can be characterized by the unique solution to Bellman's equation. We also prove the convergence of value iteration and policy iteration to the equilibrium cost. In section 5, we give an example of pursuit and evasion, illustrating our main results. Finally, in the appendix we collect some well-known results about dynamic games which are crucial to our development.

2. Mathematical formulation. Let S denote a finite state space with elements labeled $i = 1, \dots, n$. For each $i \in S$, define $U(i)$ and $V(i)$ to be the sets of actions available to the minimizer and maximizer at state i, respectively. These are collectively referred to as *control constraint sets.* The probability of transitioning from $i \in S$ to $j \in S$ under $u \in U(i)$ and $v \in V(i)$ is denoted $p_{ij}(u, v)$. The expected cost (to the minimizer) of transitioning from $i \in S$ under $u \in U(i)$ and $v \in V(i)$ is denoted $c_i(u, v)$.

We denote the sets of one-stage policies for the minimizing and maximizing players as M and N, respectively, where

$$M = \left\{ \mu : S \mapsto \bigcup_{i \in S} U(i) \quad | \quad \mu(i) \in U(i) \quad \forall\, i \in S \right\},$$

$$N = \left\{ \nu : S \mapsto \bigcup_{i \in S} V(i) \quad | \quad \nu(i) \in V(i) \quad \forall\, i \in S \right\}.$$

The sets of policies for the minimizing and maximizing players are denoted by $\bar{M}$ and $\bar{N}$, where

$$\begin{aligned} \pi_M = \{\mu^0, \mu^1, \dots\} \in \bar{M} \quad &\Longleftrightarrow \quad \mu^k \in M \quad \forall\, k, \\ \pi_N = \{\nu^0, \nu^1, \dots\} \in \bar{N} \quad &\Longleftrightarrow \quad \nu^k \in N \quad \forall\, k. \end{aligned}$$

Given $\mu \in M$ and $\nu \in N$, let $P(\mu, \nu)$ denote the transition probability matrix that results when μ and ν are in effect. That is,

$$[P(\mu, \nu)]_{ij} = p_{ij}(\mu(i), \nu(i)).$$

Let $c(\mu, \nu)$ denote the vector whose components are $c_i(\mu(i), \nu(i))$. That is,

$$[c(\mu, \nu)]_i = c_i(\mu(i), \nu(i)).$$

Given two allowable opposing policies, $\pi_M = \{\mu^0, \mu^1, \dots\} \in \bar{M}$ and $\pi_N = \{\nu^0, \nu^1, \dots\} \in \bar{N}$, we formally define the resulting cost (to the minimizer) to be

$$x(\pi_M, \pi_N) = \liminf_{t \to \infty} h^t_{\pi_M, \pi_N}, \tag{2.1}$$

where

$$h^t_{\pi_M, \pi_N} \triangleq \left\{ c(\mu^0, \nu^0) + \sum_{k=1}^{t} [P(\mu^0, \nu^0) \cdots P(\mu^{k-1}, \nu^{k-1})] c(\mu^k, \nu^k) \right\}. \tag{2.2}$$

Note that $h^t_{\pi_M, \pi_N}$ can be interpreted loosely as the expected t-stage cost vector under the policies π_M and π_N.

In establishing our main results, the definitions and assumptions in the following paragraphs will be helpful. We say that a policy $\pi_M = \{\mu^0, \mu^1, \dots\}$ for the minimizer is *stationary* if $\mu^k = \mu$ for all k. When this is the case and no confusion can arise, we use μ to denote the corresponding policy π_M, and we refer to π_M as *the stationary policy* μ. Similar definitions hold for stationary policies of the maximizer.

The state $1 \in S$ has special importance. We shall refer to it as the *terminal state.* This state is assumed to be absorbing and cost free, that is, $p_{11}(u,v) = 1$ and $c_1(u,v) = 0$ for all $u \in U(1)$ and $v \in V(1)$. Let $\pi_M = \{\mu^0, \mu^1, \dots\} \in \bar{M}$ and $\pi_N = \{\nu^0, \nu^1, \dots\} \in \bar{N}$ be an arbitrary pair of policies. We say that the corresponding Markov chain *terminates with probability one* if the following limit satisfies:

$$\lim_{t\to\infty} [P(\mu^0,\nu^0)P(\mu^1,\nu^1)\cdots P(\mu^t,\nu^t)]_{i1} = 1 \quad \forall\, i \in S. \tag{2.3}$$

(The limit above exists because the sequence (for each $i \in S$) is monotonically nondecreasing and bounded above.) We shall refer to a pair of policies (π_M, π_N) as *terminating with probability one* if the corresponding Markov chain terminates with probability one; otherwise, we refer to the pair as *prolonging.*

A stationary policy $\mu \in M$ for the minimizer is said to be *proper* if the pair (μ, π_N) is terminating with probability one for all $\pi_N \in \bar{N}$. A stationary policy μ is *improper* if it is not proper. If μ is improper, then there is a policy for the maximizer $\pi_\mu \in \bar{N}$ under which there is a positive probability that the game will never end from some initial state. The designation of proper (or improper) applies only to stationary policies for the minimizer.

It is convenient to define the set $X = \{x \in R^n \mid x_1 = 0\}$. This is the space (of cost vectors) over which our main results hold. We denote by $\mathbf{0}$ the zero vector in X. Let $\mathbf{1}_X$ denote the vector $(0,1,1,\dots 1)' \in X$. It is useful to define the following operators on X.

$$T_{\mu\nu}(x) = c(\mu,\nu) + P(\mu,\nu)x; \qquad \mu \in M, \nu \in N. \tag{2.4}$$

$$T_\mu(x) = \sup_{\nu \in N} \left[c(\mu,\nu) + P(\mu,\nu)x\right]; \qquad \mu \in M, \tag{2.5}$$

$$T(x) = \inf_{\mu \in M} \sup_{\nu \in N} \left[c(\mu,\nu) + P(\mu,\nu)x\right]. \tag{2.6}$$

$$\tilde{T}_\nu(x) = \inf_{\mu \in M} \left[c(\mu,\nu) + P(\mu,\nu)x\right]; \qquad \nu \in N, \tag{2.7}$$

$$\tilde{T}(x) = \sup_{\nu \in N} \inf_{\mu \in M} \left[c(\mu,\nu) + P(\mu,\nu)x\right]. \tag{2.8}$$

The suprema and infima in the above are taken componentwise. We use the notation $T^t_{\mu\nu}(x)$ to denote the t-fold composition of $T_{\mu\nu}$ applied to $x \in X$. Similar definitions hold for $T^t_\mu(x)$, $T^t(x)$, $\tilde{T}^t_\nu(x)$, and $\tilde{T}^t(x)$. In the appendix, we collect (and prove for completeness) some well-known results about these "T"-operators: monotonicity (Lemma A.1) and continuity (Lemma A.3).

The following are our standing assumptions.

Assumption SSP. The following are true:

1. There exists at least one proper policy for the minimizer.
2. If a pair of policies (π_M, π_N) is prolonging, then the expected cost to the minimizer is infinite for at least one initial state. That is, there is a state i for which $\lim_{t\to\infty}[h^t_{\pi_M,\pi_N}]_i = \infty$.

Assumption R (regularity). The following are true:

1. The control constraint sets are compact. That is, for each $i \in S$, $U(i)$ and $V(i)$ are compact subsets of metric spaces. (This implies that M and N are compact.)
2. The functions $p_{ij}(u,v)$ are continuous with respect to $(u,v) \in U(i) \times V(i)$, and the functions $c_i(u,v)$ are

   (a) lower semicontinuous with respect to $u \in U(i)$ (with $v \in V(i)$ fixed) and
   (b) upper semicontinuous with respect to $v \in V(i)$ (with $u \in U(i)$ fixed).

   (The Weierstrass theorem implies that the supremum and infimum in the definitions of the operators T_μ and $\tilde{T}_\nu$ are always achieved by elements of N and M, respectively. That is, for every $x \in X$, there exists $\nu \in N$ such that $T_\mu(x) = T_{\mu\nu}(x) \in X$. Similarly, for every $x \in X$, there exists $\mu \in M$ such that $\tilde{T}_\nu(x) = T_{\mu\nu}(x) \in X$.)

3. For all $x \in X$, the infimum and supremum in the definitions of the operators T and $\tilde{T}$ are achieved by elements of M and N. That is, for every $x \in X$, there exists $\mu \in M$ and $\nu \in N$ such that $T(x) = T_\mu(x) \in X$ and $\tilde{T}(x) = \tilde{T}_\nu(x) \in X$.
4. For each $x \in X$, we have $T(x) = \tilde{T}(x)$.

Note that part 4 of Assumption R is satisfied under conditions for which a minimax theorem can be used to interchange "inf" and "sup." In particular, this part, as well as the entire Assumption R, is satisfied if

1. the sets $U(i)$ and $V(i)$ are nonempty, convex, and compact subsets of Euclidean spaces,
2. the functions $p_{ij}(u,v)$ are bilinear of the form $u'Q_{ij}v$, where Q_{ij} is a real matrix of dimension commensurate with $U(i)$ and $V(i)$,
3. the functions $c_i(u,v)$ are
   (a) convex and lower semicontinuous as functions of $u \in U(i)$ with v fixed in $V(i)$, and
   (b) concave and upper semicontinuous as functions of $v \in V(i)$ with u fixed in $U(i)$.

This follows from the Sion–Kakutani theorem (see [8, p. 232] or [6, p. 397]). We will show in section 3 that dynamic games with "mixed" strategies over finite underlying action spaces satisfy this assumption.

To verify that a stationary policy $\mu \in M$ is proper, we need only check whether (μ, ν) is terminating with probability one for all *stationary* policies $\nu \in N$ for the maximizer. Furthermore, if $\mu \in M$ is improper, then we can always find a *stationary* policy $\nu \in N$ for the maximizer which is prolonging when paired with μ. This is shown in the following lemma.

LEMMA 2.1. *If $\mu \in M$ is such that (μ, ν) terminates with probability one for all $\nu \in N$, then μ is proper.*

Proof. The proof uses the analysis of [2]. Let $\mu \in M$ be a fixed policy for the minimizer, and suppose that the pair (μ, ν) is terminating with probability one for all stationary policies of the maximizer $\nu \in N$. With μ fixed, the maximizer is faced with a stochastic shortest path problem of the type considered in [2]. (The maximizer has no improper policies (against μ).) Now modify the problem such that the costs of transitioning from nonterminal states are all set to one but all of the transition probabilities are left unchanged. The assumptions of [2] remain satisfied, so the optimal expected cost for the maximizer in the new problem is bounded, even over nonstationary policies. Thus, the maximum expected number of stages to termination under μ is finite. This is true for both the modified problem and the original version of the game. This implies that μ is proper. □

One of the objectives of this paper is to show that under Assumptions SSP and R there exist policies $\pi_M^* \in \bar{M}$ and $\pi_N^* \in \bar{N}$ such that

$$x(\pi_M, \pi_N^*) \geq x(\pi_M^*, \pi_N^*) \qquad \forall\ \pi_M \in \bar{M},$$
$$x(\pi_M^*, \pi_N) \leq x(\pi_M^*, \pi_N^*) \qquad \forall\ \pi_N \in \bar{N}.$$

Such a cost vector $x^* \overset{\triangle}{=} x(\pi_M^*, \pi_N^*)$ is called the *equilibrium cost vector* (or *value*) of the stochastic shortest path game. The policies π_M^* and π_N^* form an *equilibrium solution.* Since this is a zero-sum game, we know that the equilibrium cost (if it exists) is unique. Another objective of this paper is to show that the equilibrium cost vector is characterized as the unique solution to Bellman's equation, with *stationary* equilibrium policies for the opposing players. After these results are established, we proceed to show that value iteration and policy iteration converge to the unique solution of Bellman's equation.

3. Connection to Shapley's stochastic games.

The mathematical formulation of the preceding section includes as a special case the stochastic games of Shapley. To see this, assume that the number of actions available to either player at any state is finite. As before, the players implement underlying actions simultaneously at each stage with full knowledge of the state of the system but *without* knowledge of each other's current decision. However, the players are now allowed to randomize their decisions in formulating a policy so as to keep their opponents from adapting to a deterministic policy. That is, in considering what to do at each state, the players choose *probability distributions* over underlying control sets rather than specific underlying control actions. In other words, the players use randomized or "mixed" policies.

For each $i \in S$, define $A(i)$ and $B(i)$ to be the *finite* sets of underlying actions to the minimizer and maximizer, respectively. These are the physical controls the players may ultimately implement at state i. Let $|A(i)|$ and $|B(i)|$ denote the numbers of elements in each set of actions. We define the players' "control constraint sets" for the game as

$$U(i) = \left\{ u \in R^{|A(i)|} \;\middle|\; \sum_{j \in A(i)} u_j = 1; \quad u_j \; \geq \; 0 \right\}, \tag{3.1}$$

$$V(i) = \left\{ v \in R^{|B(i)|} \;\middle|\; \sum_{j \in B(i)} v_j = 1; \quad v_j \; \geq \; 0 \right\}. \tag{3.2}$$

Thus, $U(i)$ is the set of probability distributions over control actions $A(i)$ available to the minimizer from state $i \in S$. Similarly, $V(i)$ is the set of probability distributions over underlying control actions $B(i)$ available to the maximizer from state $i \in S$. Here the functions $p_{ij}(u,v)$ and $c_i(u,v)$ are, respectively, of the form

$$p_{ij}(u,v) = \sum_{k \in A(i)} \sum_{l \in B(i)} \underline{p}_{ij}(k,l) u_k v_l, \tag{3.3}$$

$$c_i(u,v) = \sum_{j \in S} \sum_{k \in A(i)} \sum_{l \in B(i)} \underline{g}_{ij}(k,l) \underline{p}_{ij}(k,l) u_k v_l, \tag{3.4}$$

where the functions $\underline{p}_{ij}$ and $\underline{g}_{ij}$ denote the transition probabilities and costs of the underlying two-player Markov decision process. Since the sets $U(i)$ and $V(i)$ are polyhedral and the functions $c_i(u,v)$ and $p_{ij}(u,v)$ are bilinear for all i and j (and continuous) as functions of $(u,v) \in U(i) \times V(i)$, it is clear that Assumption R is satisfied. (Parts 3 and 4 are satisfied thanks to the minimax theorem of von Neumann [11].)

4. Main results.

We now develop our main results; namely, the existence and characterization of a unique equilibrium cost vector, the convergence of value iteration, and the convergence of policy iteration. In sections 4.1 and 4.2, we characterize optimal solutions for the maximizer and minimizer, respectively, for the case where the opposing player fixes a policy. After we lay this groundwork, we consider the game proper in section 4.3.

4.1. The case where the minimizer's policy is fixed. Consider the policy $\pi_M = \{\mu^0, \mu^1, \dots\} \in \bar{M}$. The *cost of* π_M is defined by

$$x(\pi_M) = \liminf_{t\to\infty} \max_{\pi_N \in \bar{N}} h^t_{\pi_M,\pi_N}. \tag{4.1}$$

The appendix shows that, with our assumptions on c and P, the maximum in (4.1) is attained for every t (see Lemma A.5). The cost of a stationary policy μ for the minimizer is denoted $x(\mu)$ and is computed according to equation (4.1) where $\pi_M = \{\mu, \mu, \dots\}$.

Given a vector $w \in R^n$ whose elements are positive, the corresponding weighted sup-norm, denoted $\|\cdot\|^w_\infty$, is defined by

$$\|x\|^w_\infty = \max_{i=1,\dots,n} x_i/w_i \qquad \forall\, x \in R^n.$$

The next lemma, similar to results derived in [1, 9], follows from the theory of one-player stochastic shortest paths.

LEMMA 4.1. *Assume that all stationary policies for the minimizer are proper. The operator T is a contraction mapping on the set $X = \{x \in R^n \mid x_1 = 0\}$ with respect to a weighted sup-norm. Moreover, if $\mu \in M$ is proper, then T_μ is a contraction mapping with respect to a weighted sup-norm.*

Proof. We will first show the result about T for the case that all stationary policies are proper. Our strategy is to identify a vector of weights w and to show that this set of weights is one for which T is a contraction with respect to $\|\cdot\|^w_\infty$.

Let us define a new one-player stochastic shortest path problem of the type considered in [2]. This problem is defined such that the transition probabilities remain unchanged and the transition costs are all set equal to -1 for all states other than the terminal state. The important difference is that the maximizer and minimizer "work together" in the sense that the decision space (for the single player of the new problem) is over $\bar{M} \times \bar{N}$. This is a stochastic shortest path problem where all stationary policies are proper. Using the results of [2], there is an optimal cost vector $\tilde{x} \in X$ which can be achieved using a stationary policy $(\tilde{\mu}, \tilde{\nu}) \in \bar{M} \times \bar{N}$. Note that, since the transition costs from all nonterminal states are set to -1 in the new stochastic shortest path problem, we have $\tilde{x}_i \le -1$ for all $i \ne 1$. Moreover, from Bellman's equation we have

$$\tilde{x} = -\mathbf{1}_X + P(\tilde{\mu}, \tilde{\nu})\tilde{x},$$

where $\mathbf{1}_X = (0, 1, 1, \dots, 1)' \in X$. Also, for all $\mu \in M$ and $\nu \in N$,

$$\tilde{x} \le -\mathbf{1}_X + P(\mu, \nu)\tilde{x}.$$

Thus, for all $\mu \in M$, $\nu \in N$, and $i \ne 1$,

$$\begin{aligned} \sum_{j=2}^{n} p_{ij}(\mu(i), \nu(i)) \cdot (-\tilde{x}_j) &\le -\tilde{x}_i - 1 \\ &\le -\tilde{x}_i \gamma, \end{aligned} \tag{4.2}$$

where $\gamma = \max_{i \neq 1}(\tilde{x}_i + 1)/\tilde{x}_i$. Since the $\tilde{x}_i \leq -1$ for all $i \neq 1$, we have that $\gamma \in [0, 1)$. Now define $w = -\tilde{x} + (1, 0, 0, \ldots, 0)'$. Note that w is a vector in R^n whose elements are all strictly positive.

Let us now resume consideration of the original stochastic shortest path game. Let x and $\bar{x}$ be any two elements of X such that $\|x - \bar{x}\|_\infty^w = c$. Let $\mu \in M$ be such that $T_\mu(x) = T(x)$, and let $\nu \in N$ be such that $T_\mu(\bar{x}) = T_{\mu\nu}(\bar{x})$. Then,

$$\begin{aligned} T(\bar{x}) - T(x) &= T(\bar{x}) - T_\mu(x) \\ &\leq T_\mu(\bar{x}) - T_\mu(x) \\ &= T_{\mu\nu}(\bar{x}) - T_\mu(x) \\ &\leq T_{\mu\nu}(\bar{x}) - T_{\mu\nu}(x). \end{aligned}$$

Thus,

$$[T(\bar{x})]_i - [T(x)]_i \leq \sum_{j=2}^{n} [P(\mu, \nu)]_{ij} (\bar{x}_j - x_j).$$

Using this, we see that for all $i \neq 1$

$$\begin{aligned} \frac{[T(\bar{x}) - T(x)]_i}{c w_i} &\leq \frac{1}{c w_i} \sum_{j=2}^{n} p_{ij}(\mu(i), \nu(i))(\bar{x}_j - x_j) \\ &\leq \frac{1}{w_i} \sum_{j=2}^{n} p_{ij}(\mu(i), \nu(i)) w_j \\ &= \frac{1}{-\tilde{x}_i} \sum_{j=2}^{n} p_{ij}(\mu(i), \nu(i))(-\tilde{x}_j) \\ &\leq \frac{1}{-\tilde{x}_i}(-\tilde{x}_i)\gamma = \gamma, \end{aligned}$$

where the last inequality follows from (4.2). Thus, we get

$$\frac{[T(\bar{x})]_i - [T(x)]_i}{w_i} \leq c\gamma \quad \forall\, i \neq 1.$$

Since $[T(\bar{x})]_1 - [T(x)]_1 = 0$,

$$\frac{[T(\bar{x})]_i - [T(x)]_i}{w_i} \leq c\gamma \quad \forall\, i.$$

Using similar arguments, we may show that

$$\frac{[T(x)]_i - [T(\bar{x})]_i}{w_i} \leq c\gamma \quad \forall\, i.$$

Combining the preceding inequalities, we see that $\|T(x) - T(\bar{x})\|_\infty^w \leq c\gamma$. Since $0 \leq \gamma < 1$, we have that T is a contraction over X with respect to $\|\cdot\|_\infty^w$.

Now suppose $\mu \in M$ is proper. By viewing T_μ as the "T"-operator in a new game where $U(i) \equiv \{\mu(i)\}$, we have the desired result. □

LEMMA 4.2. *Given a proper policy μ, the following are true.*

1. *The cost $x(\mu)$ of μ is the unique fixed point of T_μ in $X = \{x \in R^n \mid x_1 = 0\}$.*
2. $x(\mu) = \sup_{\pi_N \in \bar{N}} x(\mu, \pi_N)$.
3. *We have $T_\mu^t(x) \to x(\mu)$ for all $x \in X$ with linear convergence.*

Proof. An induction argument (cf. Lemma A.5) easily shows that

$$T_\mu^{t+1}(\mathbf{0}) = \max_{\{\nu^0,\dots,\nu^t\}} \left\{ c(\mu,\nu^0) + \sum_{k=1}^{t} [P(\mu,\nu^0)P(\mu,\nu^1)\cdots P(\mu,\nu^{k-1})]c(\mu,\nu^k) \right\},$$

where $\mathbf{0}$ is the zero vector in X. Thus, using the preceding lemma and the definition of $x(\mu)$, we have

$$x(\mu) = \lim_{t\to\infty} T_\mu^{t+1}(\mathbf{0}) = \tilde{x}_\mu,$$

where $\tilde{x}_\mu$ is the unique fixed point of the contraction mapping T_μ within X, proving statement 1.

Consider the following infinite-horizon stochastic shortest path problem for the maximizer:

$$\sup_{\pi_N \in \bar{N}} \liminf_{t\to\infty} \left\{ c(\mu,\nu^0) + \sum_{k=1}^{t} [P(\mu,\nu^0)\cdots P(\mu,\nu^{k-1})]c(\mu,\nu^k) \right\}.$$

This problem is covered by the theory developed in [2] since the fact that μ is proper implies that termination is inevitable under all policies in the maximizer's problem. The optimal cost of this problem is $\sup_{\pi_N \in \bar{N}} x(\mu, \pi_N)$, and according to the theory of [2], it is equal to the limit of the successive approximation method applied to this problem, which is $\lim_{t\to\infty} T_\mu^{t+1}(\mathbf{0})$ and is also the unique fixed point of T_μ within X. This proves statement 2.

Finally, the linear convergence of $T_\mu^{t+1}(\mathbf{0})$ follows from the contraction property of T_μ. □

LEMMA 4.3. *If $x \geq T_\mu(x)$ for some $x \in X$, then μ is proper.*

Proof. To reach a contradiction, suppose μ is improper. According to Assumption SSP and Lemma 2.1, there exists a stationary maximizer's policy $\bar{\nu} \in N$ such that $(\mu, \bar{\nu})$ is prolonging and results in unbounded expected cost from some initial state when played against μ.

Let x be an element in X such that $x \geq T_\mu(x)$. Then, applying T_μ to x, we have that

$$x \geq T_\mu(x) \geq c(\mu,\bar{\nu}) + P(\mu,\bar{\nu})x,$$

where the second inequality follows from the definition of T_μ. From the monotonicity of T_μ, we get

$$\begin{aligned} x \geq T_\mu(x) \geq T_\mu^2(x) &\geq T_\mu(c(\mu,\bar{\nu}) + P(\mu,\bar{\nu})x) \\ &\geq P(\mu,\bar{\nu})P(\mu,\bar{\nu})x + [c(\mu,\bar{\nu}) + P(\mu,\bar{\nu})c(\mu,\bar{\nu})], \end{aligned}$$

where the last inequality follows again from the definition of T_μ. Proceeding inductively and using the same steps, we have that for all t

$$x \geq T_\mu^t(x) \geq P(\mu,\bar{\nu})^{t+1}x + \sum_{k=0}^{t} P(\mu,\bar{\nu})^k c(\mu,\bar{\nu}).$$

On the other hand, because the policy $\bar{\nu}$ results in infinite expected cost (from some initial state) when played against μ, some subsequence of $\sum_{k=0}^{t} [P(\mu, \bar{\nu})]^k c(\mu, \bar{\nu})$ must have a coordinate that tends to infinity. (The term involving x remains bounded because it is just x multiplied by the product of stochastic matrices.) This contradicts the above inequality. Thus, μ must be proper. □

4.2. The case where the maximizer's policy is fixed. By Assumption SSP there exists a proper policy for the minimizer. Thus, it is impossible that a single policy for the maximizer prolongs the game for *all* policies of the minimizer. Let us define $\tilde{x}(\pi_N)$ to be *the cost of the policy* $\pi_N \in \bar{N}$, as follows:

$$\tilde{x}(\pi_N) = \liminf_{t\to\infty} \min_{\pi_M \in \bar{M}} h^t_{\pi_M, \pi_N}, \tag{4.3}$$

where $\pi_N = \{\nu^0, \nu^1, \dots\}$. The cost of a stationary policy ν for the minimizer is denoted $\tilde{x}(\nu)$ and is computed according to equation (4.3), where $\pi_N = \{\nu, \nu, \dots\}$.

LEMMA 4.4. *For any $\nu \in N$, the following are true.*

1. *The cost $\tilde{x}(\nu)$ of ν is the unique fixed point of $\tilde{T}_\nu$ in $X = \{x \in R^n \mid x_1 = 0\}$.*
2. *$\tilde{x}(\nu) = \inf_{\pi_M \in \bar{M}} x(\pi_M, \nu)$.*
3. *We have $\tilde{T}^t_\nu(x) \to \tilde{x}(\nu)$ for all $x \in X$. If for all $\mu \in M$ the pair (μ, ν) terminates with probability one, then the convergence is linear.*

Proof. This follows directly from the theory of (one-player) stochastic shortest path problems. □

4.3. Main results for the game. We now establish the main results of the paper: the existence and characterization of a unique equilibrium solution, the convergence of value iteration, and the convergence of policy iteration.

PROPOSITION 4.5. *The operator T has a unique fixed point x^* on X.*

Proof. We begin by showing that T has at most one fixed point in X. Suppose x and x' are both fixed points of T in X. We can select $\mu \in M$ and $\mu' \in M$ such that $x = T(x) = T_\mu(x)$ and $x' = T(x') = T_{\mu'}(x')$. By Lemma 4.3, we have that μ and μ' are proper. Lemma 4.2 implies that $x = x(\mu)$ and $x' = x(\mu')$. Since μ' isn't necessarily optimal with respect to x in applying the T operator, we have from the monotonicity of T that $x = T^t(x) \le T^t_{\mu'}(x)$ for all $t > 0$. Thus, by Lemma 4.2, we have that $x \le \lim_{t\to\infty} T^t_{\mu'}(x) = x(\mu') = x'$. Similarly, $x' \le x$, which implies that $x = x'$ and that T has at most one fixed point in X.

To establish the existence of a fixed point, fix a proper policy $\mu \in M$ for the minimizer. (One exists thanks to Assumption SSP.) By Lemma 4.2, we have that $x(\mu) = T_\mu(x(\mu))$. Thus, $x(\mu) \ge T(x(\mu))$. Similarly, by fixing a stationary policy $\nu \in N$ for the maximizer, we obtain from Lemma 4.4 that $\tilde{x}(\nu) = \tilde{T}_\nu(\tilde{x}(\nu))$. Thus, $\tilde{x}(\nu) \le \tilde{T}(\tilde{x}(\nu)) = T(\tilde{x}(\nu))$. We now claim that $\tilde{x}(\nu) \le x(\mu)$. To see this, note that, for every $\pi_M \in \bar{M}$, $\pi_N \in \bar{N}$, and $t > 0$,

$$h^t_{\pi_M, \pi_N} \le \max_{\tilde{\pi}_N \in \bar{N}} h^t_{\pi_M, \tilde{\pi}_N}$$

and

$$h^t_{\pi_M, \pi_N} \ge \min_{\tilde{\pi}_M \in \bar{M}} h^t_{\tilde{\pi}_M, \pi_N},$$

where we have used the notation defined in (2.2). Thus, for any $\pi_N \in \bar{N}$ and for any $\pi_M \in \bar{M}$,

$$\min_{\tilde{\pi}_M \in \bar{M}} h^t_{\tilde{\pi}_M, \pi_N} \le \max_{\tilde{\pi}_N \in \bar{N}} h^t_{\pi_M, \tilde{\pi}_N}.$$

By taking the limit inferior of both sides with respect to t, we see that $\tilde{x}(\pi_N) \leq x(\pi_M)$ for all $\pi_N \in \bar{N}$ and $\pi_M \in \bar{M}$. In particular, $\tilde{x}(\nu) \leq x(\mu)$.

Using the monotonicity of T we have that

$$\tilde{x}(\nu) \leq T(\tilde{x}(\nu)) \leq T(x(\mu)) \leq x(\mu).$$

Again from the monotonicity of T, we obtain for all $t > 1$ that

$$\tilde{x}(\nu) \leq T^{t-1}(\tilde{x}(\nu)) \leq T^t(\tilde{x}(\nu)) \leq x(\mu).$$

Thus, the sequence $\{T^t(\tilde{x}(\nu))\}$ converges to a vector $x^\infty \in X$. From the continuity of T, we have that $x^\infty = T(x^\infty)$. Thus, T has a fixed point in X. □

PROPOSITION 4.6. *The unique fixed point $x^* = T(x^*)$ is the equilibrium cost of the stochastic shortest path game. There exist stationary policies $\mu^* \in M$ and $\nu^* \in N$ which achieve the equilibrium. Moreover, if $x \in X$, $\mu \in M$, and $\nu \in N$ are such that $x = T(x) = T_\mu(x) = \tilde{T}_\nu(x)$, then*

1. $x = x(\mu, \nu)$,
2. $x(\pi_M, \nu) \geq x(\mu, \nu) \quad \forall\, \pi_M \in \bar{M}$,
3. $x(\mu, \pi_N) \leq x(\mu, \nu) \quad \forall\, \pi_N \in \bar{N}$.

Proof. That there exists a unique fixed point $x^* = T(x^*)$ follows from the preceding proposition. Let $\mu^* \in M$ be such that $x^* = T(x^*) = T_{\mu^*}(x^*)$, and let $\nu^* \in N$ be such that $x^* = T(x^*) = \tilde{T}(x^*) = \tilde{T}_{\nu^*}(x^*)$. (Such policies exist thanks to Assumption R.) By Lemma 4.3, we have that μ^* is proper. Thus, by Lemma 4.2, we have that $x^* = x(\mu^*) = \sup_{\pi_N \in \bar{N}} x(\mu^*, \pi_N)$. Similarly, by Lemma 4.4, we have that $x^* = \tilde{x}(\nu^*) = \inf_{\pi_M \in \bar{M}} x(\pi_M, \nu^*)$. Combining these results we obtain

$$\inf_{\pi_M \in \bar{M}} \sup_{\pi_N \in \bar{N}} x(\pi_M, \pi_N) \leq x^* \leq \sup_{\pi_N \in \bar{N}} \inf_{\pi_M \in \bar{M}} x(\pi_M, \pi_N).$$

Since in general we have

$$\inf_{\pi_M \in \bar{M}} \sup_{\pi_N \in \bar{N}} x(\pi_M, \pi_N) \geq \sup_{\pi_N \in \bar{N}} \inf_{\pi_M \in \bar{M}} x(\pi_M, \pi_N)$$

(a statement of the minimax inequality), we obtain the following desired result:

$$\inf_{\pi_M \in \bar{M}} \sup_{\pi_N \in \bar{N}} x(\pi_M, \pi_N) = x^* = \sup_{\pi_N \in \bar{N}} \inf_{\pi_M \in \bar{M}} x(\pi_M, \pi_N). \qquad □$$

Lemma 4.1 implies that, when all stationary policies for the minimizer are proper, the iteration $x^{t+1} = T(x^t)$ converges linearly to the equilibrium cost x^* for all $\mathbf{0} \in X$. This follows from the contraction mapping principle. In the following proposition, we extend this result to the case where not all stationary policies for the minimizer are proper.

PROPOSITION 4.7. *For every $x \in X$, there holds*

$$\lim_{t \to \infty} T^t(x) = x^*, \tag{4.4}$$

where x^* is the unique equilibrium cost vector.

Proof. The uniqueness and existence of a fixed point for T was established in Proposition 4.5. Let x^* be the unique fixed point, and let $\mu^* \in M$ (proper) be such that $T(x^*) = T_{\mu^*}(x^*)$. Our objective is to show that $T^t(x) \to x^*$ for all $x \in X$. Let Δ be the vector with coordinates

$$\Delta_i = \begin{cases} 0 & \text{if } i = 1, \\ \delta & \text{if } i \neq 1, \end{cases} \tag{4.5}$$

where δ is some scalar. Let x^Δ be the unique vector in X satisfying $T_{\mu^*}(x^\Delta) = x^\Delta - \Delta$. (Such a vector exists because μ^* is proper, making the operator $T_{\mu^*}(\cdot) + \Delta$ a contraction.) Note that

$$\begin{aligned} x^\Delta &= T_{\mu^*}(x^\Delta) + \Delta \\ &= \max_{\nu \in N} [c(\mu^*, \nu) + P(\mu^*, \nu) x^\Delta] + \Delta \\ &= \max_{\nu \in N} [c(\mu^*, \nu) + \Delta + P(\mu^*, \nu) x^\Delta]. \end{aligned}$$

Thus, x^Δ is the cost of the fixed policy μ^* with the immediate transition cost $c(\mu^*, \cdot)$ replaced with $c(\mu^*, \cdot) + \Delta$. We have that

$$x^\Delta = T_{\mu^*}(x^\Delta) + \Delta \geq T_{\mu^*}(x^\Delta).$$

Thus, from the monotonicity of T_{μ^*} we have that for all $t > 0$

$$T^t_{\mu^*}(x^\Delta) \leq x^\Delta.$$

By taking the limit as $t \to \infty$, we see that $x(\mu^*) \leq x^\Delta$.

Now using the monotonicity of T and the fact that $x^* = x(\mu^*)$, we get

$$x^* = T(x^*) \leq T(x^\Delta) \leq T_{\mu^*}(x^\Delta) = x^\Delta - \Delta \leq x^\Delta.$$

Proceeding inductively, we get

$$x^* \leq T^t(x^\Delta) \leq T^{t-1}(x^\Delta) \leq x^\Delta.$$

Hence, $\{T^t(x^\Delta)\}$ is a monotonically decreasing sequence bounded below which converges to some $\tilde{x} \in X$. By continuity of the operator T, we must have that $\tilde{x} = T(\tilde{x})$. By the uniqueness of the fixed point of T, we have that $\tilde{x} = x^*$.

We now examine the convergence of the operator T^t applied to $x^* - \Delta$. Note that

$$x^* - \Delta = T(x^*) - \Delta \leq T(x^* - \Delta) \leq T(x^*) = x^*,$$

where the first inequality follows from the fact that $P(\mu, \nu)\Delta \leq \Delta$ for all $\mu \in M$ and $\nu \in N$. Once again monotonicity of T prevails, implying that $T^t(x^* - \Delta)$ is monotonically increasing and bounded above. From the continuity of T we have that $\lim_{t \to \infty} T^t(x^* - \Delta) = x^*$.

We saw earlier that $x^\Delta = T_{\mu^*}(x^\Delta) + \Delta$ and that $x^\Delta \geq x^*$. Then, from the monotonicity of T_{μ^*},

$$x^\Delta \geq T_{\mu^*}(x^*) + \Delta = x^* + \Delta.$$

Thus, for any $x \in X$, we can find $\delta > 0$ such that $x^* - \Delta \leq x \leq x^\Delta$. By the monotonicity of T, we then have

$$T^t(x^* - \Delta) \leq T^t(x) \leq T^t(x^\Delta) \quad \forall\, t \geq 1.$$

Taking limits, we see that $\lim_{t\to\infty} T^t(x) = x^*$. □

PROPOSITION 4.8. *Given a proper stationary policy $\mu^0 \in M$, we have that*

$$x(\mu^k) \to x^*,$$

where x^* is the unique equilibrium cost vector and $\{\mu^k\}$ is a sequence of policies (generated by policy iteration) such that $T(x(\mu^k)) = T_{\mu^{k+1}}(x(\mu^k))$ for all k.

Proof. Choose $\mu^1 \in M$ such that $T_{\mu^1}(x(\mu^0)) = T(x(\mu^0))$. (Assumption SSP implies that such an initial proper policy μ^0 exists.) We have $T_{\mu^1}(x(\mu^0)) = T(x(\mu^0)) \leq T_{\mu^0}(x(\mu^0)) = x(\mu^0)$. By Lemma 4.3, μ^1 is proper. By the monotonicity of T_{μ^1} and Lemma 4.2, we have that for all t

$$x(\mu^0) \geq T(x(\mu^0)) \geq T_{\mu^1}^{t-1}(x(\mu^0)) \geq T_{\mu^1}^{t}(x(\mu^0)).$$

Thus,

$$x(\mu^0) \geq T(x(\mu^0)) \geq \lim_{t\to\infty} T_{\mu^1}^{t}(x(\mu^0)) = x(\mu^1).$$

Applying this argument iteratively, we construct a sequence $\{\mu^k\}$ of proper policies such that

$$x(\mu^k) \geq T(x(\mu^k)) \geq x(\mu^{k+1}) \geq x^* \quad \forall\, k = 0, 1, \ldots\ . \tag{4.6}$$

Since $\{x(\mu^k)\}$ is monotonically decreasing and bounded below by x^*, we have that the entire sequence converges to some vector x^∞. From (4.6) and the continuity of T, we have that $x^\infty = T(x^\infty)$. Since x^* is the unique fixed point of T on X, we have that $x(\mu^k) \to x^*$. □

5. An example of pursuit and evasion.

Consider the following two-player game, played around a table with four corners. One player, the pursuer (who is actually the minimizer), is attempting to "catch" in minimum time the other player, the evader (who is the maximizer). The game evolves in stages where, in each stage, both players implement actions simultaneously. When the players are across from one another, they each decide (independently) whether to stay where they are, move one corner clockwise, or move one corner counter-clockwise. When the two players are adjacent to one another, they each decide (independently) whether to stay where they are, move toward the other's current location, or move away from the other's current location. The pursuer catches the evader only by arranging to land on the same corner of the table as the evader. (The possibility exists that, when they are adjacent, they can both move toward each other's current location. This does not result in the evader being caught "in midair.") The evader is slower than the pursuer in the sense that when the evader decides to change location, he succeeds in doing so only with probability $p \in (0, 1)$. (With probability $1 - p$, the evader will wind up not moving at all.) Thus, the pursuer can ultimately catch the evader, provided he implements an appropriate policy (such as "always move toward the present location of the evader"). On the other hand, there exist policies for the pursuer (such as "always stay put") which allow the maximizer to prolong the game indefinitely. This results in infinite cost (i.e., infinite capture time) to the pursuer.

This game fits into our framework for stochastic shortest path games. As described above there are three states: evader caught (state 1), players adjacent to one another (state 2), and players across from one another (state 3). Thus, $S = \{1, 2, 3\}$. Once the evader is caught the game is over, so state 1 serves as the terminal state, which is zero cost and absorbing.

In state two, when the players are adjacent, the players may move toward the other's location (action 1), stay where they are (action 2), or move away from the other's location (action 3). Thus, $A(2) = B(2) = \{1, 2, 3\}$. From the description of the problem given above, it is not hard to see that

$$\begin{aligned} p_{21}(u, v) &= u_1[(v_1 + v_3)(1 - p) + v_2] + u_2 v_1 p, \\ p_{22}(u, v) &= (u_1 + u_3)(v_1 + v_3)p + u_2[(v_1 + v_3)(1 - p) + v_2], \\ p_{23}(u, v) &= u_2 v_3 p + u_3[(v_1 + v_3)(1 - p) + v_2]. \end{aligned}$$

The expected transition cost functions $c_2(u, v)$ take on the value of one for all $u \in U(2)$ and $v \in V(2)$.

In state three (when the players are on opposite corners of the table), the players may move clockwise toward the other's current location (action 1), stay where they are (action 2), or move counter-clockwise toward the other's location (action 3). Thus, $A(3) = B(3) = \{1, 2, 3\}$. It is not hard to see that

$$\begin{aligned} p_{31}(u, v) &= u_1 v_3 p + u_3 v_1 p, \\ p_{32}(u, v) &= (u_1 + u_3)[(v_1 + v_3)(1 - p) + v_2] + u_2(v_1 + v_3)p, \\ p_{33}(u, v) &= u_1 v_1 p + u_2[(v_1 + v_3)(1 - p) + v_2] + u_3 v_3 p. \end{aligned}$$

The expected transition cost functions $c_3(u, v)$ take on the value of one for all $u \in U(3)$ and $v \in V(3)$.

We will show that the equilibrium value of this stochastic shortest path game is given by

$$x^* = \left(0, \frac{1}{1 - p}, \frac{2 - p}{1 - p}\right)'$$

and that equilibrium randomized strategies for the two players are given by $\mu^* \in M$ and $\nu^* \in N$ such that

$$\begin{aligned} \mu^*(2) &= (1, 0, 0)', \\ \nu^*(2) &= (v_1, 0, v_3)', \\ \mu^*(3) &= (u_1, 0, u_3)', \\ \nu^*(3) &= (0, 1, 0)', \end{aligned}$$

where v_1, v_3, u_1, and u_3 are nonnegative and $v_1 + v_3 = 1$ and $u_1 + u_3 = 1$. Thus, any probability vector $v \in V(2)$ such that $v_2 = 0$ forms an equilibrium strategy for the evader. In other words, as long as the evader chooses not to remain at his current location (when the two players are adjacent), any mixed decision (at state 2) for the evader is optimal. The pursuer does not have the same flexibility; his optimal mixed decision is deterministic: always move toward the evader. On the other hand, any probability vector $u \in U(3)$ such that $u_2 = 0$ forms an equilibrium strategy for the pursuer. In other words, as long as the pursuer decides to not stay at his current location (when the two players are across from one another), any mixed decision for the pursuer (at state 3) is optimal. This time, it is the evader's strategy which is inflexible. His optimal action is to stay at his current location. Thus, when both players play optimally, the game will always transition from state $i = 3$ to $i = 2$ in one stage. Happily, the equilibrium cost reflects this: $x_3^* = \frac{2-p}{1-p} = 1 + x_2^*$.

To verify that these are indeed equilibrium policies, we will show that $x^* = T(x^*) = T_{\mu^*}(x^*) = \tilde{T}_{\nu^*}(x^*)$. (Notice that the policy μ^* corresponds to one where the pursuer always decides to move in the direction of the current location of the evader. This policy is clearly proper. The desired result follows from Proposition 4.6.)

Let us first consider the case where the two players are adjacent (i.e., state 2). Let a general cost-to-go vector be given as $x = (0, x_2, x_3)' \in X$. (Shortly, we shall consider the case where $x = x^*$, as suggested above.) To evaluate the second element of $T(x)$, we must compute

$$\min_{u \in U(2)} \max_{v \in V(2)} u' G_2(x) v,$$

where the matrix $G_2(x)$ is computed as

$$G_2(x) = \begin{bmatrix} 1 + px_2 & 1 & 1 + px_2 \\ 1 + (1-p)x_2 & 1 + x_2 & 1 + (1-p)x_2 + px_3 \\ 1 + px_2 + (1-p)x_3 & 1 + x_3 & 1 + px_2 + (1-p)x_3 \end{bmatrix}.$$

In other words, the second element of $T(x)$ is evaluated as the value of the matrix game (in mixed strategies) defined by $G_2(x)$. It is well known that the equilibrium cost and equilibrium strategies for a matrix game are characterized as the optimal value and solutions to a particular linear program and its dual [12]. In particular,

$$\frac{1}{[T(x)]_2} = \begin{array}{l} \min e'\breve{v} \\ \text{subject to } \; G_2(x)\breve{v} \geq e, \; \breve{v} \geq 0, \end{array}$$

$$\frac{v^*}{[T(x)]_2} \in \begin{array}{l} \arg\min e'\breve{v} \\ \text{subject to } \; G_2(x)\breve{v} \geq e, \; \breve{v} \geq 0, \end{array}$$

where e is the vector of all ones in R^3 and v^* is an equilibrium strategy for the maximizer in the matrix game. We shall refer to the linear program above as the "primal" problem. The dual of the primal problem characterizes equilibrium strategies u^* for the minimizer of the matrix game, as follows:

$$\frac{u^*}{[T(x)]_2} \in \begin{array}{l} \arg\max e'\breve{u} \\ \text{subject to } \; G_2(x)'\breve{u} \leq e, \; \breve{u} \geq 0. \end{array}$$

Now consider $G_2(x^*)$ and define

$$\begin{aligned} \breve{u}^* = \mu^*(2)/x_2^* \; &= \; (1-p)\,(1,0,0)', \\ \breve{v}^* = \nu^*(2)/x_2^* \; &= \; (1-p)\,(v_1,0,v_3)'. \end{aligned}$$

It is straightforward to verify that $\breve{v}^*$ is feasible for the primal linear program and that $\breve{u}^*$ is feasible for the dual problem. Moreover, the primal cost corresponding to $\breve{v}^*$ is exactly $1-p$, just as the dual value of $\breve{u}^*$ is also exactly $1-p$. Thus, we have found a primal/dual feasible pair for which the primal cost equals the dual value. Then, according to the duality theorem of linear programming, $\breve{v}^*$ and $\breve{u}^*$ are primal/dual optimal and the optimal values of the primal and dual problems equal $1-p$, which is exactly $\frac{1}{x_2^*}$. This verifies that $x_2^* = [T(x^*)]_2$ and that $\mu^*(2)$ and $\nu^*(2)$ form an equilibrium pair of mixed decisions at state $2 \in S$.

Let us now consider the case where the two players are across from one another (i.e., state 3). To evaluate the third element of $T(x)$ for general $x \in X$, we must compute

$$\min_{u \in U(3)} \max_{v \in V(3)} u' G_3(x) v,$$

where $G_3(x)$ is a matrix computed as

$$G_3(x) = \begin{bmatrix} 1+(1-p)x_2+px_3 & 1+x_2 & 1+(1-p)x_2 \\ 1+px_2+(1-p)x_3 & 1+x_3 & 1+px_2+(1-p)x_3 \\ 1+(1-p)x_2 & 1+x_2 & 1+(1-p)x_2+px_3 \end{bmatrix}.$$

Thus, the third element of $T(x)$ is evaluated as the value of the matrix game defined by $G_3(x)$. As before, the solution to this matrix game can be computed by solving a primal/dual pair of linear programs, as follows:

$$\begin{aligned} &\min e'\breve{v} \\ &\text{subject to } \; G_3(x)\breve{v} \geq e, \; \breve{v} \geq 0, \end{aligned}$$

$$\begin{aligned} &\max e'\breve{u} \\ &\text{subject to } \; G_3(x)'\breve{u} \leq e, \; \breve{u} \geq 0. \end{aligned}$$

Now consider the primal and dual problems given by $G_3(x^*)$. Define

$$\begin{aligned} \breve{u}^* = \mu^*(3)/x_3^* &= \frac{1-p}{2-p}\,(u_1, 0, u_3)', \\ \breve{v}^* = \nu^*(3)/x_3^* &= \frac{1-p}{2-p}\,(0, 1, 0)'. \end{aligned}$$

Again, it is straightforward to verify that $\breve{v}$ and $\breve{u}$ form a feasible primal/dual pair where the primal cost of $\breve{v}$ equals the dual value of $\breve{u}$. Thus, by the duality theorem, $\breve{v}$ and $\breve{u}$ are primal/dual optimal. This time the optimal cost works out to be $\frac{1-p}{2-p}$, which is exactly $\frac{1}{x_3^*}$. This verifies that $x_3^* = [T(x^*)]_3$ and that $\mu^*(3)$ and $\nu^*(3)$ form an equilibrium pair of mixed decisions at state $3 \in S$.

Appendix. Proofs of lemmas. We collect here some useful but well-known results. We give proofs for completeness. We require Assumption R throughout.

The following lemmas summarize some important properties of the operators $T_{\mu\nu}$, T_μ, T, $\tilde{T}_\nu$, and $\tilde{T}$.

Lemma A.1 (monotonicity). *Suppose $x \in R^n$ and $x' \in R^n$ (or $x \in X$ and $x' \in X$) are such that $x \leq x'$ is componentwise. Then*

$$T_{\mu\nu}(x) \leq T_{\mu\nu}(x'), \quad \mu \in M, \nu \in N, \tag{A.1}$$

$$T_\mu(x) \leq T_\mu(x'), \quad \mu \in M, \tag{A.2}$$

$$T(x) \leq T(x'), \tag{A.3}$$

$$\tilde{T}_\nu(x) \leq \tilde{T}_\nu(x'), \quad \nu \in N, \tag{A.4}$$

$$\tilde{T}(x) \leq \tilde{T}(x'). \tag{A.5}$$

Proof. This proof is straightforward using the definitions of various "T"-operators. □

LEMMA A.2 (cost shifting). *For all $x \in R^n$, scalars $r \in R$, integers $t > 0$, and functions $\mu^k \in M$ and $\nu^k \in N$ for $k = 1, \dots, t$ we have*

$$(T_{\mu^1\nu^1}T_{\mu^2\nu^2}\cdots T_{\mu^t\nu^t})(x + r\cdot\mathbf{1}) = (T_{\mu^1\nu^1}T_{\mu^2\nu^2}\cdots T_{\mu^t\nu^t})(x) + r\cdot\mathbf{1}, \tag{A.6}$$

$$(T_{\mu^1}T_{\mu^2}\cdots T_{\mu^t})(x + r\cdot\mathbf{1}) = (T_{\mu^1}T_{\mu^2}\cdots T_{\mu^t})(x) + r\cdot\mathbf{1}, \tag{A.7}$$

$$(TT\cdots T)(x + r\cdot\mathbf{1}) = (TT\cdots T)(x) + r\cdot\mathbf{1}, \tag{A.8}$$

where $\mathbf{1} = (1, \dots, 1)' \in R^n$. The same relationships hold for T_μ, $T_{\mu\nu}$, $\tilde{T}_{nu}$, and $\tilde{T}$.

Proof. This follows by induction and by the definition of $T_{\mu\nu}$, T_μ, and T. □

LEMMA A.3 (continuity). *The mappings T, T_μ, $T_{\mu\nu}$, $\tilde{T}_{nu}$, and $\tilde{T}$ are continuous over R^n.*

Proof. Let x and x' be any two elements of R^n, and let $r = \|x - x'\|_\infty$, where $\|\cdot\|_\infty$ denotes the usual sup-norm (l_∞-norm) on X. Then we have

$$x - r\cdot\mathbf{1} \le x' \le x + r\cdot\mathbf{1},$$

where $\mathbf{1} = (1, \dots, 1)' \in X$. Lemmas A.1 and A.2 imply that

$$\begin{aligned} T(x) - r\cdot\mathbf{1} \le T(x') &\le T(x) + r\cdot\mathbf{1}, \\ T_\mu(x) - r\cdot\mathbf{1} \le T_\mu(x') &\le T_\mu(x) + r\cdot\mathbf{1}, \\ T_{\mu\nu}(x) - r\cdot\mathbf{1} \le T_{\mu\nu}(x') &\le T_\mu(x) + r\cdot\mathbf{1}. \end{aligned}$$

Therefore,

$$\begin{aligned} \|T(x) - T(x')\|_\infty &\le \|x - x'\|_\infty, \\ \|T_\mu(x) - T_\mu(x')\|_\infty &\le \|x - x'\|_\infty, \\ \|T_{\mu\nu}(x) - T_{\mu\nu}(x')\|_\infty &\le \|x - x'\|_\infty. \end{aligned}$$

Thus, T is continuous on R^n. Similar arguments hold for T_μ, $T_{\mu\nu}$, $\tilde{T}_{nu}$, and $\tilde{T}$. □

The remainder of this appendix examines finite-horizon dynamic games where at each stage the maximizer has access to the minimizer's decision. We show that minimax and maximin versions of these games can be solved in a straightforward manner through dynamic programming. In doing so, we prove several results relevant to the main body of this paper.

LEMMA A.4. *Let M, N, T, T_μ, and $\tilde{T}_\nu$ all be defined as in previous sections. Then, for any square matrix of nonnegative elements $\bar{P}$ and any $x \in X$, we have*

$$\begin{aligned} \min_{\mu\in M}\max_{\nu\in N} \bar{P}\left[c(\mu,\nu) + P(\mu,\nu)x\right] &= \bar{P}\min_{\mu\in M}\max_{\nu\in N}\left[c(\mu,\nu) + P(\mu,\nu)x\right] = \bar{P}T(x), \\ \max_{\nu\in N} \bar{P}\left[c(\mu,\nu) + P(\mu,\nu)x\right] &= \bar{P}\max_{\nu\in N}\left[c(\mu,\nu) + P(\mu,\nu)x\right] = \bar{P}T_\mu(x), \\ \min_{\mu\in M} \bar{P}\left[c(\mu,\nu) + P(\mu,\nu)x\right] &= \bar{P}\min_{\mu\in M}\left[c(\mu,\nu) + P(\mu,\nu)x\right] = \bar{P}\tilde{T}_\nu(x). \end{aligned}$$

Proof. It is sufficient to show that the first equation holds. The remaining equations follow as corollaries by redefining the control constraint sets for the minimizing and maximizing players as $\tilde{U}(i) = \{\mu(i)\}$ and $\tilde{V}(i, u) = \{\nu(i)\}$, respectively.

The ith component of $\bar{P}\left[c(\mu,\nu) + P(\mu,\nu)x\right]$ can be expressed as

$$\sum_{s=1}^{n} \bar{p}_{is} g_s(\mu(s), \nu(s)),$$

where $\bar{p}_{is}$ is the $(i \times s)$th component of $\bar{P}$ and $g_s(u,v) \overset{\triangle}{=} c_s(u,v) + \sum_{j=1}^{n} p_{sj}(u,v)x_j$ for $u \in U(s)$ and $v \in V(s)$.

Since the min and max are taken componentwise and since the elements of P are nonnegative, we have that

$$\begin{aligned}\min_{\mu\in M}\max_{\nu\in N}\sum_{s=1}^{n}\bar{p}_{is}g_s(\mu(s),\nu(s)) &= \min_{\mu\in M}\max_{v^1\in V(1),\dots,v^n\in V(n)}\sum_{s=1}^{n}\bar{p}_{is}g_s(\mu(s),v^s)\\ &= \min_{\mu\in M}\sum_{s=1}^{n}\bar{p}_{is}\max_{v^s\in V(s)} g_s(\mu(s),v^s).\end{aligned}$$

Similarly, because the elements of $\bar{P}$ are nonnegative,

$$\begin{aligned}\min_{\mu\in M}\sum_{s=1}^{n}\bar{p}_{is}\max_{v^s\in V(s)} g_s(\mu(s),v^s) &= \min_{u^1\in U(1),\dots,u^n\in U(n)}\sum_{s=1}^{n}\bar{p}_{is}\max_{v^s\in V(s)} g_s(u^s,v^s)\\ &= \sum_{s=1}^{n}\bar{p}_{is}\min_{u^s\in U(s)}\max_{v^s\in V(s)} g_s(u^s,v^s)\\ &= \sum_{s=1}^{n}\bar{p}_{is}[T(x)]_s.\end{aligned}$$

Since this same expression applies for all $i = 1,\dots,n$, the desired result holds. □

LEMMA A.5. *Let M, N, T, T_μ, and $\tilde{T}_\nu$ all be defined as in previous sections. Then, for any $x \in X$, we have*

$$\begin{aligned}\min_{\pi_M=\{\mu^0,\dots,\mu^t\}}\max_{\pi_N=\{\nu^0,\dots,\nu^t\}}\left[h^t_{\pi_M,\pi_N} + P(\mu^0,\nu^0)\cdots P(\mu^t,\nu^t)x\right] &= T^{t+1}(x),\\ \max_{\pi_N=\{\nu^0,\dots,\nu^t\}}\left[h^t_{\mu,\pi_N} + P(\mu,\nu^0)\cdots P(\mu,\nu^t)x\right] &= T^{t+1}_\mu(x),\\ \min_{\pi_M=\{\mu^0,\dots,\mu^t\}}\left[h^t_{\pi_M,\nu} + P(\mu^0,\nu)\cdots P(\mu^t,\nu)x\right] &= \tilde{T}^{t+1}_\nu(x),\end{aligned}$$

where μ and the μ^k are elements of M and ν and the ν^k are elements of N.

Proof. It is sufficient to show that the first equation holds. The remaining equations follow as corollaries by redefining the control constraint sets for the minimizing and maximizing players as $\tilde{U}(i) = \{\mu(i)\}$ and $\tilde{V}(i) = \{\nu(i)\}$, respectively.

Notice that

$$\begin{aligned}&\min_{\pi_M=\{\mu^0,\dots,\mu^t\}}\max_{\pi_N=\{\nu^0,\dots,\nu^t\}}\left[h^t_{\pi_M,\pi_N} + P(\mu^0,\nu^0)\cdots P(\mu^t,\nu^t)x\right]\\ &\quad= \min_{\pi_M}\max_{\pi_N}\left\{h^{t-1}_{\bar{\pi}_M,\bar{\pi}_N} + \bar{P}(\bar{\pi}_M,\bar{\pi}_N)\left[c(\mu^t,\nu^t) + P(\mu^t,\nu^t)x\right]\right\}\\ &\quad= \min_{\pi_M}\max_{\bar{\pi}_N}\left\{h^{t-1}_{\bar{\pi}_M,\bar{\pi}_N} + \max_{\nu^t}\bar{P}(\bar{\pi}_M,\bar{\pi}_N)\left[c(\mu^t,\nu^t) + P(\mu^t,\nu^t)x\right]\right\}\\ &\quad= \min_{\bar{\pi}_M}\min_{\mu^t}\max_{\bar{\pi}_N}\left\{h^{t-1}_{\bar{\pi}_M,\bar{\pi}_N} + \max_{\nu^t}\bar{P}(\bar{\pi}_M,\bar{\pi}_N)\left[c(\mu^t,\nu^t) + P(\mu^t,\nu^t)x\right]\right\}\\ &\quad\geq \min_{\bar{\pi}_M}\max_{\bar{\pi}_N}\min_{\mu^t}\left\{h^{t-1}_{\bar{\pi}_M,\bar{\pi}_N} + \max_{\nu^t}\bar{P}(\bar{\pi}_M,\bar{\pi}_N)\left[c(\mu^t,\nu^t) + P(\mu^t,\nu^t)x\right]\right\}\\ &\quad= \min_{\bar{\pi}_M}\max_{\bar{\pi}_N}\left\{h^{t-1}_{\bar{\pi}_M,\bar{\pi}_N} + \min_{\mu^t}\max_{\nu^t}\bar{P}(\bar{\pi}_M,\bar{\pi}_N)\left[c(\mu^t,\nu^t) + P(\mu^t,\nu^t)x\right]\right\},\end{aligned}$$

where

$$\bar{\pi}_M \overset{\triangle}{=} \{\mu^0, \dots, \mu^{t-1}\}, \qquad \bar{\pi}_N \overset{\triangle}{=} \{\nu^0, \dots, \nu^{t-1}\},$$

and $\bar{P}(\bar{\pi}_M, \bar{\pi}_N) \overset{\triangle}{=} P(\mu^0, \nu^0) \cdots P(\mu^{t-1}, \nu^{t-1})$. (The inequality follows from the minimax inequality.)

We now prove the reverse relationship. First, we claim there exists a policy $\bar{\mu} \in M$ such that

$$\min_{\mu^t \in M} \max_{\nu^t \in N} \bar{P}(\bar{\pi}_M, \bar{\pi}_N)[c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x] = \max_{\nu^t \in N} \bar{P}(\bar{\pi}_M, \bar{\pi}_N)[c(\bar{\mu}, \nu^t) + P(\bar{\mu}, \nu^t)x].$$

To see this, notice that

$$\begin{aligned} &\min_{\mu_t \in M} \max_{\nu_t \in N} \bar{P}(\bar{\pi}_M, \bar{\pi}_N) \left[c(\mu_t, \nu_t) + P(\mu_t, \nu_t)x\right] \\ &\qquad = \bar{P}(\bar{\pi}_M, \bar{\pi}_N) \min_{\mu_t \in M} \max_{\nu_t \in N} \left(c(\mu_t, \nu_t) + P(\mu_t, \nu_t)x\right) \\ &\qquad = \bar{P}(\bar{\pi}_M, \bar{\pi}_N) \max_{\nu_t \in N} \left(c(\bar{\mu}, \nu_t) + P(\bar{\mu}, \nu_t)x\right) \\ &\qquad = \max_{\nu_t \in N} \bar{P}(\bar{\pi}_M, \bar{\pi}_N) \left[c(\bar{\mu}, \nu_t) + P(\bar{\mu}, \nu_t)x\right], \end{aligned}$$

where the first and last equalities follow from the preceding lemma and $\bar{\mu}$ is the minimax solution to $\min_{\mu_t \in M} \max_{\nu_t \in N} \left(c(\mu_t, \nu_t) + P(\mu_t, \nu_t)x\right)$. This completes the proof of our claim. Thus,

$$\begin{aligned} &\min_{\bar{\pi}_M} \min_{\mu^t} \max_{\bar{\pi}_N} \left\{ h^{t-1}_{\bar{\pi}_M, \bar{\pi}_N} + \max_{\nu^t} \bar{P}(\bar{\pi}_M, \bar{\pi}_N) \left[c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x\right] \right\} \\ &\quad = \min_{\bar{\pi}_M} \max_{\bar{\pi}_N} \left\{ h^{t-1}_{\bar{\pi}_M, \bar{\pi}_N} + \min_{\mu^t} \max_{\nu^t} \bar{P}(\bar{\pi}_M, \bar{\pi}_N) \left[c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x\right] \right\} \\ &\quad \le \min_{\bar{\pi}_M} \max_{\bar{\pi}_N} \left\{ h^{t-1}_{\bar{\pi}_M, \bar{\pi}_N} + \max_{\nu^t} \bar{P}(\bar{\pi}_M, \bar{\pi}_N) \left[c(\bar{\mu}, \nu^t) + P(\bar{\mu}, \nu^t)x\right] \right\} \\ &\quad = \min_{\bar{\pi}_M} \max_{\bar{\pi}_N} \left\{ h^{t-1}_{\bar{\pi}_M, \bar{\pi}_N} + \min_{\mu^t} \max_{\nu^t} \bar{P}(\bar{\pi}_M, \bar{\pi}_N) \left[c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x\right] \right\}. \end{aligned}$$

Combining the preceding inequalities, we see that

$$\begin{aligned} &\min_{\pi_M} \max_{\pi_N} \left[h^t_{\pi_M, \pi_N} + P(\mu^0, \nu^0) \cdots P(\mu^t, \nu^t)x\right] \\ &\quad = \min_{\bar{\pi}_M} \max_{\bar{\pi}_N} \left\{ h^{t-1}_{\bar{\pi}_M, \bar{\pi}_N} + \min_{\mu^t} \max_{\nu^t} \bar{P}(\bar{\pi}_M, \bar{\pi}_N) \left[c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x\right] \right\} \\ &\quad = \min_{\bar{\pi}_M} \max_{\bar{\pi}_N} \left\{ h^{t-1}_{\bar{\pi}_M, \bar{\pi}_N} + \bar{P}(\bar{\pi}_M, \bar{\pi}_N) \min_{\mu^t} \max_{\nu^t} \left[c(\mu^t, \nu^t) + P(\mu^t, \nu^t)x\right] \right\} \\ &\quad = \min_{\bar{\pi}_M} \max_{\bar{\pi}_N} \left[h^{t-1}_{\bar{\pi}_M, \bar{\pi}_N} + P(\mu^0, \nu^0) \cdots P(\mu^{t-1}, \nu^{t-1}) T(x)\right], \end{aligned}$$

where the second inequality follows from Lemma A.4.

Mathematical induction, repeating the same argument above, gives the desired result. □

LEMMA A.6. *Let M, N, and $\tilde{T}$ all be defined as in previous sections. Then, for any square matrix with nonnegative elements $\bar{P}$ and any $x \in X$,*

$$\max_{\nu \in N} \min_{\mu \in M} \bar{P} \left[c(\mu,\nu) + P(\mu,\nu)x \right] = \bar{P} \max_{\nu \in N} \min_{\mu \in M} \left[c(\mu,\nu) + P(\mu,\nu)x \right] = \bar{P}\tilde{T}(x).$$

Proof. The proof of this lemma is exactly analogous to that given for Lemma A.4. The interchange of the max and min has no bearing on the logical flow of the argument. □

LEMMA A.7. *Let M, N, and $\tilde{T}$ all be defined as in previous sections. Then, for any $x \in X$,*

$$\max_{\pi_n=\{\nu^0,\dots,\nu^t\}} \min_{\pi_M=\{\mu^0,\dots,\mu^t\}} \left[h^t_{\pi_M,\pi_N} + P(\mu^0,\nu^0)\cdots P(\mu^t,\nu^t)x \right] \quad = \quad \tilde{T}^{t+1}(x),$$

where the μ^k are elements of M and the ν^k are elements of N.

Proof. The proof of this is symmetrical to that given for Lemma A.5. □

Using the fact that $T(x) = \tilde{T}(x)$ (under Assumption R), we obtain

$$\begin{aligned}
&\min_{\pi_M=\{\mu^0,\dots,\mu^t\}} \max_{\pi_N=\{\nu^0,\dots,\nu^t\}} \left[h^t_{\pi_M,\pi_N} + P(\mu^0,\nu^0)\cdots P(\mu^t,\nu^t)x+ \right] \\
&\qquad = T^{t+1}(x) \\
&\qquad = \tilde{T}^{t+1}(x) \\
&\qquad = \max_{\pi_N=\{\nu^0,\dots,\nu^t\}} \min_{\pi_M=\{\mu^0,\dots,\mu^t\}} \left[h^t_{\pi_M,\pi_N} + P(\mu^0,\nu^0)\cdots P(\mu^t,\nu^t)x \right].
\end{aligned}$$

Thus, for finite-horizon games (with or without a terminal state), an equilibrium cost exists and can be found via dynamic programming iterations.

Acknowledgments. Our proof of Lemma 4.1 was inspired by an argument by John Tsitsiklis for a similar result in one-player stochastic shortest path problems. We would like to thank our anonymous SIAM reviewers who, through their persistence, have helped us to find the shortest path to establishing our results.

REFERENCES

[1] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Parallel and Distributed Computation: Numerical Methods*, Prentice–Hall, Englewood Cliffs, NJ, 1989.

[2] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Analysis of stochastic shortest path problems*, Math. Oper. Res., 16 (1991), pp. 580–595.

[3] P. R. KUMAR AND T. H. SHIAU, *Zero sum dynamic games*, in Control and Dynamic Games, C. T. Leondes, ed., Academic Press, New York, 1981, pp. 1345–1378.

[4] H. J. KUSHNER AND S. G. CHAMBERLAIN, *Finite state stochastic games: Existence theorems and computational procedures*, IEEE Trans. Automat. Control, AC-14, 1969, pp. 248–255.

[5] J. F. MERTENS AND A. NEYMAN, *Stochastic games*, Internat. J. Game Theory, 10 (1980), pp. 53–66.

[6] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[7] L. S. SHAPLEY, *Stochastic games*, Proc. Nat. Acad. Sci., Mathematics, 39 (1953), pp. 1095–1100.

[8] J. STOER AND C. WITZGALL, *Convexity and Optimization in Finite Dimensions* I, Springer-Verlag, New York, 1970.

[9] P. TSENG, *Solving H-horizon, stationary Markov decision problems in time proportional to $\log(H)$*, Oper. Res. Lett., 9 (1990), pp. 287–297.

[10] J. VAN DER WAL, *Stochastic Dynamic Programming*, Mathematical Centre Tracts 139, Mathematisch Centrum, Amsterdam, 1981.

[11] J. VON NEUMANN AND O. MORGENSTERN, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ, 1944.

[12] N. N. Vorob'ev, *Game Theory, Lectures for Economists and Systems Scientists*, Springer-Verlag, New York, 1977.

[13] O. J. Vrieze, *Stochastic Games with Finite State and Action Spaces*, CWI Tract 33, Centrum voor Wiskunde en Informatica (Center for Mathematics and Computer Science), 1009 AB Amsterdam, The Netherlands, 1987.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 3, pp. 825–843

FULLY COUPLED FORWARD-BACKWARD STOCHASTIC DIFFERENTIAL EQUATIONS AND APPLICATIONS TO OPTIMAL CONTROL[*]

SHIGE PENG[†] AND ZHEN WU[†]

Abstract. Existence and uniqueness results of fully coupled forward-backward stochastic differential equations with an arbitrarily large time duration are obtained. Some stochastic Hamilton systems arising in stochastic optimal control systems and mathematical finance can be treated within our framework.

Key words. stochastic differential equations, Hamilton system, stochastic optimal control, stochastic analysis

AMS subject classifications. 60H, 93E

PII. S0363012996313549

1. Introduction.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\{B_t\}_{t\geq 0}$ be a d-dimensional Brownian motion in this space. We denote the natural filtration of this Brownian motion by $\mathcal{F}_t$. In this paper, we consider the following fully coupled forward-backward stochastic differential equation (FBSDE):

$$\begin{aligned} x_t &= a + \int_0^t b(s, x_s, y_s, z_s)ds + \int_0^t \sigma(s, x_s, y_s, z_s)dB_s, \\ y_t &= \Phi(x_T) + \int_t^T f(s, x_s, y_s, z_s)ds - \int_t^T z_s dB_s, \quad t \in [0, T], \end{aligned}$$

where (x, y, z) takes values in $\mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^{m\times d}$, and b, and σ, and f are mappings with appropriate dimensions which are, for each fixed (x, y, z), $\mathcal{F}_t$-progressively measurable. We assume that they are Lipschitz with respect to (x, y, z); $T > 0$ is an arbitrarily prescribed number and the time interval is called the *time duration.* We look for a triple of $\mathcal{F}_t$-adapted processes (x_s, y_s, z_s) satisfying this equation.

A special situation of this problem is the well-known forward and backward ordinary differential equations (FBODEs):

$$\begin{aligned} \dot{x}(t) &= \overline{b}(x(t), y(t)), \\ -\dot{y}(t) &= \overline{f}(x(t), y(t)), \\ x(0) &= a, \quad y(T) = \overline{\Phi}(X(T)). \end{aligned}$$

The so-called two-point boundary problem for the second-order ODE is a special setting of this problem. There exist many examples in FBODE showing that, even if the uniform Lipschitz condition is imposed for all coefficients $\overline{b}$, $\overline{f}$, and $\overline{g}$, the existence and/or uniqueness of this ODE (with an arbitrarily large time duration) may fail.

One of the few types of such equations which is well understood is the Hamiltonian system motivated by the problems of classical variation method as well as the

[*] Received by the editors December 13, 1996; accepted for publication (in revised form) February 5, 1998; published electronically March 5, 1999. This research was partially supported by the National Natural Science Foundation of China, program no. 79790130.

http://www.siam.org/journals/sicon/37-3/31354.html

[†] School of Mathematics and System Sciences, Shandong University, Jinan 250100, People's Republic of China (peng@public.jn.sd.cn, peng@sdu.edu.cn, wuzhen@math.sdu.edu.cn).

maximum principle in optimal control problems (see Pontryagin et al. [Pon]). For the linear deterministic case, see Kalman [Ka] for the motivation of control problems in linear quadratic optimal regulators. Parallel to this deterministic situation, Bismut [Bis] obtained stochastic Hamiltonian systems, a kind of extension of the above Hamiltonian systems, as well as that of the Pontryagin's maximum principle (see also Kushner [Ku], Bensoussan [Ben], Haussmann [Hau], Peng [P1]). The corresponding linear case was also discussed.

Recently Antonelli gave a counterexample (see [A]) showing that the Lipschitz condition is not enough for the existence of FBSDE in an arbitrarily large time duration. It is clear that more assumptions are essentially needed.

An interesting problem is, What is the more suitable setting of the above stochastic Hamiltonian problems for which the solution exists and is unique? Is the background of optimization problem necessary? Is the constraint $m = n$ necessary?

In this paper we present a probabilistic method to treat a large kind of FBSDE with an arbitrarily prescribed time duration. Many problems like the above-mentioned FBODE and FBSDE are special situations of our setting. It is due to a new observation: for a large kind of FBSDE raised in stochastic control systems, mathematical finance, and other applied mathematics, the coefficients f, b, and σ are G-monotone, i.e., there exists a nondegenerate $m \times n$-matrix G such that, for each fixed (ω, t), the mapping

$$A(x,y,z) \overset{def}{=} \begin{pmatrix} -G^T f \\ Gb \\ G\sigma \end{pmatrix} (x,y,z): \quad R^{n+m+m\times d} \longrightarrow R^{n+m+m\times d}$$

is monotonous in (x, y, z) in the sense of (H2.2) in section 2. In this case, in order to obtain the a priori estimate for the difference of two solutions $(\hat{x}, \hat{y}, \hat{z})$, the right method is not to apply Itô's formula to $|\hat{x}_t|^2$ and $|\hat{y}_t|^2$, but to $\langle G\hat{x}_{t,}\hat{y}_t\rangle$. With this observation we succeed in obtaining the existence and uniqueness theorem under the G-monotone assumptions of A and Φ. Stochastic Hamiltonian systems introduced by Bismut [Bis] (see also Bensoussan [Ben]) can be treated as a special case.

It has been observed by the first author (see [P3]) that, coupled with a forward SDE, i.e., a classical SDE of Itô's type, the backward stochastic differential equation (BSDE) gives a probabilistic interpretation for a large kind of (systems of) second-order quasi-linear partial differential equations. This naturally generalizes the well-known Feynman–Kac formula to nonlinear case (see also [PP1], [P2], [PP2], [P4], [EPQ], [EQ], etc. for further developments in this direction).

Several major progressions have been made in the direction of fully coupled FBSDEs (see [P5], [A], [MPY]) and applications to mathematical finance (see [DE], [DMY]). It now becomes more clear that certain important problems in mathematical economics and mathematical finance, especially in the optimization problem, are formulated to be FBSDEs.

To our knowledge, actually there exist two methods in the study of FBSDEs. The first one is purely probabilistic (see Peng [P5], Antonelli [A]). Their main idea was to apply Itô's formula to $|\hat{x}|^2$ and $|\hat{y}|^2$ and then to construct a contraction mapping, where $(\hat{x}, \hat{y})$ stands for the difference of two solutions (x, y) and (x', y').

This method may be regarded as a sort of direct extension of Picard's iteration (see [A], [P3]). This was shown to be successful in treating stochastic differential equations (SDEs) of Itô's type as well as BSDEs, but the application of these techniques to FBSDEs is limited: one can only obtain the local existence and uniqueness results;

i.e., the time duration $[0, T]$ on which the solutions exist (without explosion) has to be sufficiently small.

The second method concerns a kind of "four-steps scheme" approach (see Ma, Protter, and Yong [MPY] and Duffie, Ma, and Yong [DMY]). This method may be regarded as a sort of combination of the methods of partial differential equation and probability or stochastic optimal control. With this method those authors have successfully obtained the results of existence and uniqueness of FBSDE in an arbitrarily prescribed time duration, but they need the equation to be nondegenerate in the sense that the matrix-valued coefficient σ is nondegenerate. On the other hand, as it is well known, the partial differential equation approach cannot be used to treat the case where the coefficients themselves are randomly disturbed, which is often the case in the practical situation, e.g., in financial markets.

The advantages of our method are as follows: (i) the assumptions are very easy to verify. (ii) Many existing problems of FBSDE in optimal control and Hamiltonian systems satisfy these assumptions. (iii) We do not need to impose the nondegenerate condition on the diffusion term σ; the coefficients are allowed to be randomly perturbed (e.g., $\Phi = \Phi(\omega, x)$, etc.).

This paper improves the recent result of [HP] in the following points: (i) One of typical case; i.e., the existence and uniqueness of a Hamiltonian system related to the cost function satisfying some convex conditions such as the linear quadratic optimal control problem under classical condition is a simple corollary of our result. In [HP], a similar result can also be obtained, but the assumption will become unreasonably strong. (ii) The result in [HP] is limited to the case where x and y in FBSDE take the same dimension; i.e., $m = n$ in our paper. This paper successfully removes this heavy restriction. (iii) Even in the case where $m = n$, our monotone assumptions are clearly weaker and the corresponding conclusions are stronger. An example is given in section 3 to show that some stochastic Hamiltonian systems arising from differential games are also G-monotone. (iv) In section 2, a counterexample is given showing that, without our monotone conditions, the existence result of FBSDE in an arbitrarily large time duration does not hold true.

This paper is organized as follows: In the next section we present our main result about the existence and uniqueness of FBSDE under the monotone conditions. Several examples of FBSDE related to stochastic optimal control and differential games problems are given in section 3. We also obtain the existence and uniqueness of an FBSDE corresponding linear quadratic optimal control problem in this section.

2. FBSDE: Existence and uniqueness. We consider the following FBSDEs:

$$\begin{aligned} dx_t &= b(t, x_t, y_t, z_t)dt + \sigma(t, x_t, y_t, z_t)dB_t, \\ -dy_t &= f(t, x_t, y_t, z_t)dt - z_t dB_t, \\ x_0 &= a, \quad y_T = \Phi(x_T), \end{aligned} \tag{2.1}$$

where

$$\begin{aligned} &b : \Omega \times [0, T] \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{m\times d} \longrightarrow \mathbb{R}^n, \\ &\sigma : \Omega \times [0, T] \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{m\times d} \longrightarrow \mathbb{R}^{n\times d}, \\ &f : \Omega \times [0, T] \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{m\times d} \longrightarrow \mathbb{R}^m, \\ &\Phi : \Omega \times \mathbb{R}^n \longrightarrow \mathbb{R}^m. \end{aligned}$$

We are given an $m \times n$ full-rank matrix G. We use the notations

$$u = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad A(t,u) = \begin{pmatrix} -G^T f \\ Gb \\ G\sigma \end{pmatrix}(t,u),$$

where $G\sigma = (G\sigma_1 \cdots G\sigma_d)$. We use the usual inner product and Euclidean norm in $\mathbb{R}^n$, $\mathbb{R}^m$, and $\mathbb{R}^{m\times d}$.

Definition 2.1. *A triple of process (X,Y,Z): $\Omega \times [0,T] \longrightarrow \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{m\times d}$ is called an adapted solution of FBSDE* (2.1) *if $(X,Y,Z) \in M^2(0,T;\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{m\times d})$ and it satisfies FBSDE* (2.1).

We assume that

$$\text{(H2.1)} \qquad \begin{cases} \text{(i) } A(t,u) \textit{ is uniformly Lipschitz with respect to } u; \\ \text{(ii) } \textit{for each } u,\ A(\cdot,u) \textit{ is in } M^2(0,T); \\ \text{(iii) } \Phi(x) \textit{ is uniformly Lipschitz with respect to } x \in \mathbb{R}^n; \\ \text{(iv) } \textit{for each } x,\ \Phi(x) \textit{ is in } L^2(\Omega,\mathcal{F}_T,\mathbf{P}). \end{cases}$$

The following monotone conditions are our main assumptions:

$$\text{(H2.2)} \qquad \begin{aligned} &\langle A(t,u) - A(t,\overline{u}), u - \overline{u}\rangle \le -\beta_1 |G\widehat{x}|^2 - \beta_2 |G^T\widehat{y}|^2, \\ &\langle \Phi(x) - \Phi(\overline{x}), G(x-\overline{x})\rangle \ge 0 \\ &\forall u = (x,y,z), \quad \overline{u} = (\overline{x},\overline{y},\overline{z}), \quad \widehat{x} = x - \overline{x}, \quad \widehat{y} = y - \overline{y}, \quad \widehat{z} = z - \overline{z}, \end{aligned}$$

where β_1 and β_2 are given nonnegative constants with $\beta_1 + \beta_2 > 0$. Moreover we have $\beta_1 > 0$ (resp., $\beta_2 > 0$) when $m > n$ (resp., $n > m$).

Our first result of this section is the following uniqueness theorem.

Theorem 2.2. *We assume* (H2.1) *and* (H2.2). *Then FBSDE* (2.1) *has at most one adapted solution.*

Proof. Let $u_s = (x_s, y_s, z_s)$ and $u'_s = (x'_s, y'_s, z'_s)$ be two solutions of (2.1). We set $\widehat{u} = (x - x', y - y', z - z') = (\widehat{x},\widehat{y},\widehat{z})$. We use Itô's formula applied to $\langle G\widehat{x}_s, \widehat{y}_s\rangle$:

$$\begin{aligned} \mathbb{E}\langle \Phi(x_T) - \Phi(x'_T), G\widehat{x}_T\rangle - \mathbb{E}\langle \widehat{y}_t, G\widehat{x}_t\rangle &= \mathbb{E}\int_t^T \langle A(s,u_s) - A(s,u'_s), \widehat{u}_s\rangle ds \\ &\le -\beta_1 \mathbb{E}\int_t^T \langle G\widehat{x}_s, G\widehat{x}_s\rangle ds - \beta_2 \mathbb{E}\int_t^T \langle G^T\widehat{y}_s, G^T\widehat{y}_s\rangle ds. \end{aligned}$$

This with the monotone of Φ and A implies

$$\beta_1 \mathbb{E}\int_0^T \langle G\widehat{x}_s, G\widehat{x}_s\rangle ds + \beta_2 \mathbb{E}\int_0^T \langle G^T\widehat{y}_s, G^T\widehat{y}_s\rangle ds \le 0.$$

We first treat the case where $m > n$. In this case $\beta_1 > 0$, then $\langle G\widehat{x}_s, G\widehat{x}_s\rangle \equiv 0$. We have $\widehat{x}_s \equiv 0$. Thus $x_s \equiv x'_s$. In particular, $\Phi(x_T) \equiv \Phi(x'_T)$. Thus, from the uniqueness of BSDE, it follows that $y_s \equiv y'_s$ and $z_s \equiv z'_s$.

We now discuss the second case where $m < n$. In this case $\beta_2 > 0$, then $\langle G^T\widehat{y}_s, G^T\widehat{y}_s\rangle \equiv 0$. We have $y_s \equiv y'_s$. We apply Itô's formula to $|\widehat{y}_s|^2 \equiv 0$. It follows that $\int_0^T |z_s - z'_s|^2 ds = 0$. Thus $z_s \equiv z'_s$. Finally, from the uniqueness of Itô's SDE it follows that $x_s \equiv x'_s$. Similarly to the above two cases, the result can be obtained easily in the case $m = n$. □

We now give an existence result of FBSDE (2.1) for a special case where Φ does not depend on x, i.e., $\Phi(x) \equiv \xi$.

THEOREM 2.3. *We assume* $y_T = \xi$, $\xi \in L^2(\Omega, \mathcal{F}_T, \mathbf{P})$ *and* (H2.1), (H2.2). *Then there exists a unique triple* $u_s = (x_s, y_s, z_s), s \in [0, T]$ *satisfying equations* (2.1).

The uniqueness is an immediate consequence of Theorem 2.2.

The proof of the existence is a combination of the above technique and an a priori estimate technique introduced in [P3]. We first need the following two lemmas. Consider the following family of FBSDEs parametrized by $\alpha \in [0, 1]$:

$$\begin{aligned} dx_t^\alpha &= \left[(1-\alpha)\beta_2(-G^T y_t^\alpha) + \alpha b(t, u_t^\alpha) + \phi_t\right] dt \\ &\quad + \left[(1-\alpha)\beta_2(-G^T z_t^\alpha) + \alpha\sigma(t, u_t^\alpha) + \psi_t\right] dB_t, \\ -dy_t^\alpha &= [(1-\alpha)\beta_1 G x_t^\alpha + \alpha f(t, u_t^\alpha) + \gamma_t]\, dt - z_t^\alpha dB_t, \\ x_0^\alpha &= a, \qquad y_T^\alpha = \xi, \end{aligned} \tag{2.2}$$

where ϕ, ψ, and γ are given processes in $M^2(0, T)$ with values in $\mathbb{R}^n$, $\mathbb{R}^{n\times d}$, and $\mathbb{R}^m$, resp. Clearly, when $\alpha = 1$ the existence of the solution of (2.2) implies that of (2.1) for $y_T \equiv \xi$. The following lemma gives an a priori estimate for the "existence interval" of (2.2) with respect to $\alpha \in [0, 1]$.

LEMMA 2.4. *We assume* (H2.1) *and* (H2.2). *Then there exists a positive constant* δ_0 *such that if, a priori, for an* $\alpha_0 \in [0, 1)$ *there exists a solution* $(x^{\alpha_0}, y^{\alpha_0}, z^{\alpha_0})$ *of* (2.2), *then for each* $\delta \in [0, \delta_0]$ *there exists a solution* $(x^{\alpha_0+\delta}, y^{\alpha_0+\delta}, z^{\alpha_0+\delta})$ *of* (2.2) *for* $\alpha = \alpha_0 + \delta$.

Proof. Since for each $\phi \in M^2(0, T; \mathbb{R}^n)$, $\gamma \in M^2(0, T; \mathbb{R}^m)$, $\psi \in M^2(0, T; \mathbb{R}^{n\times d})$, $\alpha \in [0, 1)$ there exists a (unique) solution of (2.2), thus, for each triple

$$u_s = (x_s, y_s, z_s) \in M^2(0, T; \mathbb{R}^{n+m+m\times d})$$

there exists a unique triple $U_s = (X_s, Y_s, Z_s) \in M^2(0, T; \mathbb{R}^{n+m+m\times d})$ satisfying the following FBSDE:

$$\begin{aligned} dX_t &= \left[(1-\alpha_0)\beta_2(-G^T Y_t) + \alpha_0 b(t, U_t) + \delta(\beta_2 G^T y_t + b(t, u_t)) + \phi_t\right] dt \\ &\quad + \left[(1-\alpha_0)\beta_2(-G^T Z_t) + \alpha_0\sigma(t, U_t) + \delta(\beta_2 G^T z_t + \sigma(t, u_t)) + \psi_t\right] dB_t, \\ -dY_t &= [(1-\alpha_0)\beta_1 G X_t + \alpha_0 f(t, U_t) + \delta(-\beta_1 G x_t + f(t, u_t)) + \gamma_t]\, dt - Z_t dB_t, \\ X_0 &= a, \qquad Y_T = \xi. \end{aligned}$$

We are going to prove that the mapping defined by

$$I_{\alpha_0+\delta}(u) = U : M^2(0, T; \mathbb{R}^{n+m+m\times d}) \to M^2(0, T; \mathbb{R}^{n+m+m\times d})$$

is a contraction.

Let $u' = (x', y', z') \in M^2(0, T; \mathbb{R}^{n+m+m\times d})$, and let $U' = (X', Y', Z') = I_{\alpha_0+\delta}(u')$. We set

$$\widehat{u} = (\widehat{x}, \widehat{y}, \widehat{z}) = (x - x', y - y', z - z'), \quad \widehat{U} = (\widehat{X}, \widehat{Y}, \widehat{Z}) = (X - X', Y - Y', Z - Z').$$

Using Itô's formula applied to $\langle G\widehat{X}_s, \widehat{Y}_s\rangle$ it yields

$$\begin{aligned}0 &= \mathbb{E}\int_0^T \langle \alpha_0(A(s,U_s)-A(s,U'_s)), \widehat{U}_s\rangle ds\\ &- (1-\alpha_0)\mathbb{E}\int_0^T (\beta_1\langle G\widehat{X}_s, G\widehat{X}_s\rangle + \beta_2\langle G^T\widehat{Y}_s, G^T\widehat{Y}_s\rangle + \beta_2\langle G^T\widehat{Z}_s, G^T\widehat{Z}_s\rangle)ds\\ &+ \delta\mathbb{E}\int_0^T (\beta_1\langle G\widehat{X}_s, G\widehat{x}_s\rangle + \beta_2\langle G^T\widehat{Y}_s, G^T\widehat{y}_s\rangle + \beta_2\langle G^T\widehat{Z}_s, G^T\widehat{z}_s\rangle\\ &+ \langle \widehat{X}_s, -G^T\bar{f}_s\rangle + \langle G^T\widehat{Y}_s, \bar{b}_s\rangle + \langle \widehat{Z}_s, G\bar{\sigma}_s\rangle)ds,\end{aligned}$$

where

$$\bar{f}_s = f(s,u_s) - f(s,u'_s), \quad \bar{b}_s = b(s,u_s) - b(s,u'_s), \quad \bar{\sigma}_s = \sigma(s,u_s) - \sigma(s,u'_s).$$

From (H2.1) and (H2.2), we can get

$$\mathbb{E}\int_0^T \left(\beta_1\langle G\widehat{X}_s, G\widehat{X}_s\rangle + \beta_2\langle G^T\widehat{Y}_s, G^T\widehat{Y}_s\rangle\right) ds \le \delta C_1\mathbb{E}\int_0^T \left(|\widehat{u}_s|^2 + |\widehat{U}_s|^2\right) ds.$$

On the other hand, since X and X' are solutions of SDE of Itô's type, applying the usual technique, the estimates for the difference $\widehat{X} = X - X'$ are obtained by

$$\sup_{0\le s\le T} \mathbb{E}|\widehat{X}_s|^2 \le C_1\delta\mathbb{E}\int_0^T |\widehat{u}_s|^2 ds + C_1\mathbb{E}\int_0^T \left(|\widehat{Y}_s|^2 + |\widehat{Z}_s|^2\right) ds,$$

$$\mathbb{E}\int_0^T |\widehat{X}_s|^2 ds \le C_1T\delta\mathbb{E}\int_0^T |\widehat{u}_s|^2 ds + C_1T\mathbb{E}\int_0^T \left(|\widehat{Y}_s|^2 + |\widehat{Z}_s|^2\right) ds.$$

Similarly, for the difference of the solutions $(\widehat{Y}, \widehat{Z}) = (Y - Y', Z - Z')$, we apply the usual technique to the BSDE:

$$\mathbb{E}\int_0^T \left(|\widehat{Y}_s|^2 + |\widehat{Z}_s|^2\right) ds \le C_1\delta\mathbb{E}\int_0^T |\widehat{u}_s|^2 ds + C_1\mathbb{E}\int_0^T |\widehat{X}_s|^2 ds.$$

Here the constant C_1 depends on the Lipschitz constants as well as G, β_1, β_2, and T. Using Itô's formula to $|\widehat{Y}_s|^2$ to the BSDE once more it yields

$$\begin{aligned}&|\widehat{Y}_0|^2 + \mathbb{E}\int_0^T |\widehat{Z}_s|^2 ds\\ &= \mathbb{E}\int_0^T 2\widehat{Y}_s[\alpha_0(f(s,U_s) - f(s,U'_s)) + (1-\alpha_0)\beta_1 G\widehat{X}_s + \delta\bar{f}_s - \delta\beta_1 G\widehat{x}_s]ds\\ &\le \frac{1}{4}\mathbb{E}\int_0^T |\widehat{Z}_s|^2 + \frac{1}{4L}\mathbb{E}\int_0^T |\widehat{X}_s|^2 ds + C_2\mathbb{E}\int_0^T |\widehat{Y}_s|^2 ds + C_2\delta\mathbb{E}\int_0^T |\widehat{u}_s|^2 ds.\end{aligned}$$

Here $L = \max(C_1T, 1)$, C_2 is a sufficiently large constant which depends on L, G, β_1, and the Lipschitz constants.

Combining the above five estimates, it is clear that, whenever $\beta_1 > 0$, $\beta_2 \ge 0$, and/or $\beta_1 \ge 0$, $\beta_2 > 0$ hold true, we always have

$$\mathbb{E}\int_0^T |\widehat{U}_s|^2 ds \le C\delta\mathbb{E}\int_0^T |\widehat{u}_s|^2 ds,$$

where the constant C depends only on C_1, C_2, L, and T. We now choose $\delta_0 = \frac{1}{2C}$. It is clear that, for each fixed $\delta \in [0, \delta_0]$, the mapping $I_{\alpha_0+\delta}$ is a contraction in the sense that

$$\mathbb{E}\int_0^T |\widehat{U}_s|^2 ds \leq \frac{1}{2}\mathbb{E}\int_0^T |\widehat{u}_s|^2 ds.$$

It follows immediately that this mapping has a unique fixed point $U^{\alpha_0+\delta} = (X^{\alpha_0+\delta}, Y^{\alpha_0+\delta}, Z^{\alpha_0+\delta})$ which is the solution of (2.2) for $\alpha = \alpha_0 + \delta$. The proof is complete. □

It remains to prove that, when $\alpha = 0$, (2.2), i.e.,

$$\begin{aligned} dx_t^0 &= \left[-\beta_2 G^T y_t^0 + \phi_t\right] dt + \left[-\beta_2 G^T z_t^0 + \psi_t\right] dB_t, \\ -dy_t^0 &= \left[\beta_1 G x_t^0 + \gamma_t\right] dt - z_t^0 dB_t, \\ x_0^0 &= a, \quad y_T^0 = \xi, \end{aligned}$$

has a unique solution.

We will treat a more general situation which can also be used in the proof of Theorem 2.6.

Lemma 2.5. *The following equation has a unique solution:*

$$\begin{aligned} dx_t &= \left[-\beta_2 G^T y_t + \phi_t\right] dt + \left[-\beta_2 G^T z_t + \psi_t\right] dB_t, \\ -dy_t &= \left[\beta_1 G x_t + \gamma_t\right] dt - z_t dB_t, \\ x_0 &= a, \quad y_T = \lambda G x_T + \xi, \end{aligned} \tag{2.3}$$

where λ is a nonnegative constant.

Proof. We observe that the matrix G is of full rank. The proof of the existence for (2.3) will be divided into two cases: $n \leq m$ and $n > m$.

For the first case, the matrix $G^T G$ is strictly positive. We set

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} x \\ G^T y \\ G^T z \end{pmatrix}, \qquad \begin{pmatrix} y'' \\ z'' \end{pmatrix} = \begin{pmatrix} (I_m - G(G^T G)^{-1} G^T) y \\ (I_m - G(G^T G)^{-1} G^T) z \end{pmatrix}.$$

Multiplying G^T on both sides of the BSDE for (y, z) yields

$$\begin{aligned} dx_t' &= \left[-\beta_2 y_t' + \phi_t\right] dt + \left[-\beta_2 z_t' + \psi_t\right] dB_t, \\ -dy_t' &= \left[\beta_1 G^T G x_t' + G^T \gamma_t\right] dt - z_t' dB_t, \\ x_0' &= a, \quad y_T' = \lambda G^T G x_T' + G^T \xi. \end{aligned} \tag{2.4}$$

Similarly, multiplying $(I_m - G(G^T G)^{-1} G^T)$ on both sides of the same equation yields

$$\begin{aligned} -dy_t'' &= (I_m - G(G^T G)^{-1} G^T)\gamma_t dt - z_t'' dB_t, \\ y_T'' &= (I_m - G(G^T G)^{-1} G^T)\xi. \end{aligned}$$

Obviously the pair (y'', z'') is uniquely determined. The uniqueness of (x', y', z') follows from Theorem 2.2. In order to solve (2.4), we introduce the following $n \times n$-symmetric matrix-valued ordinary differential equation, known as the matrix-Riccati equation:

$$\begin{aligned} -\dot{K}(t) &= -\beta_2 K^2 + \beta_1 G^T G, \quad t \in [0, T], \\ K(T) &= \lambda G^T G. \end{aligned}$$

It is well known that this equation has a unique nonnegative solution $K(\cdot) \in C^1([0,T]; S^n)$, where S^n stands for the space of all $n \times n$-symmetric matrices. We then consider the solution $(p,q) \in M^2(0,T;\mathbb{R}^{n+n\times d})$ of the following linear simple BSDE:

$$\begin{aligned} -dp_t &= [-\beta_2 K(t)p_t + K(t)\phi_t + G^T\gamma_t]dt \\ &\qquad + [K(t)\psi_t - (I_n + \beta_2 K(t))q_t]dB_t, \quad t \in [0,T], \\ p_T &= G^T\xi. \end{aligned}$$

We now let x'_t be the solution of the SDE

$$\begin{aligned} dx'_t &= [-\beta_2(K(t)x'_t + p_t) + \phi_t]\,dt + [\psi_t - \beta_2 q_t]\,dB_t, \quad t \in [0,T], \\ x'_0 &= a. \end{aligned}$$

Then it is easy to check that $(x'_{t,}y'_t, z'_t) = (x'_t, K(t)x'_t + p_t, q_t)$ is the solution of (2.4). Once (x', y', z') and (y'', z'') are resolved, then the triple (x,y,z) is uniquely obtained by

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x' \\ G(G^TG)^{-1}y' + y'' \\ G(G^TG)^{-1}z' + z'' \end{pmatrix}.$$

The proof for the case $n > m$ is analogous to the above opposite case. We observe that in this case the matrix GG^T is of full rank. We set

$$\begin{pmatrix} x' \\ x'' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} Gx \\ (I_n - G^T(GG^T)^{-1}G)x \\ y \\ z \end{pmatrix}.$$

x'' is the unique solution of the following linear forward SDE:

$$\begin{aligned} dx''_t &= (I_n - G^T(GG^T)^{-1}G)\phi_t dt + (I_n - G^T(GG^T)^{-1}G)\psi_t dB_t, \\ x''_0 &= (I_n - G^T(GG^T)^{-1}G)a. \end{aligned}$$

The triple (x', y', z') solves the FBSDE

$$\begin{aligned} dx'_t &= \left[-\beta_2 GG^T y'_t + G\phi_t\right]dt + \left[G\psi_t - \beta_2 GG^T z'_t\right]dB_t, \\ -dy'_t &= [\beta_1 x'_t + \gamma_t]\,dt - z'_t dB_t, \\ x'_0 &= Ga, \quad y'_T = \lambda x'_T + \xi. \end{aligned} \tag{2.5}$$

To solve this equation, we introduce the solution $K(\cdot) \in C([0,T]; S^m)$ of the $m \times m$ symmetric matrix-valued Riccati equation

$$\begin{aligned} -\dot{K}(t) &= \beta_1 I_m - \beta_2 KGG^TK, \quad t \in [0,T], \\ K(T) &= \lambda I_m. \end{aligned}$$

We then consider the unique solution $(p,q) \in M^2(0,T;\mathbb{R}^{m+m\times d})$ of the following linear BSDE:

$$\begin{aligned} -dp_t &= \left(-\beta_2 K(t)GG^Tp_t + K(t)G\phi_t + \gamma_t\right)dt \\ &\qquad + \left(K(t)G\psi_t - (I_m + \beta_2 K(t)GG^T)q_t\right)dB_t, \quad t \in [0,T], \\ p_T &= \xi. \end{aligned}$$

We now let x'_t be the solution of the SDE

$$\begin{aligned} dx'_t &= \left[-\beta_2 GG^T(K(t)x'_t + p_t) + G\phi_t\right] dt \\ &\quad + \left[G\psi_t - \beta_2 GG^T q_t\right] dB_t, \quad t \in [0,T], \\ x'_0 &= Ga. \end{aligned}$$

It is then easy to see that $(x'_t, y'_t, z'_t) = (x'_t, K(t)x'_t + p_t, q_t)$ is the solution of (2.5). Once (x', x'', y', z') are resolved, by the definition, the triple (x, y, z) is uniquely obtained by

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} G^T(GG^T)^{-1}x' + x'' \\ y' \\ z' \end{pmatrix}.$$

The proof is complete. □

We now proceed to give the following.

Proof of the existence of Theorem 2.3. By Lemma 2.5, when $\lambda = 0$ in (2.3), (2.2) for $\alpha_0 = 0$ has a unique solution.

It then follows from Lemma 2.4 that there exists a positive constant $\delta_0 = \delta_0(C_1, C_2, L, T)$ such that for each $\delta \in [0, \delta_0]$, (2.2) for $\alpha = \alpha_0 + \delta$ has a unique solution. Since δ_0 depends only on (C_1, C_2, L, T), we can repeat this process for N-times with $1 \le N\delta_0 < 1 + \delta_0$. It then follows that, in particular, (2.2) for $\alpha = 1$ with $\phi_s \equiv 0$, $\gamma_s \equiv 0$, and $\psi_s \equiv 0$ has a unique solution. The proof is complete. □

Now we can consider the FBSDE (2.1) for $y_T = \Phi(x_T)$. In fact, the assumption (H2.1) has to be strengthened to the following (H2.3):

$$\text{(H2.3)} \qquad \begin{aligned} &\langle A(t,u) - A(t,\overline{u}), u - \overline{u}\rangle \le -\beta_1 |G\widehat{x}|^2 - \beta_2(|G^T\widehat{y}|^2 + |G^T\widehat{z}|^2), \\ &\langle \Phi(x) - \Phi(\overline{x}), G(x - \overline{x})\rangle \ge \mu_1 |G\widehat{x}|^2 \\ &\forall u = (x,y,z), \quad \overline{u} = (\overline{x}, \overline{y}, \overline{z}), \quad \widehat{x} = x - \overline{x}, \quad \widehat{y} = y - \overline{y}, \quad \widehat{z} = z - \overline{z}, \end{aligned}$$

where β_1, β_2, and μ_1 are given nonnegative constants with $\beta_1 + \beta_2 > 0$, $\mu_1 + \beta_2 > 0$. Moreover we have $\beta_1 > 0$, $\mu_1 > 0$ (resp., $\beta_2 > 0$) when $m > n$ (resp., $n > m$).

We have the following main result of this section.

Theorem 2.6. *Let* (H2.1) *and* (H2.3) *hold. Then there exists a unique adapted solution* (X, Y, Z) *of FBSDE* (2.1).

Actually the method to prove the existence is similar to Theorem 2.3. We now consider the following (2.6) for each $\alpha \in [0,1]$:

$$\text{(2.6)} \qquad \begin{aligned} dx^\alpha_t &= \left[(1-\alpha)\beta_2(-G^T y^\alpha_t) + \alpha b(t, u^\alpha_t) + \phi_t\right] dt \\ &\quad + \left[(1-\alpha)\beta_2(-G^T z^\alpha_t) + \alpha\sigma(t, u^\alpha_t) + \psi_t\right] dB_t, \\ -dy^\alpha_t &= \left[(1-\alpha)\beta_1 G x^\alpha_t + \alpha f(t, u^\alpha_t) + \gamma_t\right] dt - z^\alpha_t dB_t, \\ x^\alpha_0 &= a, \quad y^\alpha_T = \alpha\Phi(x^\alpha_T) + (1-\alpha)Gx^\alpha_T + \xi, \end{aligned}$$

where ϕ, ψ, and γ are given processes in $M^2(0,T)$ with values in $\mathbb{R}^n$, $\mathbb{R}^{n\times d}$, and $\mathbb{R}^m$, resp., $\xi \in L^2(\Omega, \mathcal{F}_T, P)$. Clearly the existence of (2.6) for $\alpha = 1$ implies the existence of (2.1).

In order to obtain this conclusion, we also need the following lemma.

Lemma 2.7. *We assume* (H2.1) *and* (H2.3). *Then there exists a positive constant* δ_0 *such that if, a prior, for a* $\alpha_0 \in [0,1)$ *there exists a triple of solution* $(X^{\alpha_0}, Y^{\alpha_0}, Z^{\alpha_0})$

of (2.6)*, then for each* $\delta \in [0,\delta_0]$ *there exists a solution* $(X^{\alpha_0+\delta}, Y^{\alpha_0+\delta}, Z^{\alpha_0+\delta})$ (2.6) *for* $\alpha = \alpha_0 + \delta$.

Proof. Since for each $\phi \in M^2(0,T;\mathbb{R}^n)$, $\gamma \in M^2(0,T;\mathbb{R}^m)$, $\psi \in M^2(0,T;\mathbb{R}^{n\times d})$, $\xi \in L^2(\Omega,\mathcal{F}_T,P)$, $\alpha_0 \in [0,1)$ there exists a (unique) solution of (2.6), thus, for each $x_T \in L^2(\Omega,\mathcal{F}_T,P)$ and a triple $u_s = (x_s,y_s,z_s) \in M^2(0,T;\mathbb{R}^{n+m+m\times d})$ there exists a unique triple $U_s = (X_s,Y_s,Z_s) \in M^2(0,T;\mathbb{R}^{n+m+m\times d})$ satisfying the following FBSDE:

$$\begin{aligned}
dX_t &= \left[(1-\alpha_0)\beta_2(-G^TY_t) + \alpha_0 b(t,U_t) + \delta(\beta_2 G^T y_t + b(t,u_t)) + \phi_t\right] dt\\
&\qquad + \left[(1-\alpha_0)\beta_2(-G^TZ_t) + \alpha_0\sigma(t,U_t) + \delta(\beta_2 G^T z_t + \sigma(t,u_t)) + \psi_t\right] dB_t,\\
-dY_t &= [(1-\alpha_0)\beta_1 GX_t + \alpha_0 f(t,U_t) + \delta(-\beta_1 Gx_t + f(t,u_t)) + \gamma_t]\, dt - Z_t dB_t,\\
X_0 &= a, \quad Y_T = \alpha_0\Phi(X_T) + (1-\alpha_0)GX_T + \delta(\Phi(x_T) - Gx_T) + \xi.
\end{aligned}$$

We now proceed to prove that, if δ is sufficiently small, the mapping defined by

$$\begin{aligned}
&I_{\alpha_0+\delta}(u \times x_T) = U \times X_T :\\
&M^2(0,T;\mathbb{R}^{n+m+m\times d}) \times L^2(\Omega,\mathcal{F}_T,P;\mathbb{R}^n) \to M^2(0,T;\mathbb{R}^{n+m+m\times d}) \times L^2(\Omega,\mathcal{F}_T,P;\mathbb{R}^n)
\end{aligned}$$

is a contraction.

Let $u' = (x',y',z') \in M^2(0,T;\mathbb{R}^{n+m+m\times d})$, and let $U' \times X'_T = I_{\alpha_0+\delta}(u' \times x'_T)$.

We set

$$\widehat{u} = (\widehat{x},\widehat{y},\widehat{z}) = (x-x', y-y', z-z'), \quad \widehat{U} = (\widehat{X},\widehat{Y},\widehat{Z}) = (X-X', Y-Y', Z-Z').$$

Using Itô's formula applied to $\langle G\widehat{X}_s, \widehat{Y}_s\rangle$ yields

$$\begin{aligned}
&\alpha_0\mathbb{E}\langle\Phi(X_T) - \Phi(X'_T), G\widehat{X}_T\rangle + (1-\alpha_0)\mathbb{E}\langle G\widehat{X}_T, G\widehat{X}_T\rangle\\
&\qquad\qquad + \delta\mathbb{E}\langle\Phi(x_T) - \Phi(x'_T) - G\widehat{x}_T, G\widehat{x}_T\rangle\\
&= \mathbb{E}\int_0^T \langle\alpha_0(A(s,U_s) - A(s,U'_s)), \widehat{U}_s\rangle ds\\
&- (1-\alpha_0)\mathbb{E}\int_0^T (\beta_1\langle G\widehat{X}_s, G\widehat{X}_s\rangle + \beta_2\langle G^T\widehat{Y}_s, G^T\widehat{Y}_s\rangle + \beta_2\langle G^T\widehat{Z}_s, G^T\widehat{Z}_s\rangle)ds\\
&+ \delta\mathbb{E}\int_0^T (\beta_1\langle G\widehat{X}_s, G\widehat{x}_s\rangle + \beta_2\langle G^T\widehat{Y}_s, G^T\widehat{y}_s\rangle + \beta_2\langle G^T\widehat{Z}_s, G^T\widehat{z}_s\rangle\\
&\qquad + \langle\widehat{X}_s, -G^T\bar{f}_s\rangle + \langle G^T\widehat{Y}_s, \bar{b}_s\rangle + \langle\widehat{Z}_s, G\bar{\sigma}_s\rangle)ds,
\end{aligned}$$

where

$$\bar{f}_s = f(s,u_s) - f(s,u'_s), \quad \bar{b}_s = b(s,u_s) - b(s,u'_s), \quad \bar{\sigma}_s = \sigma(s,u_s) - \sigma(s,u'_s).$$

From (H2.1) and (H2.3), we can get

$$\begin{aligned}
&(\alpha_0\mu_1 + (1-\alpha_0))\mathbb{E}|G\widehat{X}_T|^2 + \beta_1\mathbb{E}\int_0^T |G\widehat{X}_s|^2 ds + \beta_2\mathbb{E}\int_0^T (|G^T\widehat{Y}_s|^2 + |G^T\widehat{Z}_s|^2)ds\\
&\le \delta K_1\mathbb{E}\int_0^T \left(|\widehat{u}_s|^2 + |\widehat{U}_s|^2\right) ds + \delta K_1\mathbb{E}|\widehat{X}_T|^2 + \delta K_1\mathbb{E}|\widehat{x}_T|^2.
\end{aligned}$$

Applying the same technique as Lemma 2.4, we have the following estimates:

$$\begin{aligned}
&\sup_{0\le s\le T}\mathbb{E}|\widehat{X}_s|^2 \le K_1\delta\mathbb{E}\int_0^T|\widehat{u}_s|^2ds + K_1\mathbb{E}\int_0^T\left(|\widehat{Y}_s|^2+|\widehat{Z}_s|^2\right)ds,\\
&\mathbb{E}\int_0^T|\widehat{X}_s|^2ds \le K_1T\delta\mathbb{E}\int_0^T|\widehat{u}_s|^2ds + K_1T\mathbb{E}\int_0^T\left(|\widehat{Y}_s|^2+|\widehat{Z}_s|^2\right)ds,\\
&\mathbb{E}\int_0^T\left(|\widehat{Y}_s|^2+|\widehat{Z}_s|^2\right)ds \le K_1\delta\mathbb{E}\int_0^T|\widehat{u}_s|^2ds + K_1\delta\mathbb{E}|\widehat{x}_T|^2\\
&\qquad\qquad + K_1\mathbb{E}\int_0^T|\widehat{X}_s|^2ds + K_1\mathbb{E}|\widehat{X}_T|^2.
\end{aligned}$$

Here the constant K_1 depends on the Lipschitz constants G, β_1, β_2, and T. If $\mu_1>0$, then $\alpha_0\mu_1+(1-\alpha_0)\ge\mu$, $\mu=\min(1,\mu_1)>0$.

Combining the above four estimates, it is clear that, whatever $\beta_1>0$, $\mu_1>0$, $\beta_2\ge0$ or $\beta_1\ge0$, $\mu_1\ge0$, $\beta_2>0$, we always have

$$\mathbb{E}\int_0^T|\widehat{U}_s|^2ds+\mathbb{E}|\widehat{X}_T|^2\le K\delta\left(\mathbb{E}\int_0^T|\widehat{u}_s|^2ds+\mathbb{E}|\widehat{x}_T|^2\right).$$

Here the constant K depends only on β_1, β_2, μ, K_1, and T. We now choose $\delta_0=\frac{1}{2K}$. It is clear that, for each fixed $\delta\in[0,\delta_0]$, the mapping $I_{\alpha_0+\delta}$ is a contraction in the sense that

$$\mathbb{E}\int_0^T|\widehat{U}_s|^2ds+\mathbb{E}|\widehat{X}_T|^2\le\frac{1}{2}\left(\mathbb{E}\int_0^T|\widehat{u}_s|^2ds+\mathbb{E}|\widehat{x}_T|^2\right).$$

It follows that this mapping has a unique fixed point $U^{\alpha_0+\delta}=(X^{\alpha_0+\delta},Y^{\alpha_0+\delta},Z^{\alpha_0+\delta})$ which is the solution of (2.6) for $\alpha=\alpha_0+\delta$. The proof is complete. □

We now give the proof of Theorem 2.6.

Proof of Theorem 2.6. The uniqueness is obvious from Theorem 2.2.

By Lemma 2.5, when $\lambda=1$, $\xi=0$ in (2.3), (2.6) for $\alpha_0=0$ has a unique solution. It then follows from Lemma 2.7 that there exists a positive constant δ_0 depending on Lipschitz constants, β_1,β_2,μ_1, and T such that, for each $\delta\in[0,\delta_0]$, (2.6) for $\alpha=\alpha_0+\delta$ has a unique solution. We can repeat this process for N-times with $1\le N\delta_0<1+\delta_0$. It then follows that, in particular, FBSDE (2.6) for $\alpha=1$ with $\xi\equiv0$ has a unique solution. The proof is complete. □

REMARK 2.8. *If $m=n$, $G=I_n$, as a special case of Theorem* 2.6, *we improve the main result in* [HP]. *Actually, our monotonicity assumptions are clearly weaker than similar assumptions in* [HP]. *Moreover, if* (2.1) *satisfy the following monotone conditions and* (H2.1),

$$\begin{aligned}
\langle A(t,u)-A(t,\bar{u}),u-\bar{u}\rangle &\le -\beta_2|\widehat{y}|^2,\\
\langle\Phi(x)-\Phi(\bar{x}),\widehat{x}\rangle &\ge \mu_1|\widehat{x}|^2,
\end{aligned}$$

where $m=n$, $G=I_n$, $\beta_2>0$, $\mu_1>0$, we can also obtain the same result as Theorem 2.6. *The proof method is similar.*

At last, in this section, we consider a simple case where the monotone condition (H2.3) is not satisfied.

Example 2.1. In FBSDE (2.1), we set

$$b = z, \quad \sigma = y, \quad f = 0, \quad \Phi = x.$$

In a given time interval $s \in [t, T]$, FBSDE takes the form

$$\begin{aligned} dx_s =& z_s ds + y_s dB_s, \\ -dy_s =& - z_s dB_s, \\ x_t =& x, \quad y_T = x_T. \end{aligned}$$

It is clear that the monotone assumption of (H2.3) fails. Indeed, though the monotone for Φ still holds true,

$$\langle \Phi(x) - \Phi(\overline{x}), x - \overline{x} \rangle = |x - \overline{x}|^2 \geq 0,$$

but for A we have

$$\begin{aligned} \langle A(t,u) - A(t,\overline{u}), u - \overline{u} \rangle &= 2(y - \overline{y})(z - \overline{z}) \\ \forall u = (x,y,z), &\quad \overline{u} = (\overline{x}, \overline{y}, \overline{z}). \end{aligned}$$

According to the local existence and uniqueness theorem (see Peng [P5] and Antonelli [A]), if the length of the interval $[t, T]$ is sufficiently small, then the FBSDE has a unique solution. As a matter of fact, this solution can be written explicitly by

$$(x_s^t, y_s^t, z_s^t) = (x_s^t, p(s)x_s^t, p^2(s)x_s^t), \quad s \in [t, T],$$

with

$$p(s) = \frac{1}{\sqrt{1 - 2(T - s)}}$$

and

$$\begin{aligned} dx_s^t &= p^2(s)x_s^t ds + p(s)x_s^t dB_s, \quad s \in [t, T], \\ x_t^t &= x. \end{aligned}$$

In particular, $y_t^t = p(t)x$, $z_t^t = p^2(t)x$, but since $p(t) \uparrow \infty$ as $t \downarrow (T - \frac{1}{2})$, (y_s^t, z_s^t) will explode when $t \downarrow (T - \frac{1}{2})$.

3. Applications of FBSDE to stochastic optimal controls and differential games.

Example 3.1. *3.1 (stochastic Hamilton system).* (stochastic Hamilton system). Let $n = m$, and let $h(x) : \mathbb{R}^n \to \mathbb{R}$. For notational simplification, we assume that $d = 1$. We consider the following "generalized Hamiltonian":

$$H(x,y,z) : \mathbb{R}^{n+n+n} \longrightarrow \mathbb{R}.$$

We assume that H and h are twice continuously differentiable such that the second derivatives are bounded. We consider the following FBSDE:

$$\begin{aligned} dx_t &= H_y(x_t, y_t, z_t)dt + H_z(x_t, y_t, z_t)dB_t, \\ -dy_t &= H_x(x_t, y_t, z_t)dt - z_t dB_t, \\ x_0 &= a, \quad y_T = h_x(x_T), \end{aligned} \tag{3.1}$$

where $(H_x \quad H_y \quad H_z)$ (resp., h_x) stands for the gradient of H (resp., h). It is clear that the stochastic Hamiltonian equations are a special case of FBSDE (2.1) with

$$f = H_x, \quad b = H_y, \quad \sigma = H_z.$$

Thus the monotone conditions, i.e., (H2.3), correspond, for this special case, to the nonnegativity of the matrix h_{xx} and

$$\begin{pmatrix} -H_{xx} & -H_{xy} & -H_{xz} \\ H_{yx} & H_{yy} & H_{yz} \\ H_{zx} & H_{zy} & H_{zz} \end{pmatrix} (x, y, z) \leq - \begin{pmatrix} \beta_1 I_n & 0 & 0 \\ 0 & \beta_2 I_n & 0 \\ 0 & 0 & \beta_2 I_n \end{pmatrix},$$

where $\beta_1 \geq 0$, $\beta_2 > 0$.

A classical optimization problem of a stochastic control system is formulated as follows: the control system is

$$\begin{aligned} dx_t &= g(x_t, v_t)dt + \mu(x_t, v_t)dB_t, \\ x_0 &= a, \end{aligned}$$

where v_s, $s \in [0, T]$, is an admissible control process, i.e., an $\mathcal{F}_t$-adapted square-integrable process taking values in a given subset U of $\mathbb{R}^k$. Here g and μ are given functions defined on $\mathbb{R}^n \times \mathbb{R}^k$ with values in $\mathbb{R}^n$. The optimal control problem is to minimize the cost function

$$J(v(\cdot)) = \mathbb{E}\left\{ \int_0^T L(x_t, v_t)dt + h(x_T) \right\}$$

over the set of admissible controls. One of the two important methods of the resolution of this optimization problem is the so-called stochastic maximum principle which may be considered as a natural generalization of the well-known Pontryagin's maximum principle to situations with uncertainty. This principle tells us that, under certain reasonable assumptions imposed on the coefficients g, μ, L, h, if $u(\cdot)$ is an admissible optimal control, and if the trajectory corresponding to $u(\cdot)$ is $x(\cdot)$, then, necessarily, there exists a pair of square-integrable adapted processes (y, z) taking values in $\mathbb{R}^n \times \mathbb{R}^n$ such that the triple of processes $(x(\cdot), y(\cdot), z(\cdot))$ satisfies the stochastic Hamiltonian system (3.1), where the Hamiltonian H is defined by

$$H(x, y, z) = \inf_{v \in U} \left\{ \langle y, g(x, v) \rangle + \langle z, \mu(x, v) \rangle + L(x, v) \right\}.$$

A coproduct of this stochastic maximum principle is that it gives us an example of the existence of the solution to FBSDE, but one of the inconveniences of such methods is that checking the conditions of the existence of an optimal control is a very hard problem. Another inconvenience is that this approach does not provide us any information about the uniqueness of the solution of (3.1).

Example 3.2. We now consider a more typical situation. We assume that in the above example g and μ are linear functions

$$g(x, v) = Ax + Bv, \quad \mu(x, v) = Cx + Dv,$$

where A and C are $n \times n$ matrices, B and D are $n \times k$ matrices. We also assume that there is no constraint imposed on the control processes $U = \mathbb{R}^k$. $L(x, v)$, $h(x)$ are twice continuously differentiable with respect to (x, v) and x; $L_x(x, v)$ are bounded by $C(1 + |x| + |v|)$.

In this case, if we assume that

(H3.1)
$$\begin{cases} \text{(i) } k \times k \text{ matrix } L_{vv}(x, v) \geq \beta I_k,\ \beta > 0, \\ \text{(ii) } n \times n \text{ matrix } L_{xx}(x, v) - L_{xv}(x, v) L_{vv}^{-1} L_{vx}(x, v) \text{ is nonnegative,} \\ \text{(iii) } n \times n \text{ matrix } h_{xx}(x) \text{ is nonnegative,} \end{cases}$$

and, if $u(\cdot)$ is an admissible optimal control, the trajectory corresponding to $u(\cdot)$ is $x(\cdot)$, then the Hamiltonian H is defined by

$$H(x, y, z) = \inf_{v \in U} \left\{ \langle y, Ax + Bv \rangle + \langle z, Cx + Dv \rangle + L(x, v) \right\}.$$

We have, necessarily,

$$L_v(x_t, u_t) + B^T y_t + D^T z_t = 0 \tag{3.2}$$

and (x, y, z) satisfy the following FBSDE:

$$\begin{aligned} dx_t &= (Ax_t + Bu_t)dt + (Cx_t + Du_t)dB_t, \\ -dy_t &= (A^T y_t + C^T z_t + L_x(x_t, u_t))dt - z_t dB_t, \\ x_0 &= a, \quad y_T = h_x(x_T). \end{aligned} \tag{3.3}$$

If the optimal control process $u(\cdot)$ does not exist, the solution of (3.3) does not exist. In optimal control theory, we can prove the $u(\cdot)$ exists and is unique. Then there exist solutions to FBSDE (3.3). This is an indirect method and we cannot know whether it is unique. Now we study FBSDE (3.3) directly. It is clear that without Assumption (H2.3), Theorem 2.6 cannot be applied to obtain the existence and uniqueness of (3.3) . Fortunately, for the special case like the following form, we have a better result.

We consider the following kind of FBSDE:

$$\begin{aligned} dx_t &= b(t, x_t, By_t, Cz_t)dt + \sigma(t, x_t, By_t, Cz_t)dB_t, \\ -dy_t &= f(t, x_t, y_t, z_t)dt - z_t dB_t, \\ x_0 &= a, \quad y_T = \Phi(x_T). \end{aligned} \tag{3.4}$$

Here B is a $k \times n$ matrix, C a is $k \times n$ matrix, $(x, y, z) \in \mathbb{R}^{n+n+n}$, and b, f, σ have appropriate dimensions. We use the notations

$$u = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad A(t, u) = \begin{pmatrix} -f \\ b \\ \sigma \end{pmatrix} (t, u)$$

and impose the following monotone conditions:

$$\begin{aligned} &\langle A(t, u) - A(t, \overline{u}), u - \overline{u} \rangle \leq -\nu_1 |\widehat{x}|^2 - \nu_2 |B\widehat{y} + C\widehat{z}|^2, \\ &\langle \Phi(x) - \Phi(\overline{x}), (x - \overline{x}) \rangle \geq 0 \\ &\forall \widehat{u} = (u - \overline{u}) = (\widehat{x}, \widehat{y}, \widehat{z}) = (x - \overline{x}, y - \overline{y}, z - \overline{z}). \end{aligned} \tag{H3.2}$$

Here $\nu_1 \geq 0$, $\nu_2 > 0$. We also assume that

(H3.3)

$$\begin{cases} \text{(i) } A(t,u) \text{ is uniformly Lipschitz with respect to } u; \\ \text{(ii) for each } u,\ A(\cdot,u) \text{ is in } M^2(0,T); \\ \text{(iii) } \Phi(x) \text{ is uniformly Lipschitz with respect to } x \in \mathbb{R}^n; \\ \text{(iv) for each } x,\ \Phi(x) \text{ is in } L^2(\Omega,\mathcal{F}_T,\mathbf{P}); \\ \text{(v) } \forall x, |l(t,x,By,Cz) - l(t,x,B\overline{y},C\overline{z})| \leq K(|B\widehat{y} + C\widehat{z}|), K > 0, l = b, \sigma. \end{cases}$$

Then we have the following result.

Theorem 3.1. *Let* (H3.2) *and* (H3.3) *hold. Then there exists a unique solution* $u_s = (x_s, y_s, z_s)$ *satisfying FBSDE* (3.4).

Proof. We first prove the uniqueness for FBSDE (3.4). Let $u_s = (x_s, y_s, z_s)$ and $u'_s = (x'_s, y'_s, z'_s)$ be two solutions of (3.4). We set $\widehat{u} = (x - x', y - y', z - z') = (\widehat{x}, \widehat{y}, \widehat{z})$. We use Itô's formula applied to $\langle \widehat{x}_s, \widehat{y}_s \rangle$:

$$\mathbb{E}\langle \Phi(x_T) - \Phi(x'_T), \widehat{x}_T \rangle - \mathbb{E}\langle \widehat{y}_t, \widehat{x}_t \rangle = \mathbb{E} \int_t^T \langle A(s,u_s) - A(s,u'_s), \widehat{u}_s \rangle ds.$$

This with the monotonicity of Φ and A implies

$$\nu_2 \mathbb{E} \int_0^T |B\widehat{y}_s + C\widehat{z}_s|^2 ds \leq 0.$$

$B\widehat{y}_s + C\widehat{z}_s \equiv 0$. Following (H3.3)(v) and a classical result for the uniqueness of Itô's FSDE, we have $\widehat{x}_s \equiv 0$. Thus $x_s \equiv x'_s$. In particular, $\Phi(x_T) \equiv \Phi(x'_T)$. Thus from the uniqueness of BSDE it follows that $y_s \equiv y'_s$ and $z_s \equiv z'_s$.

The method to prove the existence is similar to Theorems 2.3 and 2.6. We consider the following FBSDE:

$$\begin{aligned} dx_t^\alpha &= \left[(1-\alpha)(-B^T B y_t^\alpha - B^T C z_t^\alpha) + \alpha b(t, x_t^\alpha, B y_t^\alpha, C z_t^\alpha) + \phi_t\right] dt \\ &\quad + \left[\alpha \sigma(t, x_t^\alpha, B y_t^\alpha, C z_t^\alpha) + (1-\alpha)(-C^T C z_t^\alpha - C^T B y_t^\alpha) + \psi_t\right] dB_t, \\ -dy_t^\alpha &= \left[(1-\alpha) x_t^\alpha + \alpha f(t, x_t^\alpha, y_t^\alpha, z_t^\alpha) + \gamma_t\right] dt - z_t^\alpha dB_t, \\ x_0^\alpha &= a, \quad y_T^\alpha = \alpha \Phi(x_T^\alpha) + \xi, \end{aligned} \tag{3.5}$$

where ϕ, ψ, and γ are given processes in $M^2(0,T)$ with values in $\mathbb{R}^n$, $\xi \in L^2(\Omega, \mathcal{F}_T, P; \mathbb{R}^n)$. Clearly the existence of (3.5) for $\alpha = 1$ implies our conclusion. If $\alpha = 0$, we can obtain the existence and uniqueness result using Theorem 2.3 directly. So we only need to consider

$$\begin{aligned} dX_t &= [(1-\alpha_0)(-B^T B Y_t - B^T C Z_t) + \alpha_0 b(t, X_t, B Y_t, C Z_t)] dt \\ &\quad + [\delta(B^T B y_t + B^T C z_t + b(t, x_t, B y_t, C z_t)) + \phi_t] dt \\ &\quad + \left[\alpha_0 \sigma(t, X_t, B Y_t, C Z_t) + (1-\alpha_0)(-C^T C Z_t - C^T B Y_t)\right] dB_t \\ &\quad + \left[\delta(\sigma(t, x_t, B y_t, C z_t) + C^T C z_t + C^T B y_t) + \psi_t\right] dB_t, \\ -dY_t &= \left[(1-\alpha_0) X_t + \alpha_0 f(t, U_t) - \delta(x_t - f(t, u_t)) + \gamma_t\right] dt - Z_t dB_t, \\ X_0 &= a, \quad Y_T = \alpha_0 \Phi(X_T) + \delta \Phi(x_T) + \xi \end{aligned}$$

and prove that the mapping defined by

$$I_{\alpha_0+\delta}(u \times x_T) = U \times X_T :$$
$$M^2(0,T;\mathbb{R}^{n+n+n\times d}) \times L^2(\Omega,\mathcal{F}_T,P;\mathbb{R}^n) \to M^2(0,T;\mathbb{R}^{n+n+n\times d}) \times L^2(\Omega,\mathcal{F}_T,P;\mathbb{R}^n)$$

is a contraction.

For the difference $(\widehat{X},\widehat{Y}) = (X - X', Y - Y')$, using a similar technique as in Lemmas 2.4 and 2.7 to $\langle \widehat{X},\widehat{Y}\rangle$, we have

$$\begin{aligned}&\nu_2\alpha_0\mathbb{E}\int_0^T \left(|B\widehat{Y}_s + C\widehat{Z}_s|^2\right) ds + (1-\alpha_0)\mathbb{E}\int_0^T \left(|B\widehat{Y}_s + C\widehat{Z}_s|^2\right) ds \\ &\leq \delta C_1\mathbb{E}\int_0^T |\widehat{u}_s|^2 ds + \delta C_1\mathbb{E}|\widehat{x}_T|^2 + \delta C_1\mathbb{E}\int_0^T |\widehat{U}_s|^2 ds + \delta C_1 \mathbb{E}|\widehat{X}_T|^2,\end{aligned}$$

where $\widehat{U}$ and $\widehat{u}$ are defined similarly as in Lemma 2.4.

Here C_1 depends on the Lipschitz constants of b, σ, f, and Φ, $\nu_2\alpha_0+(1-\alpha_0) \geq L_1$, $L_1 = \min(1,\nu_2) > 0$, so

$$\begin{aligned}&\mathbb{E}\int_0^T \left(|B\widehat{Y}_s + C\widehat{Z}_s|^2\right) ds \\ &\leq \delta C_2\mathbb{E}\int_0^T |\widehat{u}_s|^2 ds + \delta C_2\mathbb{E}|\widehat{x}_T|^2 + \delta C_2\mathbb{E}\int_0^T |\widehat{U}_s|^2 ds + \delta C_2\mathbb{E}|\widehat{X}_T|^2.\end{aligned}$$

Here $C_2 = \frac{C_1}{L_1}$. Applying the usual technique to the $\widehat{X}_t = X_t - X'_t$, we obtain

$$\sup_{0\leq s\leq T} \mathbb{E}|\widehat{X}_s|^2 \leq C_3\delta\mathbb{E}\int_0^T |\widehat{u}_s|^2 ds + C_3\mathbb{E}\int_0^T \left(|B\widehat{Y}_s + C\widehat{Z}_s|^2\right) ds,$$

$$\mathbb{E}\int_0^T |\widehat{X}_s|^2 ds \leq C_3T\delta\mathbb{E}\int_0^T |\widehat{u}_s|^2 ds + C_3T\mathbb{E}\int_0^T \left(|B\widehat{Y}_s + C\widehat{Z}_s|^2\right) ds.$$

Similarly, for the difference of the solutions $(\widehat{Y},\widehat{Z}) = (Y - Y', Z - Z')$, applying the usual technique to the BSDE, we have

$$\mathbb{E}\int_0^T \left(|\widehat{Y}_s|^2 + |\widehat{Z}_s|^2\right) ds \leq C_3\delta\mathbb{E}\int_0^T |\widehat{u}_s|^2 ds + C_3\mathbb{E}\int_0^T |\widehat{X}_s|^2 ds + C_3\mathbb{E}|\widehat{X}_T|^2.$$

Here the constant C_3 depends on the Lipschitz constants, K, B, C, and T. Combining the above four estimates, it is clear that we always have

$$\mathbb{E}\int_0^T |\widehat{U}_s|^2 ds + \mathbb{E}|\widehat{X}_T|^2 \leq L\delta\left(\mathbb{E}\int_0^T |\widehat{u}_s|^2 ds + \mathbb{E}|\widehat{x}_T|^2\right),$$

where the constant L depends on C_1, C_2, C_3, and T. So we can choose $\delta_0 = \frac{1}{2L}$. Then for each $\delta \in [0,\delta_0]$, the mapping $I_{\alpha_0+\delta}$ is a contraction.

The other part is the same as in Theorem 2.3 or 2.6; we omit it. □

Now we can consider the FBSDE (3.3) again. Noticing (3.2), we can get

$$L_{vv}(x_t,u_t)u_y = -B^T, L_{vv}(x_t,u_t)u_z = -D^T, L_{vx}(x_t,u_t) + L_{vv}(x_t,u_t)u_x = 0.$$

Combining (H3.1), we can easily verify (3.3), satisfying the assumptions (H3.2) and (H3.3). Then applying Theorem 3.1 there exists a unique solution.

Example 3.3. *We give a more special situation of Example* 3.2: *a linear Hamiltonian system related to a classical linear quadratic optimal control problem.*

We also assume that in Example 3.1 g and μ are linear functions:

$$g(x, v) = Ax + Bv, \quad \mu(x, v) = Cx + Dv,$$

and L is a quadratic function of (x, v); h is a nonnegative quadratic function of x:

$$L(x, v) = \frac{1}{2}\langle Rx, x\rangle + \frac{1}{2}\langle Nv, v\rangle, \quad h = \frac{1}{2}\langle Qx, x\rangle,$$

where Q and R are $n \times n$ nonnegative symmetric matrices, and N is a $k \times k$ positive matrix. The control domain is $U = R^k$; in this case the Hamiltonian H has an explicit form

$$H(x, y, z) = \frac{1}{2}\langle Rx, x\rangle + \langle y, Ax\rangle + \langle z, Cx\rangle - \frac{1}{2}(y^T, z^T)\begin{pmatrix} BN^{-1}B^T & BN^{-1}D^T \\ DN^{-1}B^T & DN^{-1}D^T \end{pmatrix}\begin{pmatrix} y \\ z \end{pmatrix}.$$

Thus the Hessian of H is

$$\begin{pmatrix} -H_{xx} & -H_{xy} & -H_{xz} \\ H_{yx} & H_{yy} & H_{yz} \\ H_{zx} & H_{zy} & H_{zz} \end{pmatrix} = \begin{pmatrix} -R & -A^T & -C^T \\ A & -BN^{-1}B^T & -BN^{-1}D^T \\ C & -DN^{-1}B^T & -DN^{-1}D^T \end{pmatrix}.$$

The corresponding Hamiltonian system is

$$\begin{aligned} dx_t &= (Ax_t - BN^{-1}B^T y_t - BN^{-1}D^T z_t)dt \\ &\qquad + (Cx_t - DN^{-1}B^T y_t - DN^{-1}D^T z_t)dB_t, \\ -dy_t &= (A^T y_t + C^T z_t + Rx_t)dt - z_t dB_t, \\ x_0 &= a, \qquad y_T = Qx_T. \end{aligned} \tag{3.6}$$

It is easy to verify (3.6), satisfying assumptions (H3.2) and (H3.3). Then according to Theorem 3.1 there exists a unique solution.

The above three examples are connected with stochastic optimal control problems. Now we give an example of FBSDE under stochastic differential games circumstances. We can see one of the advantages of introducing the adjustment matrix G is that some stochastic Hamiltonian systems arising from differential games, which are in general far more difficult than stochastic optimal control, can also be dealt with using this method.

Example 3.4. *We consider the following differential game:*

$$\begin{aligned} dx_t &= g(x_t, v_t^1, v_t^2)dt + \mu(x_t, v_t^1, v_t^2)dB_t, \\ x_0 &= a, \end{aligned}$$

with

$$\inf_{v^1}\sup_{v^2}\mathbb{E}\left\{\int_0^T L(x_t, v_t^1, v_t^2)dt + h(x_T)\right\},$$

$v = (v^1, v^2) \in \mathbb{R}^k$.

We only consider a very simple case:

$$\begin{aligned} &g(x,v)=Ax+Bv, &&\mu(x,v)=Cx+Dv,\\ &Bv=B_1v^1+B_2v^2, &&Dv=D_1v^1+D_2v^2,\\ &L=\frac{1}{2}(\langle Rx,x\rangle+\langle Nv,v\rangle), &&h=\frac{1}{2}\langle Qx,x\rangle. \end{aligned}$$

In this case,

$$\begin{aligned} H(x,y,z)&=\sup_{v^2}\inf_{v^1}\left\{\frac{1}{2}(\langle Rx,x\rangle+\langle Nv,v\rangle)+\langle y,Ax+Bv\rangle+\langle z,Cx+Dv\rangle\right\}\\ &=\inf_{v^1}\sup_{v^2}\left\{\frac{1}{2}(\langle Rx,x\rangle+\langle Nv,v\rangle)+\langle y,Ax+Bv\rangle+\langle z,Cx+Dv\rangle\right\}\\ &=\frac{1}{2}(\langle Rx,x\rangle-\langle N^{-1}(B^Ty+D^Tz),B^Ty+D^Tz\rangle)+\langle y,Ax\rangle+\langle z,Cx\rangle. \end{aligned}$$

The related stochastic Hamiltonian system is

$$\begin{aligned} dx_t&=H_ydt+H_zdB_t,\\ -dy_t&=H_xdt-z_tdB_t,\\ x_0&=a,\quad y_T=Qx_T, \end{aligned}$$

or

$$\begin{cases} dx_t=(Ax_t-BN^{-1}B^Ty_t-BN^{-1}D^Tz_t)dt\\ \qquad\quad+(Cx_t-DN^{-1}B^Ty_t-DN^{-1}D^Tz_t)dB_t\\ -dy_t=(A^Ty_t+C^Tz_t+Rx_t)dt-z_tdB_t\\ x_0=a,\qquad y_T=Qx_T. \end{cases}$$

If we set

$$N=\begin{bmatrix}1&0\\0&-1\end{bmatrix},\quad R=\begin{bmatrix}1&0\\0&-1\end{bmatrix},\quad B=I_2,\quad D=I_2,\quad Q=\begin{bmatrix}1&0\\0&-1\end{bmatrix},$$

then the above equation becomes

$$\begin{aligned} dx_t&=(Ax_t-N^{-1}y_t-N^{-1}z_t)dt+(Cx_t-N^{-1}y_t-N^{-1}z_t)dB_t,\\ -dy_t&=(A^Ty_t+C^Tz_t+Rx_t)dt-z_tdB_t,\\ x_0&=a,\qquad y_T=Qx_T. \end{aligned}$$

Here N and R are indefinite. We cannot deal with this system using a traditional method. This is the same dimensional FBSDE, but if we set $G=I_2$, we cannot get the desired result. Now we set

$$G=\begin{bmatrix}1&0\\0&-1\end{bmatrix}.$$

It is easy to check that

$$\begin{bmatrix}-G^T(Rx+A^Ty+C^Tz)\\ G(Ax-N^{-1}y-N^{-1}z)\\ G(Cx-N^{-1}y-N^{-1}z)\end{bmatrix}\cdot\begin{pmatrix}x\\y\\z\end{pmatrix}=-|x|^2-|y+z|^2.$$

From Theorem 3.1, this Hamiltonian system has a unique solution.

In the example, even A and C are randomly disturbed, and the Hamiltonian system cannot be dealt with using differential games or an optimal control method. We can also adjust matrix G to get the desired result.

REFERENCES

[A] F. ANTONELLI, *Backward-forward stochastic differential equations*, Ann. Appl. Probab., 3 (1993), pp. 777–793.

[Ben] A. BENSOUSSAN, *Stochastic maximum principle for distributed parameter system*, J. Franklin Inst., 315 (1983), pp. 387–406.

[Bis] J.-M. BISMUT, *An introductory approach to duality in optimal stochastic control*, SIAM Rev., 20 (1978), pp. 62–78.

[DE] D. DUFFIE AND L. EPSTEIN, *Asset pricing with stochastic differential utilities,* Rev. Financial Stud, 5 (1992), pp. 411–436.

[DMY] D. DUFFIE, J. MA, AND J. YONG, *Black's consol rate conjecture,* Ann. Appl. Probab., 5 (1995), pp. 356–382.

[EPQ] N. EL KAROUI, S. PENG, AND M. C. QUENEZ, *Backward Stochastic Differential Equation in Finance*, Prepublication No. 260 du Lab. de Probabilite, 1994.

[EQ] N. EL KAROUI AND M.-C. QUENEZ, *Dynamic programming and pricing of contingent claims in an incomplete market*, SIAM J. Control Optim., 33 (1995), pp. 29–66.

[Hau] U. G. HAUSSMANN, *General necessary conditions for optimal control of stochastic systems*, Math. Programming Stud., 6 (1976), pp. 34–48.

[HP] Y. HU AND S. PENG, *Solution of forward-backward stochastic differential equations*, Probab. Theory Related Fields, 103 (1995), pp. 273–283.

[Ka] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.

[Ku] H. J. KUSHNER, *Necessary conditions for continuous parameter stochastic optimization problems*, SIAM J. Control, 10 (1972), pp. 550–565.

[MPY] J. MA, P. PROTTER, AND J. YONG, *Solving forward-backward stochastic differential equations explicitly—a four step scheme,* Probab. Theory Related Fields, 98 (1994), pp. 339–359.

[Pon] L. S. PONTRYAGIN, B. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[PP1] E. PARDOUX AND S. PENG, *Adapted solution of a backward stochastic differential equation*, Systems Control Lett., 14 (1990), pp. 55–61.

[PP2] E. PARDOUX AND S. PENG, *Backward Stochastic Differential Equations and Quasilinear Parabolic Partial Differential Equations*, Lecture Notes in Control and Inform. Sci. 176, Springer, New York, 1992, pp. 200–217.

[P1] S. PENG, *A general stochastic maximum principle for optimal control problems*, SIAM J. Control Optim., 28 (1990), pp. 966–979.

[P2] S. PENG, *A generalized dynamic programming principle and Hamilton-Jacobi-Bellmen equation*, Stochastics, 38 (1992), pp. 119–134.

[P3] S. PENG, *Probabilistic interpretation for systems of quasilinear parabolic partial differential equations*, Stochastics, 37 (1991), pp. 61–74.

[P4] S. PENG, *A nonlinear Feynman–Kac formula and applications*, in Proc. Symposium Control Theory; Analysis and Applications, S. Chen and J. Yong, eds., World Scientific, River Edger, NJ, 1991, pp. 173–184.

[P5] S. PENG, *Adapted Solution of Backward Stochastic Equations and Related Partial Differential Equations*, prepublication, Department of Mathematics, Shandong University, Jinan, China, 1992.

SIAM J. Control Optim.
Vol. 37, No. 3, pp. 844–868

ANALYTIC CONTROLLABILITY OF A LINEAR HYBRID SYSTEM*

BRICE ALLIBERT†

Abstract. In this paper, we consider a boundary control problem for a model of a fluid-structure hybrid system. This model has been introduced by Micu and Zuazua in connection with the works of Banks et al. They have given explicit values for the spectral data in [S. Micu and E. Zuazua, *SIAM J. Math. Anal.*, 29 (1998), pp. 967–1001] and results for the control problem in [S. Micu and E. Zuazua, *SIAM J. Control. Optim.,* 35 (1997), pp. 1614–1637]. In the latter paper, they use variable separation and Ingham inequalities to prove an observation estimate that implies, through the Hilbert uniqueness method, that initial data can be controlled within finite time. This paper improves these results by using a stronger form of Ingham inequality for low frequencies. Indeed, we prove that any analytic function is controlled in a finite time whose dependence on the analyticity can be sharply estimated.

Key words. wave equation, analyticity, boundary control, nongeometric control, reachable set

PII. S0363012997320626

1. Introduction.

1.1. Setting of the problem.

Micu and Zuazua have recently introduced in [MZ1] a PDE system as a model of a fluid-structure elastic interaction. This model follows from a series of works by Banks et al. [BFSS].

Let us denote by Ω the square $[0,1]^2$, Γ_0 the subset $\{0\} \times [0,1]$ of its boundary, and $\Gamma_1 = \partial M \setminus \Gamma_0$ the rest of it. The equations of the model are the following:

$$(1) \qquad \begin{cases} \Delta\Phi - \Phi_{tt} = 0 & \text{over } \Omega \times (0,T), \\ \frac{\partial\Phi}{\partial n} = 0 & \text{over } \Gamma_1 \times (0,T), \\ \frac{\partial\Phi}{\partial n} = W_t & \text{over } \Gamma_0 \times (0,T), \\ W_{tt} - W_{yy} + \Phi_t = \beta & \text{over } \Gamma_0 \times (0,T), \\ W_y(0,t) = W_y(1,t) = 0 & \text{over } (0,T), \\ (\Phi, \partial_t\Phi, W, \partial_t W)|_{t=0} = (\Phi_0, \Phi_1, W_0, W_1), \\ \text{with } (\Phi_0, \Phi_1, W_0, W_1) \in L^2(\Omega) \oplus H^{-1}(\Omega) \oplus L^2(\Gamma_0) \oplus H^{-1}(\Gamma_0). \end{cases}$$

For this problem, two values remain constant in time:

$$c_1 = \int_{\Gamma_0} (W_t(y) + \Phi(0,y))\, dy$$

and

$$c_2 = \int_{\Gamma_0} W(y)\, dy + \int_\Omega \Phi_t(x,y)\, dx\, dy.$$

The control problem in time T is to find a function β such that the solution of (1) reaches an equilibrium at time T, e.g.,

$$(\Phi, \partial_t\Phi, W, \partial_t W)|_{t=T} = (c_1, 0, c_2, 0).$$

*Received by the editors April 30, 1997; accepted for publication (in revised form) April 1, 1998; published electronically March 17, 1999.

http://www.siam.org/journals/sicon/37-3/32062.html

†Centre de Mathématiques, UMR 7640 du CNRS, Ecole Polytechnique, F-91128 Palaiseau, France (allibert@math.polytechnique.fr).

As the problem is linear, we may consider only initial conditions such that $c_1 = c_2 = 0$ and look for a control function β that drives the solution of (1) to

$$(\Phi, \partial_t\Phi, W, \partial_t W)|_{t=T} = (0, 0, 0, 0).$$

Then, the control function β will drive any initial condition $(\Phi_0, \Phi_1, W_0, W_1)+(c_1, 0, c_2, 0)$ to the equilibrium $(c_1, 0, c_2, 0)$.

As the geometric control hypothesis of [BLR] does not hold, the space F_T of controlled functions should be different from the space $L^2(\Omega) \oplus H^{-1}(\Omega) \oplus L^2(\Gamma_0) \oplus H^{-1}(\Gamma_0)$. Nevertheless, as the geometry is simple, we can separate the variables in order to get some results about F_T.

In [MZ1] and [MZ2], Micu and Zuazua have used this method. They get an infinite number of one-dimensional systems, indexed by an integer n. Then they apply the Hilbert uniqueness method (HUM) in order to get the adjoint problem

$$(2) \qquad \begin{cases} \Delta\Psi - \Psi_{tt} = 0 & \text{over } \Omega \times (0, T), \\ \frac{\partial\Psi}{\partial n} = 0 & \text{over } \Gamma_1 \times (0, T), \\ \frac{\partial\Psi}{\partial n} = -V_t & \text{over } \Gamma_0 \times (0, T), \\ V_{tt} - V_{yy} + \Psi_t = 0 & \text{over } \Gamma_0 \times (0, T), \\ V_y(0, t) = V_y(1, t) & \text{over } (0, T), \end{cases}$$

with initial conditions $(\underline{\Psi}, \underline{V}) = ((\Psi, \Psi_t), (V, V_t))|_{t=0}$ in

$$\Big(H_0^1([0, 1]) \times L^2([0, 1]) \times \mathbb{C} \times \mathbb{C}\Big)e^{i\pi ny}.$$

For $n \neq 0$, the HUM method, described in [MZ2, sect. 2.4], shows that if there is a constant $C(n, T)$ such that for any solution of problem (2)

$$||(\underline{\Psi}, \underline{V})||_{\mathcal{X}}^2 \leq C(n, T) \int_{-T}^{T} |V_{tt}(0, t)|^2 \ dt,$$

then for any initial data

$$(\Phi_0, \Phi_1, W_0, W_1) \in (L^2([0, 1]) \times H^{-1}([0, 1]) \times \mathbb{C} \times \mathbb{C})e^{i\pi ny}$$

a control β can be built, such that the solution of the control problem (1) satisfies

$$(\Phi, \partial_t\Phi, W, \partial_t W)|_{t=T} = 0.$$

This method naturally leads to controls β defined as $\beta := \frac{d^2}{dt^2}\rho$, with

$$(3) \qquad |\rho|_{L^2(\Gamma_0 \times (0,T))} \leq C(n, T)|(\Phi_0, \Phi_1, W_0, W_1)|_{L^2([0,1]) \times H^{-1}([0,1]) \times \mathbb{C} \times \mathbb{C}}.$$

This is why we will consider controls β in H^{-2}.

For $n = 0$, the same result holds, providing $c_1 = c_2 = 0$ (see [MZ2, sect. 2.5]).

Now suppose we take an initial condition

$$(\Phi_0, \Phi_1, W_0, W_1) \in L^2(\Omega) \oplus H^{-1}(\Omega) \oplus L^2(\Gamma_0) \oplus H^{-1}(\Gamma_0),$$

with $c_1 = c_2 = 0$ and such that its Fourier series in y

$$(\Phi_0, \Phi_1, W_0, W_1) = \sum_n (\Phi_0, \Phi_1, W_0, W_1)_n e^{i\pi ny}$$

is such that

$$\left(C(n,T)||(\Phi_0,\Phi_1,W_0,W_1)_n||_{H_0^1([0,1])\times L^2([0,1])\times\mathbb{C}\times\mathbb{C}}\right)_n \in l^2(\mathbb{Z}). \tag{4}$$

Then for each value of n, we can build a control β_n that drives $(\Phi_0,\Phi_1,W_0,W_1)_n$ to 0 in time T. Therefore, if we put $\beta := \sum_n \beta_n$, the control β drives (Φ_0,Φ_1,W_0,W_1) to 0 in time T. Notice that the control β exists and is in $H^{-2}(\Gamma_0\times(0,T))$ because of (4) and (3). Hence the space H of such initial data is a subset of F_T. So any result about the constants $C(n,T)$ gives an estimate about the space F_T.

In [MZ2], Micu and Zuazua use explicit spectral results and Ingham techniques to prove that for any real number $\alpha > 1$, there is a positive constant K_α such that for any integer n, and any time $T > 2$,

$$C(n,T) \leq K_\alpha e^{K_\alpha n^\alpha}.$$

In this case, though, Ingham techniques fail to give sharp estimates because frequencies $\nu_{n,k}$ such that $k = o(n)$ have no finite gap between each other; $\nu_{n,k+1} - \nu_{n,k} \to 0$ when n goes to infinity.

1.2. Main result. In this paper, we will replace these techniques by biorthogonal sequences methods in the bad part of the spectrum. It will allow us to get the following theorem.

THEOREM 1. *For any positive δ and ϵ, there is a time $T(\epsilon,\delta)$ smaller than $\frac{C_\delta}{\epsilon^{1+\delta}}$ and a positive constant $C_{\epsilon,\delta}$ such that for any positive integer n and any time $T > T(\epsilon,\delta)$,*

$$C(n,T) \leq C_{\epsilon,\delta}\ e^{\epsilon|n|}.$$

In other terms, this theorem means that any initial data (Φ_0,Φ_1,W_0,W_1) such that

$$\left(e^{\alpha|n|}||(\Phi_0,\Phi_1,W_0,W_1)_n||_{H_0^1([0,1])\times L^2([0,1])\times\mathbb{C}\times\mathbb{C}}\right)_n \in l^2(\mathbb{Z})$$

belongs to F_T if

$$T > T(\alpha,\delta).$$

Thus, the following corollary holds.

COROLLARY 1.

$$C^\omega \times C^\omega \times C^\omega \times C^\omega \subset \bigcup_T F_T.$$

As Ω has a boundary, C^ω functions over Ω must be understood as functions that have an analytic continuation over a neighborhood of Ω. This corollary states that any such initial condition can be controlled within finite time. Of course, this time depends on the condition.

Indeed, it would be enough to take functions that are analytic with respect to y and whose values are Sobolev functions with respect to x. Such spaces may not be familiar to the reader. However, one may keep in mind that regularity with respect to y is the key.

Proof. Let $(\Phi_0, \Phi_1, W_0, W_1)(x, y)$ be an element of $L^2(\Omega) \times H^{-1}(\Omega) \times L^2(\Gamma) \times H^{-1}(\Gamma)$, with

$$(\Phi_0, \Phi_1, W_0, W_1)(x, y) = \sum_n (\Phi_0, \Phi_1, W_0, W_1)_n(x) e^{i\pi n y}.$$

First, notice that we can subtract $(c_1, 0, c_2, 0)$ from the initial condition in order to ensure that $c_1 = c_2 = 0$.

The Fourier coefficients can be computed by

$$(\Phi_0, \Phi_1, W_0, W_1)_n(x) = \int_0^1 (\Phi_0, \Phi_1, W_0, W_1)(x, y) e^{-i\pi n y}\, dy. \tag{5}$$

If $(\Phi_0, \Phi_1, W_0, W_1)$ is analytic (analytic with respect to y is enough), then there is a positive real number ϵ such that

$$y \mapsto (\Phi_0, \Phi_1, W_0, W_1)(., y + i\epsilon), \tag{6}$$

and

$$y \mapsto (\Phi_0, \Phi_1, W_0, W_1)(., y - i\epsilon). \tag{7}$$

are analytic and hence L^2.

Now we can shift the integration contour in (5) to the imaginary direction (because $(\Phi_0, \Phi_1, W_0, W_1)$ is periodic with respect to y). Thus,

$$(\Phi_0, \Phi_1, W_0, W_1)_n(x) = \int_0^1 (\Phi_0, \Phi_1, W_0, W_1)(x, y + i\epsilon) e^{-i\pi n y + n\pi\epsilon}\, dy.$$

Thus, $(\Phi_0, \Phi_1, W_0, W_1)_n e^{n\pi\epsilon}$ is the nth Fourier coefficient of function (6) that is in L^2. Therefore,

$$((\Phi_0, \Phi_1, W_0, W_1)_n e^{n\pi\epsilon})_n \in l^2(\mathbb{Z}).$$

For symmetric reasons

$$((\Phi_0, \Phi_1, W_0, W_1)_n e^{-n\pi\epsilon})_n \in l^2(\mathbb{Z}).$$

Thus,

$$((\Phi_0, \Phi_1, W_0, W_1)_n e^{|n|\pi\epsilon})_n \in l^2(\mathbb{Z}).$$

Therefore, by Theorem 1, if $T > T(\pi\epsilon, 1)$, $(\Phi_0, \Phi_1, W_0, W_1) \in F_T$. Therefore,

$$(\Phi_0, \Phi_1, W_0, W_1) \in \bigcup_T F_T. \qquad \square$$

This paper aims at proving Theorem 1. In section 2, we will recall Micu and Zuazua's spectral results and give more detail about the eigenvalues of the problem. Then in section 3 we will state two propositions, dealing with high and low frequencies, and show how we can prove the theorem out of them. Finally, we will prove those propositions in sections 4 and 5.

2. Notations and preliminaries. At first, let us recall a few notations and results from [MZ1]. $\mathcal{X}$ is the energy space $H^1(\Omega)\oplus L^2(\Omega)\oplus H^1(\Gamma_0)\oplus L^2(\Gamma_0)$. Reference [MZ1] introduces a skew-adjoint operator over $\mathcal{X}$, denoted $\mathcal{A}_C$, whose domain is a subset of $H^2(\Omega)\times H^1(\Omega)\times H^2(\Gamma_0)\times H^1(\Gamma_0)$. It is defined by

$$\mathcal{A}_C(\Psi_0,\Psi_1,V_0,V_1)=(-\Psi_1,-\Delta\Psi_0,-V_1,-V_{0yy}+V_1+\Psi_1).$$

It can be diagonalized over an orthogonal basis of eigenvectors

$$\xi_\nu\cos n\pi y=\begin{pmatrix}\xi_\nu^1\cos n\pi y\\ \xi_\nu^2\cos n\pi y\\ \xi_\nu^3\cos n\pi y\\ \xi_\nu^4\cos n\pi y\end{pmatrix}=\begin{pmatrix}\frac{1}{\nu}\cosh(\sqrt{n^2\pi^2+\nu^2}(x-1))\cos n\pi y\\ -\cosh(\sqrt{n^2\pi^2+\nu^2}(x-1))\cos n\pi y\\ \frac{-\sqrt{n^2\pi^2+\nu^2}}{\nu^2}\sinh(\sqrt{n^2\pi^2+\nu^2})\cos n\pi y\\ \frac{\sqrt{n^2\pi^2+\nu^2}}{\nu}\sinh(\sqrt{n^2\pi^2+\nu^2})\cos n\pi y\end{pmatrix},$$

and the solution of (2) with initial condition $\xi_\nu\cos n\pi y$ is such that

$$\begin{pmatrix}\Psi(x,y,t)\\ \Psi_t(x,y,t)\\ V(y,t)\\ V_t(y,t)\end{pmatrix}=\xi_\nu(x)\cos n\pi y\ e^{\nu t}.$$

As this basis is not normalized, we will denote $\Xi_\nu=||\xi_\nu||_{\mathcal{X}}$.

The values ν that belong to $i\mathbb{R}$ are the eigenvalues of $\mathcal{A}_C$. For any integer n they are the roots of

$$\nu^2+n^2\pi^2=-z^2,$$

with

$$\tan z=\frac{z^2+n^2\pi^2}{z^3}. \tag{8}$$

For $n=0$ only, $\nu=0$ belongs to the spectrum. Otherwise, (8) has two kinds of solutions:

• Either z is real, denoted $z_{n,k}$ with k in $\mathbb{N}^*$. This corresponds to eigenvalues $\nu_{n,k}$ of modulus greater than $n\pi$. We denote $\nu_{n,-k}=-\nu_{n,k}$.

• Or $z=it$ is imaginary, with $\tanh t=\frac{-t^2+n^2\pi^2}{t^3}$. We denote it $z_{n,*}$. It gives rise to two eigenvalues denoted $\nu_{n,*}$ and $\nu_{n,**}=-\nu_{n,*}$ whose moduli are smaller than $n\pi$.

For any $(\underline{\Psi},\underline{V})$ in $\mathcal{X}$,

$$\begin{pmatrix}\underline{\Psi}(x,y)\\ \underline{V}(y)\end{pmatrix}=\sum_{\substack{n\in\mathbb{Z}\\ k\in\mathbb{Z}^*\cup\{*,**\}}}\frac{a_{n,k}}{\Xi_{n,k}}\ \xi_{n,k}(x,y),$$

with $(a_{n,k})_{n,k}\in l^2$.

We will denote $\mathcal{X}_{-1}$ the space of functions defined by the above formula, with the condition $(\frac{a_{n,k}}{\nu_{n,k}})_{n,k}\in l^2$, and we will denote $\mathcal{W}$ the same type of space with the condition $(a_{n,k}(1+|n|))_{n,k}\in l^2$.

If we denote (Ψ,V) as the solution of problem (2) with the initial conditions $(\Psi,\Psi_t)|_{t=0}=\underline{\Psi}$, $(V,V_t)|_{t=0}=\underline{V}$, the well-posedness of the problem implies that

$$||\Psi||_{H^1(\Omega\times(0,T))}\leq C||(\underline{\Psi},\underline{V})||_{\mathcal{X}}$$

and

$$||\Psi||_{L^2(\Omega\times(0,T))} \leq C||(\underline{\Psi},\underline{V})||_{\mathcal{X}_{-1}}.$$

To end with, we will also write for any $(\underline{\Psi},\underline{V})$ in $\mathcal{X}$ that

$$(\underline{\Psi},\underline{V}) \in \mathcal{X}^{n_0} \text{ if } n \neq n_0 \Rightarrow a_{n,k} = 0,$$
$$(\underline{\Psi},\underline{V}) \in \mathcal{X}^{(1)} \text{ if } |k| > |n| \Rightarrow a_{n,k} = 0,$$
$$(\underline{\Psi},\underline{V}) \in \mathcal{X}^{(2)} \text{ if } \Big(|k| \leq |n| \text{ or } k \in \{*, **\}\Big) \Rightarrow a_{n,k} = 0.$$

For any $(\underline{\Psi},\underline{V}) \in \mathcal{X}$, we have the following decomposition:

$$(\underline{\Psi},\underline{V}) = (\underline{\Psi},\underline{V})^{(1)} + (\underline{\Psi},\underline{V})^{(2)},$$

with $(\underline{\Psi},\underline{V})^{(1)} \in \mathcal{X}^{(1)}$ and $(\underline{\Psi},\underline{V})^{(2)} \in \mathcal{X}^{(2)}$.

Moreover, I will be the set of (k,n) such that $k \in \{*, **\}$ or $|k| \leq |n|$, and we will agree that $** = -*$.

Notice that for any integer n, $\mathcal{X}^n \subset \mathcal{W}$.

The following lemma gives a few results about the numbers $\nu_{n,k}$.

Lemma 1. *For all integers n and k, the following inequalities hold:*

$$\text{(i)} \quad \sqrt{n^2\pi^2 + (k-1)^2\pi^2} \leq |\nu_{n,k}| \leq \sqrt{n^2\pi^2 + \left(k - \frac{1}{2}\right)^2 \pi^2},$$

$$\text{(ii)} \quad \left|\sqrt{|\nu_{n,k}|^2 - n^2\pi^2} - \left(k - \frac{1}{2}\right)\pi\right| \geq \frac{C}{1+n^2},$$

$$\text{(iii)} \quad \left|\sqrt{|\nu_{n,k}|^2 - n^2\pi^2} - (k-1)\pi\right| \geq \frac{C}{k}.$$

Furthermore, with n going to infinity, we also have

$$\text{(iv)} \quad \pi - \frac{|\nu_{n,*}|}{n} \sim \frac{1}{2}\pi^{\frac{1}{3}} n^{-\frac{2}{3}}.$$

Proof. We denote, as in [MZ2], $z_{n,k} = \sqrt{|\nu_{n,k}|^2 - n^2\pi^2}$. $z_{n,k}$ satisfies

$$\tan z_{n,k} = \frac{z_{n,k}^2 + n^2\pi^2}{z_{n,k}^3}.$$

(See Figure 2.1.)

Inequality (i) follows from $(k-1)\pi \leq z_{n,k} \leq (k - \frac{1}{2})\pi$ and $|\nu_{n,k}| = \sqrt{z_{n,k}^2 + n^2\pi^2}$.

As we have $|z_{n,k} - ((k-1)\pi + \frac{\pi}{2})| \geq |z_{n,1} - \frac{\pi}{2}|$, (ii) follows from $|z_{n,1} - \frac{\pi}{2}| \geq \frac{C}{1+n^2}$. To prove (iii), it is enough to bound $|y_{n,k} - (k-1)\pi|$ from below, where $y_{n,k}$ satisfies $\frac{y_{n,k}^2}{y_{n,k}^3} = C(y_{n,k} - (k-1)\pi)$. The latter equation has an explicit solution that ensures $|y_{n,k} - (k-1)\pi| \leq \frac{C}{k\pi}$.

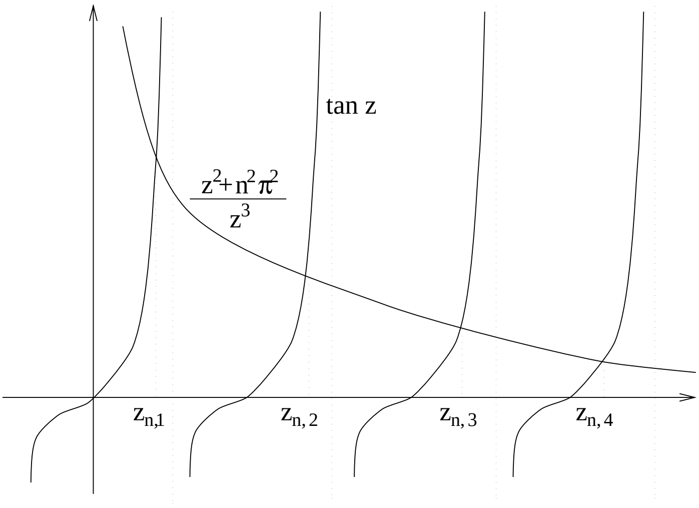


FIG. 2.1. *Determination of* $z_{n,k}$.

To end with, if we denote $z_{n,*} = \sqrt{n^2\pi^2 - |\nu_{n,*}|^2}$, we have

$$e^{2z_{n,*}} = \frac{z_{n,*}^3 - z_{n,*}^2 + \pi^2 n^2}{z_{n,*}^3 + z_{n,*}^2 - \pi^2 n^2}.$$

Thus,

$$z_{n,*} \sim (\pi^2 n^2)^{\frac{1}{3}} \Rightarrow \pi - \frac{|\nu_{n,*}|}{n} \sim \frac{1}{2}\pi^{\frac{1}{3}} n^{-\frac{2}{3}}(1 + o(1)). \qquad \square$$

From Lemma 1 and in view of the explicit value of $\xi_{n,k}$, we can see that there is a constant C such that for any (n,k) we have

$$\Xi_{n,k} \le Ce^{Cn^{\frac{2}{3}}}. \tag{9}$$

3. Proof of Theorem 1. In order to prove this theorem, we will need two estimates, for high and low frequencies.

PROPOSITION 1 (high frequencies). *There exists a constant C and a positive time T_1 such that for any initial condition $(\underline{\Psi}, \underline{V})$ in $\mathcal{W}$, we have*

$$||(\underline{\Psi}, \underline{V})^{(2)}||_{\mathcal{W}}^2 \le C\Big(||\Psi||^2_{L^2((-2T_1,2T_1)\times\Omega)} + ||V_{tt}||^2_{L^2((0,T_1)\times\Gamma_0)}\Big).$$

This proposition looks like the usual estimate given by microlocal analysis when the geometric control hypothesis of Bardos–Lebeau–Rauch holds (see [Le]). This is not surprising because the rays that correspond to those frequencies always hit the controlled boundary within a uniform finite time.

PROPOSITION 2 (low frequencies). *For any positive ϵ and δ, there exists a constant $C_{\epsilon,\delta}$, an integer $n_1(\epsilon)$, and a positive time $T_2(\epsilon,\delta) \le \frac{C_\delta}{\epsilon^{1+\delta}}$ such that for any integer n greater than $n_1(\epsilon)$ and any initial condition $(\underline{\Psi}, \underline{V})$ in $\mathcal{X}^n$*

$$||(\underline{\Psi}, \underline{V})^{(1)}||_{\mathcal{X}}^2 \le C_{\epsilon,\delta}\; e^{2\epsilon|n|} \int_{-T_2(\epsilon,\delta)}^{T_2(\epsilon,\delta)} |V_{tt}(0,t)|^2 \; dt.$$

The difficult part of the problem is concentrated in this part of the spectrum, where the gap between the frequencies is not bounded from below. This is why the relevant constants appear in that proposition. Not much is known about those constants. When ϵ goes to 0, $C_{\epsilon,\delta}$ certainly goes to infinity; if it were bounded, we could take the limit and prove that any initial data can be controlled, without any hypothesis. This certainly is not true because of Bardos–Lebeau–Rauch's theorem.

To the author's knowledge, no such thing can be said about the constant C_δ when δ goes to 0.

We also have the following lemma.

Lemma 2. *There is a constant C such that for any integer n and any initial condition $(\underline{\Psi}, \underline{V})$ in $\mathcal{X}^n$*

$$||(\underline{\Psi}, \underline{V})^{(2)}||^2_{\mathcal{X}_{-1}} \leq \frac{C}{1+n^2}\ ||(\underline{\Psi}, \underline{V})^{(2)}||^2_{\mathcal{X}}.$$

Proof of Lemma 2. All frequencies in $(\underline{\Psi}, \underline{V})^{(2)}$ of nonzero amplitude are such that $|k| > |n|$.

Now, it follows from Lemma 1 that $|k| > |n| \Rightarrow |\nu_{n,k}|^2 \geq C(1+n^2)$. Furthermore,

$$||(\underline{\Psi}, \underline{V})||^2_{\mathcal{X}_{-1}} = \sum_k \frac{1}{|\nu_{n,k}|^2}|a_{n,k}|^2, \qquad ||(\underline{\Psi}, \underline{V})||^2_{\mathcal{X}} = \sum_k |a_{n,k}|^2. \qquad \square$$

Let us first prove that Propositions 1 and 2 and Lemma 2 imply that Theorem 1 holds.

Proof. Let ϵ and δ be two positive real numbers. Out of Propositions 1 and 2, we get two positive times, denoted T_1 and $T_2(\epsilon, \delta)$. Let us define

$$T(\epsilon, \delta) = \sup(T_1, T_2(\epsilon, \delta)).$$

Let n be a positive integer and $(\underline{\Psi}, \underline{V})$ any initial condition in $\mathcal{X}^n$. Then we have

$$\begin{aligned} ||(\underline{\Psi}, \underline{V})||^2_{\mathcal{X}} &= ||(\underline{\Psi}, \underline{V})^{(1)}||^2_{\mathcal{X}} + ||(\underline{\Psi}, \underline{V})^{(2)}||^2_{\mathcal{X}}, \\ &\leq ||(\underline{\Psi}, \underline{V})^{(1)}||^2_{\mathcal{X}} + ||(\underline{\Psi}, \underline{V})^{(2)}||^2_{\mathcal{W}}. \end{aligned}$$

Hence, by Propositions 1 and 2, for $n \geq n_1(\epsilon)$,

$$||(\underline{\Psi}, \underline{V})||^2_{\mathcal{X}} \leq C_{\epsilon,\delta}\ e^{2\epsilon|n|} \int_{-T_2(\epsilon,\delta)}^{T_2(\epsilon,\delta)} |V_{tt}(0,t)|^2\ dt$$

$$+\ C\ \left[\int_0^{T_1} |V_{tt}(0,t)|^2\ dt\ +\ ||\Psi||^2_{L^2((-2T_1,2T_1)\times\Omega)}\right].$$

Thus,

$$||(\underline{\Psi}, \underline{V})||^2_{\mathcal{X}} \leq C'_{\epsilon,\delta}\ e^{2\epsilon|n|} \int_{-T(\epsilon,\delta)}^{T(\epsilon,\delta)} |V_{tt}(0,t)|^2\ dt + C\ ||\Psi||^2_{L^2((-2T_1,2T_1)\times\Omega)}.$$

As the problem is well posed,

$$\begin{aligned}||(\underline{\Psi},\underline{V})||^2_{\mathcal{X}} &\leq C'_{\epsilon,\delta}\ e^{2\epsilon|n|}\int_{-T(\epsilon,\delta)}^{T(\epsilon,\delta)}|V_{tt}(0,t)|^2\ dt + C'\ ||(\underline{\Psi},\underline{V})||^2_{\mathcal{X}_{-1}},\\ &\leq C'_{\epsilon,\delta}\ e^{2\epsilon|n|}\int_{-T(\epsilon,\delta)}^{T(\epsilon,\delta)}|V_{tt}(0,t)|^2\ dt + C'\ ||(\underline{\Psi},\underline{V})^{(1)}||^2_{\mathcal{X}_{-1}} + C'\ ||(\underline{\Psi},\underline{V})^{(2)}||^2_{\mathcal{X}_{-1}},\\ &\leq C'_{\epsilon,\delta}\ e^{2\epsilon|n|}\int_{-T(\epsilon,\delta)}^{T(\epsilon,\delta)}|V_{tt}(0,t)|^2\ dt + C'\ ||(\underline{\Psi},\underline{V})^{(1)}||^2_{\mathcal{X}} + C'\ ||(\underline{\Psi},\underline{V})^{(2)}||^2_{\mathcal{X}_{-1}}.\end{aligned}$$

Hence by Proposition 2,

$$||(\underline{\Psi},\underline{V})||^2_{\mathcal{X}} \leq C'_{\epsilon,\delta}\ e^{2\epsilon|n|}\int_{-T(\epsilon,\delta)}^{T(\epsilon,\delta)}|V_{tt}(0,t)|^2\ dt\ +\ C'C_{\epsilon,\delta}\ e^{2\epsilon|n|}\int_{-T_2(\epsilon,\delta)}^{T_2(\epsilon,\delta)}|V_{tt}(0,t)|^2\ dt$$

$$+C'\ ||(\underline{\Psi},\underline{V})^{(2)}||^2_{\mathcal{X}_{-1}}.$$

Therefore, by Lemma 2,

$$||(\underline{\Psi},\underline{V})||^2_{\mathcal{X}} \leq C''_{\epsilon,\delta}\ e^{2\epsilon|n|}\int_{-T(\epsilon,\delta)}^{T(\epsilon,\delta)}|V_{tt}(0,t)|^2\ dt + \frac{C}{1+n^2}\ ||(\underline{\Psi},\underline{V})^{(2)}||^2_{\mathcal{X}}.$$

Thus,

$$||(\underline{\Psi},\underline{V})||^2_{\mathcal{X}} \leq C''_{\epsilon,\delta}\ e^{2\epsilon|n|}\int_{-T(\epsilon,\delta)}^{T(\epsilon,\delta)}|V_{tt}(0,t)|^2\ dt + \frac{C}{1+n^2}\ ||(\underline{\Psi},\underline{V})||^2_{\mathcal{X}}.$$

For $|n| \geq \sqrt{2C-1}$, $\frac{C}{1+n^2} \leq \frac{1}{2}$.

Thus, if $n \geq n_2(\epsilon) = \sup\{n_1(\epsilon), \sqrt{2C-1}\}$,

$$||(\underline{\Psi},\underline{V})||^2_{\mathcal{X}} \leq 2\ C''_{\epsilon,\delta}\ e^{2\epsilon|n|}\int_{-T(\epsilon,\delta)}^{T(\epsilon,\delta)}|V_{tt}(0,t)|^2\ dt.$$

As we can increase the constant to take care of the first $n_2(\epsilon)$ values of n, we have proved Theorem 1. (Note that it is enough to prove estimates for each $\mathcal{X}^n$ because of the orthogonality of the solutions for different values of n.) □

We still have to prove the two propositions. This will be dealt with in the following two sections.

4. Proof of Proposition 1 (high frequencies). A simple application of Ingham techniques would allow us to prove a slightly weaker form of this proposition; specifically, it would prove that

$$||(\underline{\Psi},\underline{V})^{(2)}||_{\mathcal{W}} \leq C\Big(||\Psi||_{L^2((-2T_1,2T_1)\times\Omega)} + ||V^{(2)}{}_{tt}||_{L^2((0,T_1)\times\Gamma_0)}\Big).$$

In order to get V instead of $V^{(2)}$ in the right-hand side of the estimate, we will use microlocal techniques.

The proof is based upon the techniques that are used when the geometric control hypothesis holds, e.g., when every ray hits the boundary. In our case, we will have to get rid of uncontrolled rays by cutting off their frequencies. We cannot take a brutal cutoff because it would not be pseudodifferential. Thus, we will have to introduce smooth pseudodifferential cutoffs, which we will denote P_1 and P_2. These will be chosen to ensure $P_1 + P_2 = Id$, but their images will have a nonzero intersection. Dealing with this intersection will be the main source of dissemblance between the proof of this proposition and the classical one (see [Le], for instance).

Let T_1 be a positive time that will be fixed later. Let us introduce a few functional spaces.

Let $\mathcal{Z}$ be the set of all (Ψ, V) in $L^2((-2T_1, 2T_1) \times \Omega) \times L^2((0, T_1) \times \Gamma_0)$ such that

$$\begin{cases} \Box\Psi = 0 \text{ over } (-2T_1, 2T_1) \times \Omega, \\ \frac{\partial \Psi}{\partial n} = 0 \text{ over } \Gamma_1 \times (-2T_1, 2T_1), \\ \frac{\partial \Psi}{\partial n} = -V_t \text{ over } \Gamma_0 \times (-2T_1, 2T_1), \\ V_{tt} - V_{yy} + \Psi_t = 0 \text{ over } \Gamma_0 \times (-2T_1, 2T_1), \end{cases}$$

and

$$V_{tt} \text{ belongs to } L^2((0, T_1) \times \Gamma_0).$$

The norm over $\mathcal{Z}$ will be defined as

$$||(\Psi, V)||^2_{\mathcal{Z}} = ||\Psi||^2_{L^2((-2T_1,2T_1)\times\Omega)} + ||V_{tt}||^2_{L^2((0,T_1)\times\Gamma_0)}.$$

The functions of $\mathcal{Z}$ can always be written as

$$\Psi(x, y, t) = \sum_{\substack{n\in\mathbb{Z} \\ k\in\mathbb{Z}^*\cup\{*,**\}}} \frac{a_{n,k}}{\Xi_{n,k}} e^{\nu_{n,k}t} \cos \pi ny \; \xi^1_{n,k}(x),$$

$$V(y, t) = \sum_{\substack{n\in\mathbb{Z} \\ k\in\mathbb{Z}^*\cup\{*,**\}}} \frac{a_{n,k}}{\Xi_{n,k}} e^{\nu_{n,k}t} \cos \pi ny \; \xi^3_{n,k},$$

with $(\frac{a_{n,k}}{\nu_{n,k}})_{n,k} \in l^2$.

Therefore, we can separate the high frequencies from the low ones. Let us denote

$$(\Psi, V) \in \mathcal{Z}^{(1)} \text{ if } |k| > |n| \Rightarrow a_{n,k} = 0.$$

(Note for easier understanding: The functions of $\mathcal{Z}^{(1)}$ have the same frequencies as those of $\mathcal{X}^{(1)}$. Nevertheless, $\mathcal{X}^{(1)} \neq \mathcal{Z}^{(1)}$.)

To end with, for a $(\underline{\Psi}, \underline{V})$ in $\mathcal{W}$, we shall write $(\underline{\Psi}, \underline{V}) \in \mathcal{W}^{(3)}$ if

$$\left(|k| \leq \frac{n}{2} \text{ or } |k| = *\right) \Rightarrow a_{n,k} = 0.$$

If $(\underline{\Psi}, \underline{V})$ belongs to $\mathcal{W} \cap \mathcal{X}^{(2)}$, then $(\underline{\Psi}, \underline{V})$ belongs to $\mathcal{W}^{(3)}$.

Let S map any initial condition $(\underline{\Psi}, \underline{V})$ in $\mathcal{X}_{-1}$ to the solution of problem (2) over $(-2T_1, 2T_1) \times \Omega$ with $(\underline{\Psi}, \underline{V})$ as initial conditions.

In view of Proposition (2.1) of [MZ2], if we put $(\underline{\Psi}, \underline{V})(x, y) = \sum_n (\underline{\Psi}, \underline{V})_n(x) e^{iny}$, we have

$$\int_0^{T_1} |(V_n)_{tt}|^2 \; dt \leq C(n^4 + 1)||(\underline{\Psi}, \underline{V})_n||^2_{H^1(0,1)\times L^2(0,1)\times\mathbb{R}\times\mathbb{R}}.$$

Therefore,

$$S(\mathcal{W}) \subset \mathcal{Z}.$$

Let us define the mapping $\tilde{\imath}$ as follows:

$$\tilde{\imath}:\left|\begin{array}{l}\mathcal{W}^{(3)}\oplus\mathcal{Z}^{(1)}\to\mathcal{Z}\\ ((\underline{\Psi},\underline{V}),(\Phi,W))\mapsto S(\underline{\Psi},\underline{V})+(\Phi,W),\end{array}\right.$$

$$\tilde{\imath}((\underline{\Psi},\underline{V})\ \ (\Phi,W))=0\Leftrightarrow S(\underline{\Psi},\underline{V})+(\Phi,W)=0\Leftrightarrow S(\underline{\Psi},\underline{V})=-(\Phi,W).$$

Therefore, the only nonzero coefficients $a_{n,k}$ are such that $\frac{|n|}{2}<|k|\leq|n|$. $\tilde{\imath}$ induces an injective mapping i from $G=(\mathcal{W}^{(3)}\oplus\mathcal{Z}^{(1)})/\ker\tilde{\imath}$ to $\mathcal{Z}$.

LEMMA 3. *If T_1 is big enough, i is onto.*

We will prove Lemma 3 later on.

First, let us see how we can end the proof of the proposition out of this lemma for such a T_1.

As i is a one-to-one continuous morphism, its reciprocal mapping is also continuous. Thus, there exists a constant C such that for any couple $((\underline{\Psi},\underline{V}),\ (\Phi,W))$ in $\mathcal{W}^{(3)}\oplus\mathcal{Z}^{(1)}$, we have

$$||[(\underline{\Psi},\underline{V}),\ (\Phi,W)]||_G\leq C\ ||\tilde{\imath}((\underline{\Psi},\underline{V}),\ (\Phi,W))||_{\mathcal{Z}}. \tag{10}$$

Here, $[(\underline{\Psi},\underline{V}),\ (\Phi,W)]$ denotes the class of $((\underline{\Psi},\underline{V}),\ (\Phi,W))$ in G, and the norm on this quotient space has the usual definition:

$$||[(\underline{\Psi},\underline{V}),\ (\Phi,W)]||_G^2=\inf_{(w_1,w_2)\in\ker\tilde{\imath}}||((\underline{\Psi},\underline{V}),\ (\Phi,W))+(w_1,w_2)||^2_{\mathcal{W}^{(3)}\oplus\mathcal{Z}^{(1)}}.$$

Let $(\underline{\Psi},\underline{V})$ be any element of $\mathcal{W}$.

We can split it into $(\underline{\Psi},\underline{V})=(\underline{\Psi},\underline{V})_1+(\underline{\Psi},\underline{V})_3$ with $(\underline{\Psi},\underline{V})_3\in\mathcal{W}^{(3)}$ and $S(\underline{\Psi},\underline{V})_1=(\Psi_1,V_1)\in\mathcal{Z}^{(1)}$ (by deciding for each frequency whether $|k|\leq\frac{|n|}{2}$).

Indeed, $S(\underline{\Psi},\underline{V})_1$ has the same frequencies as $(\underline{\Psi},\underline{V})_1$ that belongs to $\mathcal{W}$. Thus,

$$\begin{aligned}||\tilde{\imath}((\underline{\Psi},\underline{V})_3,(\Psi_1,V_1))||^2_{\mathcal{Z}}&=||S(\underline{\Psi},\underline{V})_3+S(\underline{\Psi},\underline{V})_1||^2_{\mathcal{Z}}\\&=||\Psi||^2_{L^2((-2T_1,2T_1)\times\Omega)}+||V_{tt}||^2_{L^2((0,T_1)\times\Gamma_0)}.\end{aligned}\tag{11}$$

On the other hand,

$$||[(\underline{\Psi},\underline{V})_3\ ,\ (\Psi_1,V_1)]||_G^2=\inf_{Sw_2=-w_1}\Big(||(\underline{\Psi},\underline{V})_3+w_2||^2_{\mathcal{W}^{(3)}}+||(\Psi_1,V_1)+w_1||^2_{\mathcal{Z}^{(1)}}\Big).$$

The nonzero coefficients in w_2 are such that $\frac{|n|}{2}<|k|\leq|n|$. They correspond to elements of the basis that are orthogonal to $\mathcal{W}\cap\mathcal{X}^{(2)}$. So, as the basis is orthogonal, for any initial condition $(\underline{\Psi},\underline{V})$ in $\mathcal{W}$

$$\begin{aligned}||(\underline{\Psi},\underline{V})^{(2)}||^2_{\mathcal{W}}&=\inf_{Sw_2=-w_1}||(\underline{\Psi},\underline{V})_3+w_2||^2_{\mathcal{W}^{(3)}},\\&\leq\inf_{Sw_2=-w_1}\Big(||(\underline{\Psi},\underline{V})_3+w_2||^2_{\mathcal{W}^{(3)}}+||(\Psi_1,V_1)+w_1||^2_{\mathcal{Z}^{(1)}}\Big).\end{aligned}$$

Therefore, by definition of the norm over G,

$$||(\underline{\Psi},\underline{V})^{(2)}||^2_{\mathcal{W}}\leq||[(\underline{\Psi},\underline{V})_3,\ (\Psi_1,V_1)]||^2_G.$$

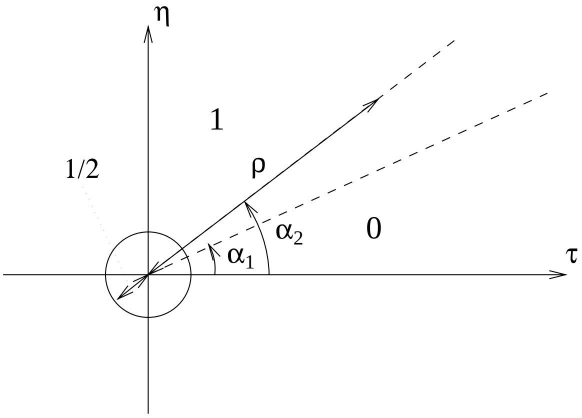


FIG. 4.1. *Angles α_1 and α_2.*

Thus, out of (10),

$$||(\underline{\Psi},\underline{V})^{(2)}||^2_{\mathcal{W}} \le C\ ||\tilde{\imath}((\underline{\Psi},\underline{V})_3,\ (\Psi_1,V_1))||^2_{\mathcal{Z}}.$$

Then, out of (11),

$$||(\underline{\Psi},\underline{V})^{(2)}||^2_{\mathcal{W}} \le C\ \Big(||\Psi||^2_{L^2((-2T_1,2T_1)\times\Omega)} + ||V_{tt}||^2_{L^2((0,T_1)\times\Gamma_0)}\Big). \qquad \square$$

Proof of Lemma 3. In order to prove that i is surjective, we have to prove that for any couple (Φ,W) in $\mathcal{Z}$, we can find an element $((\underline{\Psi},\underline{V})_3,\ (\Phi_1,W_1))$ of $\mathcal{W}^{(3)}\oplus\mathcal{Z}^{(1)}$ such that

$$\tilde{\imath}((\underline{\Psi},\underline{V})_3,\ (\Phi_1,W_1)) = (\Phi,W).$$

Let (Φ,W) be an element in $\mathcal{Z}$. It can be written as

$$\Phi(x,y,t) = \sum_{\substack{n\in\mathbb{Z}\\ k\in\mathbb{Z}^*\cup\{*,**\}}} \frac{a_{n,k}}{\Xi_{n,k}}\ e^{\nu_{n,k}t}\cos\pi ny\ \xi^1_{n,k}(x),$$

$$W(y,t) = \sum_{\substack{n\in\mathbb{Z}\\ k\in\mathbb{Z}^*\cup\{*,**\}}} \frac{a_{n,k}}{\Xi_{n,k}}\ e^{\nu_{n,k}t}\cos\pi ny\ \xi^3_{n,k}.$$

We shall consider Φ as defined over $\mathbb{R}_t\ \times\ \mathbb{R}_x\ \times\ \mathbb{R}_y$ by continuing the functions $\xi(x)$ by $\xi(x)=0$ if $x\notin[0,1]$.

Let us denote $\tilde\Omega=(0,1)\times\mathbb{R}$. Φ is a distribution that satisfies $\Box\Phi=0$ over $\tilde\Omega\times\mathbb{R}_t$ and whose restriction outside $\tilde\Omega\times\mathbb{R}_t$ is zero. Moreover, Φ belongs to $\mathcal{S}'(\mathbb{R}^3)$.

We shall introduce two symbols. Let γ_1 and γ_2 be two real numbers such that $\frac{1}{2}<\gamma_2<\gamma_1<1$ and denote $\alpha_k=\frac{1}{\sqrt{1+\gamma_k^2}}$, $k=1,2$. See Figure 4.1.

We define the symbol $\sigma_1(\eta,\tau)$ (the degree of which is zero) as follows.

If $\eta^2+\tau^2$ is greater than $\frac{1}{2}$, σ_1 is homogeneous of degree 0 and is an even, C^∞ function of $\alpha=\frac{\eta}{\tau}$. It is equal to 1 in the neighborhood of the set $|\frac{\eta}{\tau}|\ge\alpha_2$ and to 0 in the neighborhood of $|\frac{\eta}{\tau}|\le\alpha_1$. Furthermore, on the other regions, we demand that $\sigma_1(\eta,\tau)$ remain C^∞.

Let us denote $P_1 = Op(\sigma_1)$ as the pseudodifferential operator associated with the symbol σ_1. It is tangential with respect to x and of degree zero. It is operating over $\mathcal{S}'(\mathbb{R}^3)$.

If we put $\sigma_2 = 1 - \sigma_1$ and $P_2 = Op(\sigma_2)$, we get another operator with the same properties, such that $P_1 + P_2 = Id$.

Let us define $\Phi_1 = P_1\Phi$ and $\Phi_2 = P_2\Phi$. We have $\Phi_1 + \Phi_2 = \Phi$.

We may compute the explicit form of Φ_2:

$$\widehat{\Phi} = \sum_{n,k} \frac{a_{n,k}}{\Xi_{n,k}}\ \widehat{\xi^1_{n,k}}(\xi)\ \frac{\delta_{\pi n} + \delta_{-\pi n}}{2}(\eta)\ \delta_{-i\nu_{n,k}}(\tau),$$

so

$$\sigma_2(\eta,\tau)\widehat{\Phi} = \sum_{n,k} \frac{a_{n,k}}{\Xi_{n,k}}\ \sigma_2(\pi n, -i\nu_{n,k})\ \widehat{\xi^1_{n,k}}(\xi)\ \frac{\delta_{\pi n} + \delta_{-\pi n}}{2}(\eta)\ \delta_{-i\nu_{n,k}}(\tau),$$

then

$$\Phi_2 = P_2\Phi = \sum_{n,k} \frac{a_{n,k}}{\Xi_{n.k}}\ \sigma_2(\pi n, -i\nu_{n,k})\ \xi^1_{n,k}(x)\ \cos ny\pi\ e^{t\nu_{n,k}}.$$

Like Φ, Φ_2 satisfies $\Box\Phi_2 = 0$ in $\Omega \times \mathbb{R}_t$.

Furthermore, if we denote

$$W_2 = P_2 W = \sum_{n,k} \frac{a_{n,k}}{\Xi_{n,k}}\ \sigma_2(\pi n, -i\nu_{n,k})\ \xi^3_{n,k}\ \cos ny\pi\ e^{t\nu_{n,k}},$$

we can notice that (Φ_2, W_2) still is a solution to problem (2).

The same computation and remark can be done about Φ_1 and W_1.

Let us study the singularities of W_2.

As P_2 is of order 0 and W_{tt} belongs to $L^2((0,T_1)\times\Gamma_0)$, W_{2tt} also belongs to $L^2((0,T_1)\times\Gamma_0)$. Furthermore, as P_2 is tangential with respect to x and σ_2 is zero near $|\frac{\eta}{\tau}| \geq \alpha_2$, we have

$$WF(W_2) \subset \left\{|\frac{\eta}{\tau}| \leq \alpha_2\right\}. \tag{12}$$

Now, on the part $\{|\frac{\eta}{\tau}| \leq \alpha_2\}$ of the cotangent set, we have $\tau \neq 0$, so as W_{2tt} belongs to $L^2((0,T_1)\times\Gamma_0)$ we know that if $\sigma \in ((0,T_1)\times\Gamma_0)\times\{|\frac{\eta}{\tau}| \leq \alpha_2\}$, W_2 belongs to H^2_σ. Together with the inclusion (12), this means that W_2 belongs to $H^2((0,T_1)\times\Gamma_0)$; thus,

$$W_{2yy} \in L^2((0,T_1)\times\Gamma_0).$$

This means that

$$\int_0^{T_1}\int_0^1 \left|\sum_{n,k} \frac{n^2 a_{n,k}}{\Xi_{n,k}} \sigma_2(\pi n, -i\nu_{n,k})\xi^3_{n,k}\ \cos ny\pi\ e^{t\nu_{n,k}}\right|^2\ dy\ dt < +\infty.$$

Now, as $\sigma_2(\pi n, -i\nu_{n,k})$ is 0 if $|k| = *$ or $|k| \leq \gamma_2 n$, the sum ranges only on $|k| \geq \gamma_2|n|$; therefore, through Lemma 1, there is a positive ϵ such that for any n, k, $|\nu_{n,k+1}| -$

$|\nu_{n,k}| \geq \epsilon$. Thus, according to Ingham's inequality (see, for instance, [Z, p. 222]), if we put $T_1 > \frac{2\pi}{\epsilon}$,

$$\sum_{n,k} \left| \frac{n^2 a_{n,k} \sigma_2(\pi n, -i\nu_{n,k}) \xi^3_{n,k}}{\Xi_{n,k}} \right|^2 < +\infty.$$

Now Lemma 1 proves that for $|k| \geq \gamma_2 |n|$

$$c \leq \Xi_{n,k} \leq C$$

and

$$\frac{c}{n} \leq \xi^3_{n,k} \leq \frac{C}{n}.$$

Thus, we get

$$\sum_{n,k} |n a_{n,k} \sigma_2(\pi n, -i\nu_{n,k})|^2 < +\infty.$$

Together with the same computation based upon $W_2 \in L^2((0,T_1) \times \Gamma_0)$, we get

$$\Big((1+|n|) a_{n,k} \sigma_2(\pi n, -i\nu_{n,k})\Big)_{n,k} \in l^2.$$

Thus, we have proved that (Φ_2, W_2) corresponds to a couple $(\underline{\Psi}, \underline{V})_3$ of $\mathcal{W}$ such that $(\Phi_2, W_2) = S(\underline{\Psi}, \underline{V})_3$. Moreover, the set of frequencies allowed by σ_2 in $(\underline{\Psi}, \underline{V})_3$ shows that $(\underline{\Psi}, \underline{V})_3$ belongs to $\mathcal{W}^{(3)}$.

To conclude, as $(\Phi_1, W_1) = (\Phi, W) - (\Phi_2, W_2)$ and (Φ_2, W_2) belongs to $S(\mathcal{W}) \subset \mathcal{Z}$, (Φ_1, W_1) also belongs to $\mathcal{Z}$. As it has the same frequencies as Φ_1, we even know that it belongs to $\mathcal{Z}^{(1)}$.

Thus, we have found a couple $((\underline{\Psi}, \underline{V})_3,\ (\Phi_1, W_1))$ in $\mathcal{W}^{(3)} \oplus \mathcal{Z}^{(1)}$ such that

$$\tilde{\imath}((\underline{\Psi}, \underline{V})_3,\ (\Phi_1, W_1)) = S(\underline{\Psi}, \underline{V})_3 + (\Phi_1, W_1) = (\Phi, W).$$

Hence, we have proved Lemma 3 and Proposition 1. □

5. Proof of Proposition 2 (low frequencies). In low frequencies of the spectrum, Ingham techniques fail because the gap $\nu_{n,k+1} - \nu_{n,k}$ goes to 0. In order to get a sharp estimate, we will use biorthogonal sequences.

The idea is to build functions $h_{n,k}$ for which $\widehat{h}_{n,k}(\nu_{n,k_0}) = \delta_k^{k_0}$, supported by $[-T, T]$ and such that their L^2 norm is bounded from above.

The usual technique is based upon the functions

$$\prod_k \Big(1 - \frac{z^2}{\nu^2_{n,k}}\Big).$$

We will build that kind of function (which we will call functions f) by more explicit means in section 5.1.

These functions have a problem, though: Their L^2 norms behave like $e^{\pi|n|}$ and, therefore, are not bounded. Nevertheless, the zone where this L^2 norm is big is concentrated in the frequency interval $[-n, n]$ that includes only two eigenfrequencies: $\nu_{n,*}$ and $\nu_{n,**}$. Thus, in section 5.2 we will build a sequence of functions $g_{n,k}$ such that

$||\widehat{g}_{n,k}||_{L^2}$ is exponentially small in $[-n,n]$ and bounded from below outside. Then we will put $h=f*g$ in order to compensate the size of f without hindering its properties at frequencies $\nu_{n,k}$.

Indeed, we will prove the following lemma.

LEMMA 4. *For any odd integer q and any positive real number ϵ, there exists a time $T_2(q,\epsilon)$ smaller than $C_q\epsilon^{\frac{q+1}{1-q}}$ such that we can find a biorthogonal sequence $(h_{\epsilon,q}^{k_0,n})_{(n,k_0)\in\mathbb{N}^*\times(\mathbb{Z}^*\cup\{*,**\})}$ of L^2 functions for which the following properties hold:*

(i) *$h_{\epsilon,q}^{k_0,n}$ is supported by $[-T_2(q,\epsilon),T_2(q,\epsilon)]$;*

(ii) *$||h_{\epsilon,q}^{k_0,n}||_{L^2}^2\leq C\ e^{2\epsilon|n|}$;*

(iii) *if $k\neq\pm k_0$, $\int h_{\epsilon,q}^{k_0,n}(t)e^{t\nu_{n,k}}\ dt=0$;*

(iv) *if $n\geq n_1(\epsilon,q)$ and $(k_0,n)\in I=\{(k,n)\mid|k|=*\text{ or }|k|\leq n\}$,*
$|\int h_{\epsilon,q}^{k_0,n}(t)e^{t\nu_{n,\pm k_0}}\ dt\ |\geq\frac{c}{n^{N_q}}$.

The constants depend only on q and ϵ. Moreover, the functions h can be chosen as even or odd. We will denote them $h_{e_{\epsilon,q}}^{k_0,n}$ or $h_{o_{\epsilon,q}}^{k_0,n}$.

Let us first prove Proposition 2 from this lemma.

Let n be an integer greater than $n_1(\epsilon)$ and $(\underline{\Psi},\underline{V})$ an initial condition in $\mathcal{X}^n$. Let us denote (Ψ,V) the solution of (2) with these initial data. We will denote K as the operator that maps $(\underline{\Psi},\underline{V})$ in $\mathcal{X}^n$ to $V_{tt}(y=0,.)$.

If we denote $a_{n,k}=\langle(\underline{\Psi},\underline{V}),\frac{\xi_{n,k}}{\Xi_{n,k}}\rangle$, we notice that

$$V(y,t)=\sum_{k\in\mathbb{Z}^*\cup\{*,**\}}a_{n,k}\ \frac{\xi_{n,k}^3}{\Xi_{n,k}}\cos\pi ny\ e^{\nu_{n,k}t}.$$

Thus,

$$K(\underline{\Psi},\underline{V})(t)=\sum_{k\in\mathbb{Z}^*\cup\{*,**\}}a_{n,k}\ \frac{\xi_{n,k}^3}{\Xi_{n,k}}\nu_{n,k}^2e^{\nu_{n,k}t}.$$

Now for (k_0,n) in I and L in $\mathbb{N}^*$, as $\widehat{h_e}$ is even,

$$\int h_{e_{\epsilon,q}}^{k_0,n}(t)K\Big(\sum_{\substack{|k|=*\\|k|\leq L}}a_{n,k}\ \frac{\xi_{n,k}}{\Xi_{n,k}}\Big)(t)\ dt=\sum_{\substack{k=*\\1\leq k\leq L}}(a_{n,k}+a_{n,-k})\frac{\xi_{n,k}^3}{\Xi_{n,k}}\nu_{n,k}^2\int h_{e_{\epsilon,q}}^{k_0,n}(t)e^{\nu_{n,k}t}\ dt.$$

From (iii), if $L\geq k_0$,

$$\int h_{e_{\epsilon,q}}^{k_0,n}(t)K\Big(\sum_{\substack{|k|=*\\|k|\leq L}}a_{n,k}\ \frac{\xi_{n,k}}{\Xi_{n,k}}\Big)(t)\ dt=(a_{n,k_0}+a_{n,-k_0})\frac{\xi_{n,k_0}^3}{\Xi_{n,k_0}}\nu_{n,k_0}^2\int h_{e_{\epsilon,q}}^{k_0,n}(t)e^{\nu_{n,k_0}t}\ dt.$$

From (iv), we get that

$$\left|\int h_{e_{\epsilon,q}}^{k_0,n}(t)K\Big(\sum_{\substack{|k|=*\\|k|\leq L}}a_{n,k}\frac{\xi_{n,k}}{\Xi_{n,k}}\Big)(t)\ dt\right|\geq|a_{n,k_0}+a_{n,-k_0}|\ \left|\frac{\xi_{n,k_0}^3}{\Xi_{n,k_0}}\right|\ |\nu_{n,k_0}|^2\ \frac{c}{n^{N_q}}.$$

Now Lemma 1 implies that

$$|\xi_{n,k_0}^3|=\frac{|z_{n,k_0}|}{|\nu_{n,k_0}|}|\sin z_{n,k_0}|\geq\frac{C_\gamma}{n^N}.$$

Thus,

$$\left|\int h_{e\epsilon,q}^{k_0,n}(t)K\Big(\sum_{\substack{|k|=*\\|k|\leq L}} a_{n,k}\frac{\xi_{n,k}}{\Xi_{n,k}}\Big)(t)\ dt\right| \geq |a_{n,k_0}+a_{n,-k_0}|\ C\ e^{-Cn^{\frac{2}{3}}}.$$

If we take the limit with $L \to +\infty$,

$$\left|\int h_{e\epsilon,q}^{k_0,n}(t)K(\underline{\Psi},\underline{V})(t)\ dt\right| \geq |a_{n,k_0}+a_{n,-k_0}|\ C\ e^{-Cn^{\frac{2}{3}}}.$$

We can prove the same way that

$$\left|\int h_{o\epsilon,q}^{k_0,n}(t)K(\underline{\Psi},\underline{V})(t)\ dt\right| \geq |a_{n,k_0}-a_{n,-k_0}|\ C\ e^{-Cn^{\frac{2}{3}}}.$$

Thus, by summing conveniently,

$$(13)\quad |a_{n,k_0}| \leq C\ e^{Cn^{\frac{2}{3}}}\left[\left|\int h_{e\epsilon,q}^{k_0,n}(t)K(\underline{\Psi},\underline{V})(t)\ dt\right| + \left|\int h_{o\epsilon,q}^{k_0,n}(t)K(\underline{\Psi},\underline{V})(t)\ dt\right|\right].$$

For any n greater than $n_1(\epsilon)$

$$||(\underline{\Psi},\underline{V})^{(1)}||_{\mathcal{X}}^2 = \sum_{\substack{|k|=*\\|k|\leq|n|}} |a_{n,k}|^2.$$

From (13)

$$||(\underline{\Psi},\underline{V})^{(1)}||_{\mathcal{X}}^2 \leq C \sum_{\substack{|k|=*\\|k|\leq|n|}} e^{Cn^{\frac{2}{3}}}\ \left|\int h_{e\epsilon,q}^{k,n}(t)\ V_{tt}(0,t)\ dt\right|^2 + \text{same with } h_o.$$

Hence, from (i)

$$||(\underline{\Psi},\underline{V})^{(1)}||_{\mathcal{X}}^2 \leq Ce^{Cn^{\frac{2}{3}}} \sum_{\substack{|k|=*\\|k|\leq|n|}} \int \left|h_{\epsilon,q}^{k,n}(t)\right|^2 dt \int_{-T_2(q,\epsilon)}^{T_2(q,\epsilon)} |V_{tt}(0,t)|^2\,dt.$$

Thus, from (ii)

$$||(\underline{\Psi},\underline{V})^{(1)}||_{\mathcal{X}}^2 \leq C\ e^{Cn^{\frac{2}{3}}}\ e^{2\epsilon|n|} \int_{-T_2(q,\epsilon)}^{T_2(q,\epsilon)} |V_{tt}(0,t)|^2\,dt.$$

When q goes to infinity, if $\frac{q+1}{1-q} = -1-\delta$, δ goes to 0. Thus, we have proved Proposition 2. □

We still have to prove Lemma 4. We will proceed by two steps.

First, we will introduce a sequence of functions $f^{k_0,n}$ that will satisfy conditions (i), (iii), and (iv) but whose L^2 norms will behave like $e^{n\pi}$, which is too big for (ii). We will notice, however, that these norms will be mostly concentrated within $[-\pi n, \pi n]$ on the Fourier side.

Then we will build a sequence of functions g of which we will know, by stationary phases computations, that their norms on the Fourier side are exponentially small over $[-\pi n, \pi n]$ and reasonably bounded outside.

We will then put $h = f * g$ and prove that h satisfies (i) to (iv) for suitable parameters.

5.1. Construction of functions $\boldsymbol{f}$. Let $f_0^n(z) = (\sqrt{z^2-n^2\pi^2}^3 \tan\sqrt{z^2-n^2\pi^2} - z^2)\cos\sqrt{z^2-n^2\pi^2}$.

The following properties hold:

(f-i) $f_0^n \in \mathcal{O}(\mathbb{C})$.

(f-iii) From (8), for any k in $\mathbb{Z}^* \cup \{*,**\}$,

$$f_0^n(\pm|\nu_{n,k}|) = (z_{n,k}^3 \tan z_{n,k} - z_{n,k}^2 - n^2\pi^2)\cos z_{n,k} = 0.$$

Let us evaluate $f_0^{n\prime}(|\nu_{n,k}|)$. We shall write $F(z) = z^3\tan z - z^2 - n^2\pi^2$ and $G(z) = \sqrt{z^2-n^2\pi^2}$. We have

$$f_0^n(z) = F(G(z)) \cdot \cos G(z) \text{ and } G(|\nu_{n,k}|) = z_{n,k}.$$

So

$$f_0^{n\prime}(z) = G'(z)\cdot F'(G(z))\cdot\cos G(z) - F(G(z))\cdot G'(z)\cdot\sin G(z).$$

Thus,

$$|f_0^{n\prime}(|\nu_{n,k}|)| = |G'(|\nu_{n,k}|)\cdot\cos z_{n,k}|\ |F'(z_{n,k})|. \tag{14}$$

Now

$$F'(z) = 3z^2\tan z + z^3(1+\tan^2 z) - 2z$$

and

$$\tan z_{n,k} = \frac{z_{n,k}^2 + n^2\pi^2}{z_{n,k}^3}.$$

So for any (n,k) in $\mathbb{N}^* \times (\mathbb{N}^* \cup \{*\})$,

$$F'(z_{n,k}) = 2z_{n,k} + \frac{5\pi^2 n^2}{z_{n,k}} + z_{n,k}^3 + \frac{\pi^4 n^4}{z_{n,k}^3} \geq 2z_{n,k} + z_{n,k}^3 \geq 1.$$

Moreover,

$$|G'(|\nu_{n,k}|)| = \frac{|\nu_{n,k}|}{z_{n,k}} \geq \frac{1}{2}.$$

At last, from Lemma 1 (ii),

$$|\cos z_{n,k}| \geq \frac{1}{5}|z_{n,k} - \left(|k| - \frac{1}{2}\right)\pi| \geq \frac{C}{1+n^2}.$$

Thus, from (14), we get that for any couple (n,k) in $\mathbb{N}^* \times (\mathbb{Z}^* \cup \{*,**\})$,

$$|f_0^{n\prime}(|\nu_{n,k}|)| \geq \frac{C}{1+n^2}. \tag{15}$$

Let us put for any k in $\mathbb{Z}^* \cup \{*,**\}$

$$f_0^{k,n}(z) = f_0^n(z)\ \frac{1}{z^2 - |\nu_{n,k}|^2}\ \left(\frac{\sin\sqrt{z^2-\pi^2n^2}}{\sqrt{z^2-\pi^2n^2}}\right)^2.$$

(The last term ensures that f remains in L^2.)

We have kept properties (f-i) and (f-iii):

(f-i) $f_0^{k,n} \in \mathcal{O}(\mathbb{C})$. Moreover, $f_0^{k,n} \in L^2(\mathbb{R})$ and, for any complex number z, $|f_0^{k,n}(z)| \leq C\ e^{3|\Im m\ z|}$.

(f-iii) for any k in $(\mathbb{Z}^* \cup \{*, **\})\backslash\{\pm k_0\}$, $f_0^{k_0,n}(\pm|\nu_{n,k}|) = 0$.

We have one more property; from the upper bound we got for the derivative at the pole

$$f_0^{k,n}(\pm|\nu_{n,k}|) = f_0^{n\prime}(\pm|\nu_{n,k}|)\Big(\frac{\sin z_{n,k}}{z_{n,k}}\Big)^2 \frac{1}{\mp 2|\nu_{n,k}|}.$$

Thus,

$$|f_0^{k,n}(\pm|\nu_{n,k}|)| \geq \frac{C}{(1+n^2)|k|^3}|\sin z_{n,k}|^2.$$

Therefore, from Lemma 1 (i) and (iii),
(f-iv) $f_0^{k,n}(\pm|\nu_{n,k}|) \geq \frac{C}{(1+n^2)|k|^{N_2}}$.

Unfortunately, $||f^{k,n}||_{L^2} \geq C\ e^n$ so (ii) doesn't hold for functions f. However, the following estimate holds for any integers k, n and $z \in [-\pi n, \pi n]$:

$$|f^{k,n}(z)| \leq Ce^{3n\sqrt{\pi^2-(\frac{z}{n})^2}}. \tag{16}$$

5.2. Introduction of functions *g*. Let q be an odd integer and let us denote $h_q(x)$ the solution of $y' = 1 + y^{q-1}$ for which $y(0) = 0$. This function is defined over $(-x_q, x_q)$ for a positive x_q. It is odd, strictly increasing and analytic. Moreover, we have $h_q(x) = x + \alpha_q x^q + o(x^q)$ when x is close to 0, with a positive α_q, and when x goes to x_q, h_q goes to infinity.

We shall denote by H_q the reciprocal function to h_q. It is defined over $\mathbb{R}$, odd, strictly increasing, and bounded by x_q. We have $H_q(x) = x - \alpha_q x^q + o(x^q)$ if x is close to 0. Let δ be a real number, greater than 1 and close to 1, that will be fixed later.

Let us introduce functions g as follows:

$$g_{+T,q}^{\ n}(t) = \mathbf{1}_{(-T,T)} e^{in\frac{T}{\delta x_q}h_q(\frac{x_q}{T}t)}.$$

Then

$$\widehat{g_{+T,q}^{\ n}}(\tau) = \int_{-T}^{T} e^{in\frac{T}{\delta x_q}h_q(\frac{x_q}{T}t)-i\tau t}\ dt.$$

If we put

$$\Psi_q(s) = \frac{T}{x_q}H_q(\frac{\delta x_q}{T}s),$$

we can write

$$\widehat{g_{+T,q}^{\ n}}(\tau) = \int_{-\infty}^{+\infty} e^{ins-i\tau\Psi_q(s)}\ \Psi_q'(s)\ ds.$$

Let us also denote

$$\theta_q(s) = \frac{1}{x_q}H_q(\delta x_q s),$$

then

$$\begin{aligned}\widehat{g_{+}{}^{n}_{T,q}}(\tau) &= \int_{-\infty}^{+\infty} \theta_q'\Big(\frac{s}{T}\Big) e^{inT\left(\frac{s}{T}-\frac{\tau}{n}\theta_q(\frac{s}{T})\right)}\, ds,\\ &= T\int_{-\infty}^{+\infty} \theta_q'(v) e^{inT\left(v-\frac{\tau}{n}\theta_q(v)\right)}\, dv.\end{aligned}$$

We will eventually prove the following lemma about functions $g_{+}{}^{n}_{T,\delta}$.

LEMMA 5. *For T big enough, there exist three positive constants $C_q^1, C_{q,T}^2, c_{q,T,\delta}^3$ and two integers $r_q, n(q,\delta)$ such that*

(i) *for any positive integer n and any real number τ smaller than $\frac{n}{\delta}$,*

$$\left|\widehat{g_{+}{}^{n}_{T,q}}(\tau)\right| \le C_{q,T}^2\ e^{-Tn\ C_q^1 \min\{(\frac{1}{\delta}-\frac{\tau}{n})^{\frac{q}{q-1}},1\}};$$

(ii) *for any integer n greater than $n(q,\delta)$, if $k_0 = *$ or $1 \le k_0 \le n$, there is a time T_{n,k_0} in $[T, T+1]$ such that*

$$\left|\widehat{g_{+}{}^{n}_{T_{n,k_0},q}}\left(\frac{|\nu_{n,k_0}|}{\pi}\right)\right| \ge \frac{c_{q,T,\delta}^3}{\sqrt{n}}.$$

Let us see how this lemma allows us to construct a sequence of functions h for which the properties of Lemma 4 hold.

5.2.1. Construction of *h*. First, let us notice that we can prove a lemma that is similar to Lemma 5 for functions $g_{-}{}^{n}_{T,q}(t) = \mathbf{1}_{(-T,T)} e^{-in\frac{T}{\delta x_q}h_q(\frac{x_q}{T}t)}$ by changing t into $-t$.

As $g_{-}{}^{n}_{T,q} = \overline{g_{+}{}^{n}_{T,q}}$, we have $T_{n,k_0,+} = T_{n,k_0,-}$. Thus, if we put

$$g_{e}{}^{n}_{T,q}(t) = \mathbf{1}_{(-T,T)} \cos\left(n\frac{T}{\delta x_q} h_q\left(\frac{x_q}{T}t\right)\right) = \Re e\ (g_{+}{}^{n}_{T,q}) = \frac{1}{2}(g_{+}{}^{n}_{T,q}(t) + g_{-}{}^{n}_{T,q}(t)),$$

the following inequalities hold.

If $|\frac{\tau}{n}| \le \frac{1}{\delta}$,

$$\left|\widehat{g_{e}}{}^{n}_{T,q}(\tau)\right| \le C_{q,T}\ e^{-Tn\ C_q(\frac{1}{\delta}-|\frac{\tau}{n}|)^{\frac{q}{q-1}}}. \tag{17}$$

And if $n \ge n(q,\delta)$ and ($|k_0| \le n$ or $|k_0| = *$), as $C_{q,T}^2\ e^{-nT_uC_q^1} \le \frac{c_{q,T,\delta}^3}{2\sqrt{n}}$ if n is big enough,

$$\left|\widehat{g_{\pm}{}^{n}_{T_{n,k_0},q}}(\tau)\right| \le \frac{1}{2}\left|\widehat{g_{\mp}{}^{n}_{T_{n,k_0},q}}(\tau)\right| \text{ for } \tau = \mp\frac{|\nu_{n,k_0}|}{\pi}.$$

As we can increase C to cope with the finite number of (n,k) in I for which n is not big enough, we get for (n,k_0) in $I, \tau = \pm\frac{|\nu_{n,k_0}|}{\pi}$

$$\left|\widehat{g}_{e}{}_{T_{n,k_0},q}(\tau)\right| \ge \frac{c_{q,T,\delta}}{\sqrt{n}}. \tag{18}$$

Thus, we have the following lemma for g_e.

LEMMA 6. *For T big enough, there exist three positive constants $C_q^1, C_{q,T}^2, c_{q,T,\delta}^3$ and two integers $r_q, n(q,\delta)$ such that*

(i) *for any positive integer* n *and any real number* τ *in* $[-\frac{n}{\delta},\frac{n}{\delta}]$,

$$\left|\widehat{g_e}^{\,n}_{T,q}(\tau)\right| \le C^2_{q,T}\ e^{-Tn\ C^1_q\left(\frac{1}{\delta}-|\frac{\tau}{n}|\right)^{\frac{q}{q-1}}};$$

(ii) *for any integer* n *greater than* $n(q,\delta)$, *if* $|k_0| = *$ *or* $1 \le |k_0| \le n$, *there is a time* T_{n,k_0} *in* $[T, T+1]$ *such that*

$$\left|\widehat{g_e}^{\,n}_{T_{n,k_0},q}\left(\frac{|\nu_{n,k_0}|}{\pi}\right)\right| \ge \frac{c^3_{q,T,\delta}}{\sqrt{n}}.$$

Of course, we get the same kind of results for

$$g_o{}^{n}_{T,q}(t) = \mathbf{1}_{(-T,T)}\sin\left(n\frac{T}{\delta x_q}h_q\left(\frac{x_q}{T}t\right)\right).$$

Obviously, g_e is even and g_o is odd.

We shall now define the functions h. Let ϵ be a positive real number. Pick δ_ϵ such that

$$3\pi\sqrt{1-\left(\frac{1}{\delta_\epsilon}\right)^2} = \frac{\epsilon}{2}$$

and T^ϵ such that

$$\sup_{\beta\in[0,\frac{1}{\delta_\epsilon}]}\left(3\pi\sqrt{1-\beta^2} - C^1_q\ T^\epsilon\ \left(\frac{1}{\delta_\epsilon}-\beta\right)^{\frac{q}{q-1}}\right) \le \epsilon. \tag{19}$$

Indeed, the derivative of the function under the sup is

$$\frac{-3\pi\beta}{\sqrt{1-\beta^2}} + \frac{q}{q-1}T^\epsilon C^1_q\left(\frac{1}{\delta_\epsilon}-\beta\right)^{\frac{1}{q-1}},$$

so we get the right estimate if we choose T^ϵ such that this derivative is 0 for β_ϵ such that

$$3\pi\sqrt{1-\beta_\epsilon^2} = \epsilon.$$

We have $\delta_\epsilon = 1 + \frac{\epsilon^2}{72\pi^2} + o(\epsilon^2)$, $\beta_\epsilon = 1 - \frac{\epsilon^2}{18\pi^2} + o(\epsilon^2)$ so $\frac{1}{\delta_\epsilon} - \beta_\epsilon \sim \frac{\epsilon^2}{24\pi^2}$; hence,

$$T^\epsilon \sim C_q\epsilon^{\frac{q+1}{1-q}}.$$

Let us define positive times T^ϵ_{n,k_0} as follows.

For integers k_0 such that $|k_0| \le |n|$ or $|k_0| = *$, we take the time T^ϵ_{n,k_0} given by Lemma 6 with $T = T^\epsilon$, and for $|k_0| > |n|$, we put $T^\epsilon_{n,k_0} = T^\alpha$.

$$T^\epsilon_{n,k_0} \in [T^\epsilon, T^\epsilon + 1],\ \text{so}\ c^1_q\epsilon^{\frac{q+1}{1-q}} \le T^\epsilon_{n,k_0} \le c^2_q\epsilon^{\frac{q+1}{1-q}}.$$

Let us denote

$$\widehat{h_{e_{\epsilon,q}}}^{k_0,n}(\tau) = f^{k_0,n}(\tau)\cdot\ \widehat{g_e}^{\,n}_{T^\epsilon_{n,k_0},q}\left(\frac{\tau}{\pi}\right),$$

$$\widehat{h_{o\epsilon,q}}^{k_0,n}(\tau) = f^{k_0,n}(\tau) \cdot \ \widehat{g}_{oT^{\epsilon}_{n,k_0},q}^{\ n}\left(\frac{\tau}{\pi}\right).$$

The subscript means that h is even or odd. We will drop this subscript when it is not required.

We shall now prove step by step that $h_{\epsilon,q}^{k_0,n}$ satisfies all the properties of Lemma 4.

• Proof of (i): $h_{\epsilon,q}^{k_0,n}$ is a convolution product of $\widehat{f}^{k_0,n}$, whose support is located within $[-3\pi, 3\pi]$ and $g^n_{T^{\epsilon}_{n,k_0},q}$ that is supported by $[-T^{\epsilon}_{n,k_0}, T^{\epsilon}_{n,k_0}]$. So we can put $T_2(q,\epsilon) = 3\pi + c_q^2 \epsilon^{\frac{q+1}{1-q}}$ so that $h_{\epsilon,q}^{k_0,n}$ is supported by $[-T_2(q,\epsilon), T_2(q,\epsilon)]$. The estimate on T^{ϵ}_{n,k_0} ensures that $T_2(q,\epsilon) \leq C_q \epsilon^{\frac{q+1}{1-q}}$.

• Proof of (ii): We will use results about the small size of $||g||$ that will compensate $||f||$. Outside of $[-\pi n, \pi n]$, the L^2 norm of f is bounded by a polynomial in n and $|\widehat{g}|_{L^\infty}$ is bounded by $2T_2(q,\epsilon)$, so the problems are located within this interval.

We must estimate $\int_{-n}^{n} |\widehat{h}_{\epsilon,q}^{k_0,n}(\tau)|^2 \ d\tau$.

Now, by (16) we know that if $\frac{\tau}{n}$ belongs to $[-\pi, \pi]$, we have

$$|f^{k_0,n}(\tau)|^2 \leq C \ e^{6n\sqrt{\pi^2 - |\frac{\tau}{n}|^2}} = C \ e^{6\pi n\sqrt{1-|\frac{\tau}{\pi n}|^2}}.$$

Thus, if $|\frac{\tau}{\pi n}| \geq \frac{1}{\delta_\epsilon}$,

$$|\widehat{h}_{\epsilon,q}^{k_0,n}(\tau)|^2 \leq Ce^{\epsilon n}.$$

Moreover, by Lemma 6, if $|\frac{\tau}{\pi n}|$ is smaller than $\frac{1}{\delta_\epsilon}$, we have

$$|\widehat{g}^n_{T^{\epsilon}_{n,k_0},q}\left(\frac{\tau}{\pi}\right)|^2 \leq C \ e^{-2T^{\epsilon}_{n,k_0} n C_q^1 (\frac{1}{\delta_\epsilon} - |\frac{\tau}{\pi n}|)^{\frac{q}{q-1}}}.$$

Therefore, from (19), we get that if $|\frac{\tau}{\pi n}|$,

$$|\widehat{h}_{\epsilon,q}^{k_0,n}(\tau)|^2 \leq C \ e^{2\epsilon n}.$$

Thus,

$$||\widehat{h}_{\epsilon,q}^{k_0,n}||^2_{L^2} \leq Ce^{2\epsilon n}.$$

• Proof of (iii): This is a simple consequence of property (f-iii) for functions f. Indeed, for any integer k different from k_0, $f^{k_0,n}(|\nu_{n,k}|) = 0$. Thus, by definition of h we also have $\widehat{h}_{\epsilon,q}^{k_0,n}(|\nu_{n,k}|) = 0$, which is exactly the Fourier transcription of (iii).

• Proof of (iv): For any couple (n, k_0) in I, from (f-iv) and (18) we get

$$\left|\widehat{h}_{\epsilon,q}^{k_0,n}(\pm|\nu_{n,k_0}|)\right| \geq \frac{C}{n^N}\frac{c_{q,T_\epsilon,\delta_\epsilon}}{\sqrt{n}} \geq \frac{C_{q,\epsilon}}{n^{N'}}.$$

This again is the Fourier transcription of the needed result.

We shall now prove Lemma 5 to end our demonstration.

5.2.2. Proof of Lemma 5. Let us keep in mind that we have put

$$\widehat{g_{+T,q}^{n}}(\tau) = T\int_{-\infty}^{\infty} \theta_q'(v) e^{inT(v - \frac{\tau}{n}\theta_q(v))} \ dv.$$

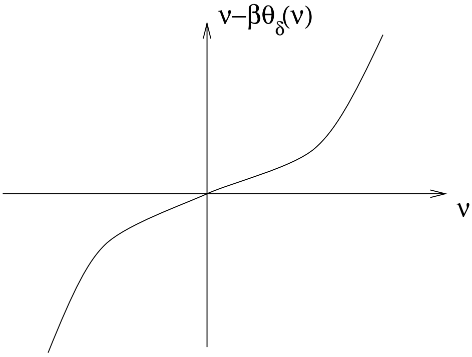
Let us put $\alpha = nT$ and $\beta = \frac{\tau}{n}$. We will estimate

$$\phi(\alpha, \beta) = \int \theta'_\delta(v) e^{i\alpha\left[v - \beta\theta_\delta(v)\right]}\, dv,$$

when α goes to infinity.

There will be two kinds of estimates depending upon the value of β as compared to $\frac{1}{\delta}$: $\beta < \frac{1}{\delta}$ and $\beta \geq 1 (> \frac{1}{\delta})$.

If $\beta < \frac{1}{\delta}$,



In this zone, the phase is nonstationary. Thus, we will get an exponential decrease.

Let us shift slightly in the imaginary direction. For any real number v, any β smaller than $\frac{1}{\delta}$, and any small ϵ we get

$$\begin{aligned}
\Im m\ \Big(v + i\epsilon - \beta\theta_q(v + i\epsilon)\Big) &= \epsilon - \beta\ \Im m\ \theta_q(v + i\epsilon),\\
&= \epsilon - \beta\ \Im m\ \Big(\theta_q(v + i\epsilon) - \theta_q(v)\Big),\\
&= \epsilon - \beta\ \Im m\ \int_v^{v+i\epsilon} \theta'_q(z)\, dz,\\
&= \epsilon - \beta\ \Im m\ \int_v^{v+i\epsilon} \frac{\delta dz}{1 + \delta^{q-1} x_q^{q-1} z^{q-1}},\\
&= \epsilon - \beta\epsilon\delta\ \Re e\ \int_0^1 \frac{du}{1 + \delta^{q-1} x_q^{q-1} (v + i\epsilon u)^{q-1}}.
\end{aligned}$$

Thus, if $\beta \leq 0$,

$$\Im m\ \Big(v + i\epsilon - \beta\theta_q(v + i\epsilon)\Big) \geq \epsilon \text{ if } \beta \leq 0.$$

If β is positive,

$$\Im m\ \Big(v + i\epsilon - \beta\theta_q(v + i\epsilon)\Big) \geq \epsilon - \beta\epsilon\delta \left| \int_0^1 \frac{du}{1 + \delta^{q-1} x_q^{q-1} (v + i\epsilon u)^{q-1}} \right|.$$

Now for any real number v,

$$\underbrace{\left| \int_0^1 \frac{du}{1 + \delta^{q-1} x_q^{q-1} (v + i\epsilon u)^{q-1}} \right|}_{I} \leq \frac{1}{1 - c_q \epsilon^{q-1}},$$

because either $v \gg \epsilon$ and then $I \leq \frac{c}{1+v^{q-1}} \leq 1$ or $v \leq M_q \epsilon$ and then

$$|v + i\epsilon u|^{q-1} \leq C_q \epsilon^{q-1} \Rightarrow |1 + \delta^{q-1} x_q^{q-1} (v + i\epsilon u)^{q-1}| \geq 1 - c_q \epsilon^{q-1} \Rightarrow I \leq \frac{1}{1 - c_q \epsilon^{q-1}}.$$

Thus,

$$\begin{aligned} \Im m \, \Big(v + i\epsilon - \beta \theta_q (v + i\epsilon) \Big) &\geq \epsilon - \frac{\beta \delta \epsilon}{1 - c_q \epsilon^{q-1}}, \\ &\geq \epsilon (1 - \delta \beta) - c_q' \beta \epsilon^q, \\ &\geq \epsilon (\frac{1}{\delta} - \beta) - c_q' \beta \epsilon^q. \end{aligned}$$

Now

$$\begin{aligned} \max_{\epsilon} \epsilon \left(\frac{1}{\delta} - \beta \right) - c_q \beta \epsilon^q &= c_q' \left(\frac{1}{\delta} - \beta \right)^{\frac{q}{q-1}} \beta^{\frac{1}{1-q}}, \\ &\geq c_q'' \left(\frac{1}{\delta} - \beta \right)^{\frac{q}{q-1}}. \end{aligned}$$

We can choose a real number ϵ and a very small c_q such that for any real number v

$$\begin{cases} \Im m \, \Big(v + i\epsilon - \beta \theta_q (v + i\epsilon) \Big) \geq c_q^{te} (\frac{1}{\delta} - \beta)^{\frac{q}{q-1}} \text{ if } \beta \in (0, \frac{1}{\delta}], \\ \Im m \, \Big(v + i\epsilon - \beta \theta_q (v + i\epsilon) \Big) \geq c_q^{te} \text{ if } \beta \leq 0. \end{cases}$$

Now we can shift the integration line over v from $\mathbb{R}$ to $\mathbb{R} + i\epsilon$. Then

$$\phi(\alpha, \beta) = \int \theta_q' (v + i\epsilon) e^{i\alpha \left[v + i\epsilon - \beta \theta_q (v + i\epsilon) \right]} \, dv.$$

To conclude, as

$$\theta_q' (v + i\epsilon) = \frac{\delta}{1 + (\delta x_q (v + i\epsilon))^q - 1},$$

we get

$$|\theta_q' (v + i\epsilon)| \leq \frac{C_q}{1 + v^{q-1}}.$$

Hence, for any real number α and any $\beta \leq \frac{1}{\delta}$,

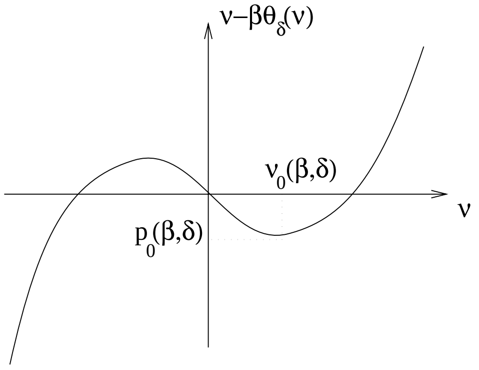
$$|\phi(\alpha, \beta)| \leq \int \frac{C_q}{1 + v^{q-1}} \, e^{-\alpha \ c_q \min \left\{ (\frac{1}{\delta} - \beta)^{\frac{q}{q-1}}, 1 \right\}} \, dv \leq C_q \ e^{-\alpha \ c_q \min \left\{ (\frac{1}{\delta} - \beta)^{\frac{q}{q-1}}, 1 \right\}}.$$

Thus, if $\frac{\tau}{n} \leq \frac{1}{\delta}$,

$$\left| \widehat{g_{+T,q}^{n}} (\tau) \right| \leq C_q \ T \ e^{-nTc_q \min \left\{ (\frac{1}{\delta} - \frac{\tau}{n})^{\frac{q}{q-1}}, 1 \right\}}. \tag{20}$$

That is the first part of Lemma 5.

If $\beta \geq 1 \ \ (> \frac{1}{\delta})$,



.

Through the stationary phase formula, we get

$$\phi(\alpha,\beta) = \Big(|H_{\beta,\delta}| \ \cos \alpha p_0(\beta,\delta)\Big) \left[\frac{\theta_q'(v_0(\beta,\delta))}{\sqrt{\alpha}} + \sum_{j=1}^{N} \frac{a_j(\beta,\delta)}{\alpha^j \sqrt{\alpha}} \right] + r_{\beta,\delta}(\alpha),$$

with $r_{\beta,\delta}(\alpha) \leq \frac{C_\beta}{\alpha^{N+1}}$ and $\alpha \geq A_{\beta,\delta}$. $H_{\beta,\delta}$ is the square root of the Hessian at the critical points.

Moreover, in this formula, C and A are continuous with respect to β and δ, and $a_j(\beta,\delta)$ depends on the first $2j+1$ derivatives of $v \mapsto \theta_q(v)$ at $v = v_0(\beta,\delta)$.

Let us compute $p_0(\beta,\delta)$:

$$\begin{aligned} \frac{\partial}{\partial v}\Big(v - \beta\theta_q(v)\Big) = 0 \Leftrightarrow 1 - \frac{\beta\delta}{1 + \delta^{q-1}x_q^{q-1}v^{q-1}} &= 0, \\ \Rightarrow 1 + \delta^{q-1}x_q^{q-1}v_0^{q-1}(\beta,\delta) &= \beta\delta, \\ \Rightarrow v_0(\beta,\delta) &= \frac{1}{\delta x_q}(\delta\beta - 1)^{\frac{1}{q-1}}. \end{aligned}$$

If β takes the values $\frac{|\nu_{n,k}|}{n\pi}$ for any couple (n,k) such that $|k| \leq n$, we have $1 \leq \beta \leq \pi\sqrt{2}$ through point (i) of Lemma 1.

Moreover, if $\beta = \frac{|\nu_{n,*}|}{n\pi}$, by point (iv) of Lemma 1 $\beta \geq 1 - \frac{C}{\sqrt{n}} \geq \frac{1}{2}(1 + \frac{1}{\delta})$ as soon as $n \geq n_0(\delta)$.

Therefore, for any n greater than $n_0(\delta)$, if (n,k) belongs to I and $\beta = \frac{|\nu_{n,k}|}{n\pi}$,

$$C \geq v_0(\beta,\delta), |p_0(\beta,\delta)|, |H_{\beta,\delta}| \geq c_\delta;$$

thus,

$$1 \geq \theta_q'(v_0(\beta,\delta)) \geq c_q.$$

Moreover,

$$a_j(\beta,\delta) \leq C_{j,\delta}.$$

Let T be a positive real time. As $|p_0(\beta,\delta)| \geq c_\delta$, for any n greater than $n_0(\delta)$ and k_0 such that (n,k_0) belongs to I, one can pick a time T_{n,k_0} in $[T, T+1]$ such that

$$\cos\left(nT_{n,k_0}p_0\left(\frac{|\nu_{n,k_0}|}{n\pi},\delta\right)\right) \geq c_\delta'.$$

Thus, for $T > T_u,\ n \geq n(q,\delta),\ \alpha = Tn,\ (k_0, n) \in I$, and $\beta = \frac{|\nu_{n,k_0}|}{n\pi}$,

$$\left| \frac{\theta_q'(v_0(\beta,\delta))}{\sqrt{\alpha}} + \sum_{j=1}^{N} \frac{a_j(\beta,\delta)}{\alpha^j \sqrt{\alpha}} \right| \geq \frac{|\theta_q'(v_0(\beta,\delta))|}{2\sqrt{\alpha}},$$

$$|r_{\beta,\delta}(\alpha)| \leq c_\delta' \frac{|H_{\beta,\delta}|\ \theta_q'(v_0(\beta,\delta))}{4\sqrt{\alpha}}.$$

In the same conditions, there is a time T_{n,k_0} in $[T, T+1]$ such that

$$\left| \phi\left(nT_{n,k_0}, \frac{|\nu_{n,k_0}|}{n\pi} \right) \right| \geq \frac{c_\delta' |H|\ \theta_q'(v_0(\frac{\nu_{n,k_0}}{n\pi},\delta))}{4\sqrt{n}\sqrt{T_{n,k_0}}} \geq \frac{C}{\sqrt{n}}.$$

We have proved that for any time T greater than T_u, for any n bigger than $n(q,\delta)$ and k_0 such that $|k_0| = *$ or $|k_0| \leq n$, there is a time T_{n,k_0} in $[T, T+1]$ such that

$$\left| \widehat{g^n_{+T_{n,k_0},q}}\left(\frac{|\nu_{n,k_0}|}{\pi} \right) \right| \geq \frac{C_{T,q,\delta}}{\sqrt{n}}. \tag{21}$$

This proves the second part of Lemma 5 and ends the proof of Proposition 2. □

REFERENCES

[BFSS] H. T. Banks, W. Fang, R. J. Silcox, and R. C. Smith, *Approximation methods for control of acoustic/structure models with piezoceramic actuators*, J. Intell. Material Systems Structures, 4 (1993), pp. 98–116.

[BLR] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[Ha] A. Haraux, *Séries lacunaires et contrôle semi-interne des vibrations d'une plaque rectangulaire*, J. Math. Pures Appl., 68 (1989), pp. 457–465.

[Le] G. Lebeau, *Control for hyperbolic equations*, Journées "équations aux dérivées partielles" Saint Jean de Monts, 1992, École Polytechnique, Palaiseau, 1992.

[Li] J.-L. Lions, *Contrôlabilité exacte, perturbation et stabilisation des systèmes distribués*, Rech. Math. Appl. 8-9, Masson, Paris, 1988.

[M] S. Micu, *Analisis de un sistema hibrido bidimensional fluido-estructura*, Ph.D. thesis, Universidad Complutense de Madrid, Madrid, Spain, 1996.

[MZ1] S. Micu and E. Zuazua, *Asymptotics for the spectrum of a fluid/structure hybrid system arising in the control of noise*, SIAM J. Math. Anal., 29 (1998), pp. 967–1001.

[MZ2] S. Micu and E. Zuazua, *Boundary controllability of a linear hybrid system arising in the control of noise*, SIAM J. Control Optim., 35 (1997), pp. 1614–1637.

[Z] A. Zygmund, *Trigonometric Series*, Cambridge University Press, Cambridge, UK, 1968.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 3, pp. 869–891

CONVERGENCE RESULTS FOR THE FLUX IDENTIFICATION IN A SCALAR CONSERVATION LAW[*]

FRANÇOIS JAMES[†] AND MAURICIO SEPÚLVEDA [‡]

Abstract. Here we study an inverse problem for a quasilinear hyperbolic equation. We start by proving the existence of solutions to the problem which is posed as the minimization of a suitable cost function. Then we use a Lagrangian formulation in order to formally compute the gradient of the cost function introducing an adjoint equation. Despite the fact that the Lagrangian formulation is formal and that the cost function is not necessarily differentiable, a viscous perturbation and a numerical approximation of the problem allow us to justify this computation. When the adjoint problem for the quasi-linear equation admits a smooth solution, then the perturbed adjoint states can be proved to converge to that very solution. The sequences of gradients for both perturbed problems are also proved to converge to the same element of the subdifferential of the cost function. We evidence these results for a large class of numerical schemes and particular cost functions which can be applied to the identification of isotherms for conservation laws modeling distillation or chromatography. They are illustrated by numerical examples.

Key words. inverse problem, scalar conservation laws, adjoint state, gradient method

AMS subject classifications. 35R30, 35L65, 65K10, 49M07

PII. S0363012996272722

1. Introduction.

In this paper, we are interested in the following inverse problem: Consider the scalar hyperbolic conservation law

$$\partial_t w + \partial_x f(w) = 0, \qquad x \in \mathbb{R},\ t > 0, \tag{1.1}$$

together with the Cauchy data

$$w(x,0) = w^0(x) \in BV(\mathbb{R}) \cap L^\infty(\mathbb{R}). \tag{1.2}$$

It is well known that there exists one and only one entropy solution in $L^\infty(\mathbb{R}_+, BV(\mathbb{R})) \cap L^\infty(\mathbb{R} \times (-\infty, +\infty))$ of (1.1)–(1.2) (see [6], [18]), and we emphasize the fact that the unique entropy solution to (1.1) depends continuously (in a sense which we shall describe precisely) on the smooth function f by denoting it w_f. The question we address is whether, given an observation w^{obs} at time $T > 0$, one can identify the nonlinearity f such as w_f at time T is as close as possible to w^{obs}.

It is quite natural to formulate this problem more or less like an optimal control problem: For any function $v : \mathbb{R} \to \mathbb{R}$ we define a cost function $J(v)$, and we look for an f solving

$$\min_f J(w_f(.,T)), \tag{1.3}$$

thus giving a precise meaning to the phrase "as close as possible." Therefore we are led to the constrained optimization problem of minimizing $J(w(.,T))$ under the constraint

[*] Received by the editors October 23, 1996; accepted for publication (in revised form) April 1, 1998; published electronically March 17, 1999. This research was partially supported by Program A on Numerical Analysis of FONDAP in Applied Mathematics, Fondecyt 1980123, and the Universidad de Concepción (P.I. 97.13.09-1.2 and 97.013.011-1.In).

http://www.siam.org/journals/sicon/37-3/27272.html

[†] Mathématiques, Applications et Physique Mathématique d'Orléans, UMR CNRS 6628, Université d'Orléans, B.P. 6759, F-45067 Orléans Cedex 2, France (james@cmapx.polytechnique.fr).

[‡] Departamento de Ingeniería Matemática, Facultad de Ciencias Físicas y Matemáticas, Universidad de Concepción, Casilla 160-C, Concepción, Chile (mauricio@ing-mat.udec.cl).

for w to satisfy the partial differential equation (1.1)–(1.2). This problem can be viewed as well as an unconstrained minimization problem: If we set $\tilde{J}(f) = J(w_f)$, then problem (1.3) boils down to minimizing $\tilde{J}$ on a suitable set of functions.

In theory, this inverse problem is in general ill-posed in uniqueness when there are discontinuities in the solution. For instance, a well-known undesirable case appears when we try to identify f over a shock wave with a propagation speed equal to σ: there are infinitely many functions f giving the same entropic solution w_f of (1.1)–(1.2) equal to the shock wave (see [4] for more details). Yet, as far as applications are concerned some interesting practical problems can be found: It is possible to resolve the identification of f (or "a part of f") via a gradient technique in order to compute numerically the minimum of $\tilde{J}$. This was achieved in a preceding paper [9] in which we considered the identification problem arising from a model of diphasic propagation in chromatography. Therefore we dealt with a system of conservation laws, and we obtained successful numerical results because the function f was given a precise analytic form, so the minimization occurred on $\mathbb{R}^n$, and we chose adequate criteria for the cost function linked to the physical parameter of the problem.

The classical gradient technique used in order to obtain the gradient of $\tilde{J}$ consists of writing a Lagrangian formulation for the constrained problem and introducing the adjoint state. This has to be done at two levels.

First we consider a formal level; that is, we take a solution of the continuous equation (1.1) and perform the computations. We obtain a backward linear hyperbolic equation for the adjoint state. The trouble is that this equation is ill-posed as soon as the solution of (1.1) is not smooth—which is of course the case in most of the applications. This is related to the fact that the inverse problem is ill-posed in uniqueness when there are discontinuities in the solution. Thus, in general, the computation of the gradient of $\tilde{J}$ remains formal. Furthermore, it is easy to find some counterexamples where the gradient does not exist.

On the other hand, we can perform the same computations at a discrete level, that is, when both (1.1) and the cost function J are discretized. This introduces a "discrete adjoint state" which we call the adjoint scheme, and we obtain the gradient for the discretization of $\tilde{J}$, which is well defined. Thus we are able to perform numerical computations, using standard conjugate gradient techniques, and the numerical evidence is that the method seems to converge (see [9, section 5] and [11] for application on real data).

The aim of this paper is to interpret and justify the convergence of the method in the scalar case and in a particular case, namely, when the solution of the adjoint state is Lipschitz continuous. We shall consider two modified problems: First we add a viscous term to (1.1), then we turn to the discretized problem. In both cases, we prove that the perturbed adjoint states converge to the solution of the original problem. That enables us to pass to the limit in the approximation of the gradient, and we prove that both approximations tend to the same limit. This limit is not necessarily the gradient of the cost function because the gradient does not exist a priori. In fact, we also prove by means of convexity hypotheses that it is an element of the subdifferential of $\tilde{J}$. This result gives an interpretation of the formal computation of the gradient for continuous cost functions including some cases when the gradient does not exist.

Therefore the paper is organized as follows. First we precisely state the problem, in particular concerning the cost function we consider, which is not the standard least square function. Then we consider the identification problem for a parabolic regularization of the conservation equation and, in particular, we prove the differentiability of the cost function. We also prove the convergence of the sequence of the perturbed gradients to an element of the subdifferential of $\tilde{J}$. Finally we prove that we can obtain the same element of the subdifferential via a discretized problem and for a large class of numerical schemes, and we illustrate these results by a numerical application on experimental data.

2. The identification problem.

2.1. The cost function. A classical example of cost function J arises in the well-known output least square method (see [4] for instance):

$$J_0(w) = \frac{1}{2}\int_{x\in\mathbb{R}} |w(x,T) - w^{\mathrm{obs}}(x)|^2 dx. \tag{2.1}$$

For practical reasons, in [9] the following modified cost function J_ρ was used:

$$J_\rho(w) = J_0(w) + \frac{\rho}{2}|\mu_1\big(w(\cdot,T)\big) - \mu_1(w^{\mathrm{obs}})|^2, \tag{2.2}$$

where ρ is a constant parameter to be adjusted, and where $\mu_1(\mathrm{X})$ is the first moment of the function $X : \mathbb{R} \to \mathbb{R}$:

$$\mu_1(\mathrm{X}) = \int_{\mathbb{R}} x\mathrm{X}(x)dx. \tag{2.3}$$

Roughly speaking, the advantage of J_ρ over J_0 is that it is more sensitive to the localization of the observed signal on the x-axis, whereas J_0 essentially takes care of the shape of the signal over its localization.

Notice that we shall always consider initial data with compact support so that, by the finite velocity of propagation, the solution at any time $t > 0$ will also have a compact support. Thus all the integrals in the definitions of J_0 and of J_ρ have a meaning as soon as w is in $L^\infty(\mathbb{R})$.

For practical reasons, what follows is essentially focused on the study of criteria (2.1) or (2.2)–(2.3).

2.2. Existence and Lipschitz continuous dependence. We can assure the existence of at least one solution f of our identification problem (1.3) when we search it in a compact of the Lipschitz continuous functions. In fact, this is a consequence of the following theorem, the proof of which is contained in a paper by Lucier (see [13]).

THEOREM 2.1. *The application $f \mapsto w_f$ is Lipschitz continuous from the space of Lipschitz functions to L^1, that is,*

$$\|w_f(.,t) - w_g(.,t)\|_{L^1} \leq t\|f-g\|_{\mathrm{Lip}}\|w^0\|_{BV}, \tag{2.4}$$

where $\|\cdot\|_{BV}$ is the usual norm of $BV(\mathbb{R})$.

COROLLARY 2.2. *The function $\tilde{J} : f \mapsto J_\rho(w_f)$ is Lipschitz continuous from* $\mathrm{Lip}(\mathbb{R})$ *to* $\mathbb{R}_+$.

Proof. Concerning the function J_0, the estimate follows immediately from the L^∞ estimate which holds uniformly in g:

$$\forall t > 0, \quad \|w_g(.,t)\|_{L^\infty} \leq \|w^0\|_{L^\infty}. \tag{2.5}$$

Indeed we have

$$\begin{aligned} |J_0(g) - J_0(f)| &\leq \frac{1}{2}\int_{\mathbb{R}} |(w_g - w_f)(x,T)|\,|(w_g + w_f)(x,T) - 2w^{\mathrm{obs}}(x)|\,dx \\ &\leq \max(\|w^0\|_{L^\infty}, \|w^{\mathrm{obs}}\|_{L^\infty})\,\|w_g(,.T) - w_f(.,T)\|_{L^1} \end{aligned}$$

by Hölder's inequality. The result follows by (2.4). Now for the momentum criterion, if we assume all the supports included in $[-L, L]$ for some L large enough, we get, again using (2.5):

$$\begin{aligned} |J_1(g) - J_1(f)| &\leq \frac{1}{2}|\mu_1(w_g(.,T)) - \mu_1(w_f(.,T))|\,|\mu_1(w_g + w_f(.,T)) - 2\mu_1(w^{\mathrm{obs}})| \\ &\leq L^3 \max(\|w^0\|_{L^\infty}, \|w^{\mathrm{obs}}\|_{L^\infty})\,\|w_g(.,T) - w_f(.,T)\|_{L^1}. \qquad \square \end{aligned}$$

By considering a minimizing sequence for J_ρ, we easily deduce form Corollary 2.2 the following existence result.

COROLLARY 2.3. *If $f \in \mathcal{F}$ is a compact set of the Lipschitz continuous functions, then there exists at least one solution of the identification problem* (1.3)*, with the cost function defined by* (2.1) *or* (2.2)–(2.3).

Remarks. The compactness of the set $\mathcal{F}$ is a necessary hypothesis, but it is not a restrictive condition for many practical identification problems: the function f can have a precise analytic form so that the minimization occurs on a bounded subset of $\mathbb{R}^n$ (for instance, see [9]). Another way to obtain Lipschitz compactness is to seek f, for instance, in $W^{2,\infty}(\mathbb{R})$.

We cannot ensure the uniqueness of the solution in Corollary 2.3 for the reasons given in the introduction. Obviously, we can try to modify the cost functions to obtain a strictly convex functional and, for instance, search the flux f with minimal $W^{2,\infty}(\mathbb{R})$ norm. Yet in general, this is an arbitrary mathematical condition. Thus we prefer to deal with the cost functions defined by (2.1), (2.2), and (2.3), which have a realistic physical sense and to leave the uniqueness problem as an open problem.

2.3. Remarks on differentiability.

Since we are concerned with the problem of minimizing $\tilde{J}(f) = J_\rho(w_f)$ with respect to f, differentiability is of course of crucial importance: optimality conditions, gradient algorithms rely on it. For the function $\tilde{J}$, the problem is open in general, and we are going to point out precisely the difficulties. These will come of course from the operator $f \mapsto w_f$, since the functions J_0 and J_1 are smooth convex functions. Let us first consider this operator in the nice case where all the involved functions are smooth. Indeed, if $w^0 \in C^1(\mathbb{R})$ and $f \in C^1(\mathbb{R})$, then there exists $T > 0$ such that $w_f \in C^1(\mathbb{R} \times [0, T])$. Let us use the notations of the calculus of variations, denoting by δw the variation of w_f corresponding to some variation δf of f. Then δw has to solve

$$(2.6) \qquad \left\{ \begin{aligned} \partial_t \delta w + \partial_x[f'(w_f)\delta w] + \partial_x[\delta f(w_f)] &= 0, \\ \delta w(.,0) &= 0. \end{aligned} \right.$$

Under the above assumptions, we are faced with a standard linear conservation equation with smooth coefficients. Actually, by applying the inverse functions theorem to

$$\begin{aligned} F \quad : \quad C^1(\mathbb{R}) \times \mathrm{Lip}(\mathbb{R}) &\rightarrow C^0(\mathbb{R}) \times C^0(\mathbb{R}), \\ (w, f) &\mapsto (\partial_t w + \partial_x f(w), w(.,0) - w^{\mathrm{obs}}), \end{aligned}$$

we can prove that $f \mapsto w_f$ is differentiable in a strong sense from $\mathrm{Lip}(\mathbb{R})$ to $C^1(\mathbb{R})$. Thus $\tilde{J}$ is also differentiable, with

$$(2.7) \qquad \begin{aligned} \tilde{J}'(f)\delta f \quad = \quad & \int_{\mathbb{R}} (w_f(x,T) - w^{\mathrm{obs}}(x))\delta w(x,T)\, dt \\ & + \rho\,(\mu_1(w_f(.,T)) - \mu_1(w^{\mathrm{obs}}))\,\mu_1(\delta w(.,T)). \end{aligned}$$

This, by the way, justifies in this case the integrations by parts we perform in the next section.

We would like to point out now that if w_f happens to be discontinuous, then the resolution of (2.6) is much more difficult. Indeed it becomes a conservation law with discontinuous coefficient $(f'(w_f))$ and a measure-valued source term $(\partial_x \delta f(w_f))$. The solution therefore has to be sought in the class of measures on $\mathbb{R}$, generalizing the results obtained by Bouchut and James [2], [3] in the absence of a source term. Until now, this has been possible only if f is convex. In this context, it is reasonable to hope for some very weak differentiability results, the involved topology being the usual weak convergence of measures. The general case of a nonconvex f is still completely open.

However, even the differentiability of $f \mapsto w_f$ does not settle the problem for $\tilde{J}(f)$. Indeed since δw is a measure, (2.7) is meaningful if $w_f(.,T) - w^{\text{obs}} \in L^1(\delta w(.,T))$, which is a priori not obvious. In the same way, $\mu_1(\delta w(.,T))$ has to be defined. We refer to the next section for an explicit example and further comments.

For all these reasons, we leave the problem of a rigorous differentiation of $\tilde{J}$ (and possibly the choice of theoretically more convenient cost functions) to a future work. In the following, we focus on the study of two approximated problems where all the involved quantities are well defined. We prove, when it is possible, the convergence of these problems to the original one. We begin by a formal computation of the gradient of $\tilde{J}$, as it was done in [9]. The basic tool for that is to consider the constrained minimization problem and its Lagrangian formulation.

2.4. Lagrangian formulation and the adjoint problem. In [9] and [17], we formally obtain the gradient through the following Lagrangian formulation for the constrained minimization problem:

$$\mathcal{L}\left(w,p;f\right) \overset{def}{=} J(w) - E\left(w,p;f\right), \tag{2.8}$$

where $E\left(w,p;f\right)$ is a weak form of (1.1), defined by

$$E\left(w,p;f\right) = -\int_0^T \int_0^L \left(w\partial_t p + f(w)\partial_z p\right) + \int_{t=T} wp - \int_{t=0} w^0 p. \tag{2.9}$$

We are interested in canceling $(\partial\mathcal{L})/(\partial w)$. For that purpose, we take p solution to the following backward adjoint problem:

$$\begin{cases} \partial_t p + f'(w_f)\partial_x p = 0, & x \in \mathbb{R},\ t \le T, \\ p(x,T) = p^T(x), \end{cases} \tag{2.10}$$

where $f'(w_f)$ represents the derivative of f with respect to w evaluated on the solution w_f of (1.1). The function p^T depends on J, w, and w^{obs}. More precisely, we have

$$\int_{\mathbb{R}} p^T \delta w = D_w J(w_f)\delta w \qquad \forall \delta w \in \mathcal{D}(\mathbb{R}), \tag{2.11}$$

where $D_w J\delta w$ represents the derivative of J in the direction δw. The problem (2.10) is called the *adjoint problem* associated with the direct problem (1.1).

Thus we can compute the gradient of $\tilde{J}$ by the formula

$$D\tilde{J}(f)\delta f = \int_0^T \int_{-\infty}^{+\infty} \partial_x p\, \delta f(w_f) dx dt, \tag{2.12}$$

where p is solution of (2.10), in a sense which has to be made precise (see Remark below).

Remark. Formula (2.12) is a formal result, since the derivative of function $\widetilde{J}$ does not necessarily exist, as a rule. For instance, consider the entropic solution of the Riemann problem

$$w_k(x,t) = \begin{cases} 1 & \text{if } x \geq kt, \\ 0 & \text{otherwise} \end{cases}$$

for the Burgers equation—$f(w) = kw^2$—and suppose $w^{\text{obs}}(x) = w_{k_0}(x,T)$ for some given k_0. Then we have that the cost function given by the criterion of the norm L^2 (2.1) is equal to

$$\tilde{J}_0(k) = J_0(w_k) = T|k - k_0|, \tag{2.13}$$

which is not differentiable in k_0. Moreover, the backward equation defining p is ill-posed as soon as discontinuities occur in the solution of the direct problem. Actually the solution is not uniquely defined by the characteristics.

If we assume that we are in the neighborhood of a minimum, the function $\tilde{J}$ is locally convex, so the subdifferential $\partial\tilde{J}$ is a nonempty set. We may hope that $D\tilde{J}(f)$ defined by (2.12) is an element of $\partial\tilde{J}$ when the adjoint equation is ill-posed. We are not going to answer this question here. We will restrict ourselves to the particular case where there exists a smooth solution to the adjoint equation. In this case we shall study whether (2.12) is well defined and whether it is an element of the subdifferential. First, we give an existence result for Lipschitz solutions to the adjoint problem (2.10), then, we give conditions for the uniqueness.

Definition 2.4. *The function $a(x,t)$ verifies the one-side-Lipschitz-continuous condition (OSLC) when there exists a function $m \in L^1(0,T)$ such as*

$$L^+\left(a(x,t)\right) \overset{def}{=} \text{ess} \sup_{x \neq y} \left(\frac{a(x,t) - a(y,t)}{x - y} \right)^+ \leq m(t). \tag{2.14}$$

In other words, (OSLC) means that the function $a(\cdot,t)$ must be Lipschitz continuous for all t, when $a(\cdot,t)$ is increasing, and it allows decreasing jumps $a(x-,t) > a(x+,t)$. This condition has been used by several authors, e.g., Oleinik [14], Conway [5], Hoff [8], and Tadmor [19], to prove the existence of at least one Lipschitz continuous solution to the adjoint problem (2.10), when $p^T \in W^{1,\infty}(\mathbb{R})$. A refined version of Oleinik's entropy condition (Hoff [8]) states that $f'(w)$ verifies (OSLC) when w accepts only entropic shocks as discontinuities and when f is convex. We need this result to make sense out of solving p by the characteristics method. So far, we do not know a general existence result for any f. The problem with this result is that its solution is ill-posed in uniqueness. For instance, if we assume that $f'(w_f)(x,t) = -\text{sign}(x)$, and $\|p^T\|_{W^{1,\infty}(\mathbb{R})} > 0$, then it is easy to verify that

$$p(x,t) = \begin{cases} p^T\left(x - \text{sign}(x)(T-t)\right) & \text{when } t + |x| \geq T, \\ \varphi(t + |x|) & \text{otherwise} \end{cases}$$

are Lipschitz continuous solutions of adjoint problem (2.10), for any $\varphi(t)$ Lipschitz continuous function such that $\varphi(T) = p^T(0)$.

We recall briefly here the definition of the so-called "reversible solutions" introduced in [2], [3], for which uniqueness holds. First, we define the set $\mathcal{E}$ of "exceptional solutions" as the space of all the Lipschitz continuous solutions of (2.10) with $p^T = 0$. Next, we introduce the open set called "support of exceptional solutions":

$$\mathcal{V} \overset{def}{=} \{(t,x) \in]0,T[\times\mathbb{R} \quad | \quad \exists p_e \in \mathcal{E},\ p_e(t,x) \neq 0\}. \tag{2.15}$$

The following result is obtained in [3].

THEOREM 2.5. *Let p be a Lipschitz continuous solution of* (2.10). *Then, the following properties are equivalent:*

(i) *p is locally constant in $\mathcal{V}$;*

(ii) *there exist p_1 and p_2 in $\mathrm{Lip}_{\mathrm{loc}}([0,T]\times\mathbb{R})$, verifying $\partial_t p_i + f'(w)\partial_x p_i = 0$ and $\partial_x p_i \geq 0$, such that $p = p_1 - p_2$.*

Furthermore, for all $p^T \in W^{1,\infty}(\mathbb{R})$ there exists one and only one Lipschitz continuous solution p of (2.10) *verifying one of these properties and the following estimate:*

$$\|p(\cdot,t)\|_{W^{1,\infty}(\mathbb{R})} \leq \|p^T\|_{W^{1,\infty}(\mathbb{R})} \cdot \exp\left\{\int_t^T m(\tau)d\tau\right\} \qquad \forall t \leq T. \tag{2.16}$$

This function p is called the reversible solution *of* (2.10).

Remarks. According to property (i) of Theorem 2.5, we can choose a constant for p in each *fan-wise set* defined by all the characteristic straight lines which converge to a discontinuity of w. This constant is equal to the value of $p^T(x)$, where (x,T) is a point in the discontinuity of w. Property (ii) is a monotonicity property of the reversible solution.

Choosing the reversible solution in counterexample (2.13) is equivalent to choosing the characteristic element of the subdifferential of $\tilde{J}_0(k_0)$ equal to zero. This is not an arbitrary choice. At the limit, we will see that we obtain an element of the subdifferential of $\tilde{J}$ characterized by the reversible solution, when we consider viscosity approximation and some classical numerical schemes.

When we consider the cost function (2.1) or (2.3), the hypothesis $p^T \in W^{1,\infty}(\mathbb{R})$ is equivalent to $(w(\cdot,T) - w^{\mathrm{obs}}) \in W^{1,\infty}(\mathbb{R})$. In practice, this is a very restrictive hypothesis in the sense that it means a regular solution w, or at least a solution in which the shocks at $t = T$ are canceled with the shocks of the observation w^{obs}. When w is regular (without shocks), it is clear that we do not need the notion of reversible solution: In this case we have seen that the cost function is differentiable. The resolution of (2.10) when $p^T \notin W^{1,\infty}(\mathbb{R})$ remains an open problem.

3. Artificial viscosity.

We introduce the classical viscous regularization of (1.1):

$$\begin{cases} \partial_t w^\varepsilon + \partial_x f(w^\varepsilon) = \varepsilon \partial_{xx}^2 w^\varepsilon, & x \in \mathbb{R},\ t > 0, \\ w^\varepsilon(x,0) = w^0(x). \end{cases} \tag{3.1}$$

It is well known that (3.1) admits a unique smooth solution which approaches the entropy solution of (1.1) in the following sense (see, e.g., Smoller [18]).

THEOREM 3.1. *We suppose that $w^0 \in BV(\mathbb{R})$. Then,*

(i) *Problem* (3.1) *has a solution $w^\varepsilon(\cdot,t) \in BV(\mathbb{R})$ for all $t > 0$. This solution is C^1 on $\mathbb{R}\times(0,\infty)$. Furthermore, for all $t > 0$,*

$$\|w^\varepsilon\|_{L^\infty(\mathbb{R})} \leq \textit{constant}, \qquad TV(w^\varepsilon(\cdot,t)) \leq TV(w^0).$$

(ii) *The only accumulation point in $L^\infty(\mathbb{R}\times(0,+\infty))$-weakly$*$, and $L^1_{\mathrm{loc}}(\mathbb{R}\times(0,+\infty))$-strong, of the sequence w^ε is the entropic solution w of* (1.1).

Now we shall limit ourselves to the (restrictive) case where the final data for the adjoint state are in $W^{1,\infty}(\mathbb{R})$ and consider the following minimization problem:

$$\min_f \widetilde{J}^\varepsilon(w_f^\varepsilon), \qquad \text{where } \widetilde{J}^\varepsilon(w_f^\varepsilon) = J_\rho(w_f^\varepsilon), \tag{3.2}$$

w^ε_f is the solution of the parabolic problem (3.1), and the cost function J_ρ is defined by (2.2).

We are going to proceed in three steps. First we compute the *exact* derivative of the function $\widetilde{J}^\varepsilon$. Then we shall prove that the adjoint equation is well-posed, which will give a characterization of the derivative. Finally, we are going to prove the convergence of the adjoint state and of the derivative of the viscous problem to the corresponding quantities for the hyperbolic problem, when ε tends to 0.

3.1. Derivative of the viscous cost function. We want to determine the Gateaux derivative of $\widetilde{J}^\varepsilon$. We shall prove that the directional derivative

$$D\widetilde{J}^\varepsilon(f)\delta f = \lim_{h\to 0} \delta_h \widetilde{J}^\varepsilon(f) = \lim_{h\to 0} \frac{\widetilde{J}^\varepsilon(f+h\delta f) - \widetilde{J}^\varepsilon(f)}{h} \tag{3.3}$$

exists for all Lipschitz continuous function δf (δf is called a Lipschitz direction). We then have the following propostion.

PROPOSITION 3.2. *Let w^ε_f be the solution of the viscous problem* (3.1). *We suppose that the flux $f = f(w^\varepsilon)$ is of class C^1 and Lipschitz continuous with respect to w^ε with a Lipschitz constant C_{Lip}. Then the limit* (3.3) *exists for all Lipschitz direction δf (which we can suppose with the same Lipschitz constant C_{Lip}). It is characterized by*

$$D\widetilde{J}^\varepsilon(f)\delta f = D_w J_\rho(w^\varepsilon_f) w^\varepsilon_1, \tag{3.4}$$

where w^ε_1 is the solution in $L^2(0,T,L^2(\mathbb{R})) \cap C^1(\mathbb{R}\times(0,+\infty))$ of the following linear parabolic problem:

$$\begin{cases} \partial_t(w^\varepsilon_1) + \partial_x\left(f'(w^\varepsilon_f)w^\varepsilon_1 + \delta f(w^\varepsilon_f)\right) = \varepsilon\partial^2_{xx}w^\varepsilon_1, \\ w^\varepsilon_1(x,0) = 0. \end{cases} \tag{3.5}$$

Proof. Recall that $J_\rho = J_0 + \rho J_1$. Then we can write

$$\begin{cases} D_w J_0(w^\varepsilon_f)w^\varepsilon_1 = \displaystyle\int_{\mathbb{R}} \left(w^\varepsilon_f(x,T) - w^{\text{obs}}(x)\right) w^\varepsilon_1(x,T)dx, \\ D_w J_1(w^\varepsilon_f)w^\varepsilon_1 = \left(\mu_1\big(w^\varepsilon_f(\cdot,T)\big) - \mu_1(w^{\text{obs}})\right)\mu_1\big(w^\varepsilon_1(\cdot,T)\big), \end{cases} \tag{3.6}$$

and we remark that

$$\begin{cases} \delta_h J_0(w^\varepsilon_f) = \displaystyle\int_{\mathbb{R}} \left(\frac{w^\varepsilon_f + w^\varepsilon_{f^h}}{2}(x,T) - w^{\text{obs}}(x)\right)\delta_h w^\varepsilon(x,T)dx, \\ \delta_h J_1(w^\varepsilon_f) = \left(\dfrac{\mu_1(w^\varepsilon_f(\cdot,T)) + \mu_1(w^\varepsilon_{f^h}(\cdot,T))}{2} - \mu_1(w^{\text{obs}})\right)\mu_1(\delta_h w^\varepsilon(\cdot,T)), \end{cases} \tag{3.7}$$

where $f^h = f + h\delta f$ and $\delta_h w^\varepsilon = w^\varepsilon_{f^h} - w^\varepsilon_f/h$.

We first prove the convergence of the derivative of the L^2-criterion:

$$\delta_h \widetilde{J}_0(w^\varepsilon_f) \to D_w J_0(w^\varepsilon_f)w^\varepsilon_1, \qquad \text{when } h \to 0. \tag{3.8}$$

We have that w^ε_f and $w^\varepsilon_{f^h}$ are solutions of the parabolic problem (3.1) with the respective nonlinear flux f and f^h. Thus, if we take the difference between the equation in w_f^ε and the equation in $w_{f^h}^\varepsilon$, we deduce that $\delta_h w^\varepsilon$ is the solution of

$$\partial_t \delta_h w^\varepsilon + \partial_x \left(\frac{f^h(w_{f^h}^\varepsilon) - f^h(w_f^\varepsilon)}{h} + \delta f(w_f^\varepsilon) \right) = \varepsilon \partial_{xx}^2 \delta_h w^\varepsilon. \tag{3.9}$$

We multiply this equation by $\delta_h w^\varepsilon$ and we integrate by parts. Using classical estimate arguments and the Gronwall lemma, we prove

$$\|\delta_h w^\varepsilon(\cdot,t)\|_{H^1} \leq C(\varepsilon) \quad \forall t \in (0,T), \tag{3.10}$$

where $C(\varepsilon)$ is a constant which does not depend on h. Thus, up to a subsequence, there exists $w_1^\varepsilon(\cdot,t)$, such that

$$\delta_h w^\varepsilon(\cdot,T) \rightharpoonup w_1^\varepsilon(\cdot,T) \text{ in } H^1(\mathbb{R})\text{-weak}, \qquad \text{when } h \to 0.$$

We can easily verify that $w_{f^h}^\varepsilon(\cdot,T) \to w_f^\varepsilon(\cdot,T)$ in $L^2(\mathbb{R})$-strong when $h \to 0$. Hence we deduce (3.8).

Next we prove the convergence of the derivative of the first moment criterion:

$$\delta_h \widetilde{J}_1(w_f^\varepsilon) \to D_w J_1(w_f^\varepsilon) w_1^\varepsilon, \qquad \text{when } h \to 0. \tag{3.11}$$

We deduce from the compact support of w^0, and from the maximum principle applied to (3.1), (3.5), and (3.9), that the functions $|w_f^\varepsilon(\cdot,T)|$, $|w_{f^h}^\varepsilon(\cdot,T)|$, $|\delta_h w^\varepsilon(\cdot,T)|$, and $|w_1^\varepsilon(\cdot,T)|$ are bounded by a function $g(x) = C\exp(-r|x|)$, where C, r are constants which depend only on T and $\|w^0\|_{L^\infty}$, and thus are independent of h. We obtain that

$$\begin{aligned}
&|\mu_1(\delta_h w^\varepsilon(\cdot,T)) - \mu_1(w_1^\varepsilon(\cdot,T))| \\
&\qquad \leq \int_{-R}^{R} |x\,(\delta_h w^\varepsilon - w_1^\varepsilon)\,(x,T)|dx \; + \int_{|x|>R} |xg(x)|dx.
\end{aligned}$$

Using the convergence L^2-weak of $\delta_h w^\varepsilon(\cdot,T)$ we have that the first term of the right-hand side converges to 0. The second term uniformly converges to 0 in h, when $R \to \infty$. This implies the convergence $\mu_1(\delta_h w^\varepsilon(\cdot,T)) \to \mu_1(w_1^\varepsilon(\cdot,T))$, when $h \to 0$. In the same way, and using the convergences L^2-strong of $w_{f^h}^\varepsilon(\cdot,T)$, we have that $\mu_1(w_{f^h}^\varepsilon(\cdot,T)) \to \mu_1(w_f^\varepsilon(\cdot,T))$, when $h \to 0$. From these convergences of the moments, we deduce the result (3.11).

Finally, we prove that w_1^ε is the solution of (3.5). Using (3.10) and the compact injection $H^1_{\mathrm{loc}}(\mathbb{R}) \hookrightarrow L^2_{\mathrm{loc}}(\mathbb{R})$, we have the following strong convergence:

$$\delta_h w^\varepsilon(\cdot,t) \to w_1^\varepsilon(\cdot,t) \text{ in } L^2_{\mathrm{loc}}(\mathbb{R})\text{-strong}, \qquad \text{when } h \to 0. \tag{3.12}$$

We multiply (3.9) by a test function $\varphi \in C_0^1([0,+\infty) \times \mathbb{R})$ and we integrate by parts. Hence, by passing to the limit in (3.9), we obtain that w_1^ε is a weak solution of (3.5). Therefore, by the existence and uniqueness of the solution of the linear parabolic problem (3.5), we have that function w_1^ε is the strong solution in $L^2(0,T,L^2(\mathbb{R})) \cap C^\infty(\mathbb{R} \times (0,+\infty))$ of this equation. □

3.2. Viscous adjoint problem. We showed that the Gateaux derivative of $\widetilde{J}^\varepsilon(f)$ is well defined for cost function J_ρ defined by (2.2). Now we shall use the Lagrangian formulation in order to give a characterization of this derivative. First we define the weak form associated with (3.1) by

$$\begin{aligned}
E^\varepsilon\left(w^\varepsilon, p^\varepsilon, f\right) \overset{def}{=} &- \int_0^T \int_{-\infty}^{+\infty} \left(w^\varepsilon \partial_t p^\varepsilon + f(w^\varepsilon)\partial_x p^\varepsilon - \varepsilon \partial_x w^\varepsilon \partial_x p^\varepsilon\right) \\
&+ \int\limits_{t=T} w^\varepsilon p^\varepsilon - \int\limits_{t=0} w^\varepsilon p^\varepsilon
\end{aligned}$$

and the Lagrangian by $L^\varepsilon(w^\varepsilon, p^\varepsilon, f) = J(w^\varepsilon) - E^\varepsilon(w^\varepsilon, p^\varepsilon, f)$, where p^ε is a regular function in (x,t). We take p^ε equal to the solution of the following backward parabolic equation (called *viscous adjoint problem*):

$$\left\{\begin{array}{ll} \partial_t p^\varepsilon + f'(w^\varepsilon)\partial_x p^\varepsilon = -\varepsilon\partial^2_{xx}p^\varepsilon, & x\in\mathbb{R},\ t<T, \\ p^\varepsilon(x,T) = p^T(x), & \end{array}\right. \tag{3.13}$$

where the final condition p^T is defined by (2.11). In this case, the derivative of the Lagrangian with respect to w^ε is equal to zero, and the Gateaux-derivative of the cost function is characterized by

$$D\widetilde{J}^\varepsilon(f)\delta f = \int_0^T \int_{-\infty}^{+\infty} \partial_x \mathrm{p}^\varepsilon \delta f(w^\varepsilon_f). \tag{3.14}$$

Equation (3.13) is a parabolic linear equation, and it is known that it admits only one solution in $L^2(0,T,\,L^2(\mathbb{R}))\cap C^1(\mathbb{R}\times(0,\infty))$ which depends on w^ε.

On the other hand, if $\beta \geq f'' \geq \alpha > 0$, we can prove the following (OSLC) estimate (see [19]) by a maximum principle argument applied to (3.1):

$$L^+\left(f'(w^\varepsilon(x,t))\right) \leq \frac{\beta L^+\left(w^0\right)}{1+\alpha t L^+\left(w^0\right)} = m(t) \in L^1(0,T). \tag{3.15}$$

From this result, we shall deduce BV and $W^{1,\infty}$ estimates on the adjoint state.

THEOREM 3.3. *We consider the solution $p^\varepsilon(x,t)$ of linear parabolic problem* (3.13)*, with $p^T \in W^{1,\infty}(\mathbb{R})\cap BV(\mathbb{R})$. We suppose $\beta \geq f'' \geq \alpha > 0$, with α, β constants independent of w. Then $(x,t)\mapsto f'(w_f(x,t))$ verifies the* (OSLC)*, and*

$$\|p^\varepsilon(\cdot,t)\|_{L^\infty(\mathbb{R})} \leq \|p^T\|_{L^\infty(\mathbb{R})}, \tag{3.16}$$

$$\|\partial_x p^\varepsilon(\cdot,t)\|_{L^1(\mathbb{R})} \leq \|p^T\|_{BV(\mathbb{R})}, \tag{3.17}$$

$$\|\partial_x p^\varepsilon(\cdot,t)\|_{L^\infty} \leq \|p^T\|_{W^{1,\infty}} \cdot \exp\left\{\int_t^T m(T-\tau)d\tau\right\} \tag{3.18}$$

for all $t \leq T$, where $m \in L^1(0,T)$ is the function defined in (3.15)*.*

Proof. Estimates (3.16) and (3.17) are classical results of the theory of nonlinear hyperbolic equations, and the proofs can be found in [12] and [6]. The proof of the estimate (3.18) is very similar to the arguments used by Tadmor [19]. Let us recall them briefly.

We consider p and p^ε solutions to the problems (2.10) and (3.13), respectively. Let $q^\varepsilon(x,t) = p^\varepsilon(-x, T-t)$ and $\psi_\varepsilon(x,t) = \partial_x q^\varepsilon(x,t)$. We differentiate (3.13), and we notice that the function ψ_ε verifies

$$\begin{aligned} &\partial_t\psi_\varepsilon + f'(w^\varepsilon(-x,T-t))\partial_x\psi_\varepsilon \\ &\qquad = -\left(\partial_x f'(w^\varepsilon(-x,T-t))\right)\psi_\varepsilon + \varepsilon\partial^2_{xx}\psi_\varepsilon. \end{aligned} \tag{3.19}$$

Let $\lambda \geq 2$ be an even integer. We multiply (3.19) by $\lambda\psi_\varepsilon^{\lambda-1}$ and we integrate by parts. We obtain

$$\begin{aligned} \frac{d}{dt}\|\psi_\varepsilon(\cdot,t)\|^\lambda_{L^\lambda} = &-(\lambda-1)\int_{x\in\mathbb{R}}\left\{\partial_x\frac{\partial f}{\partial w}(w^\varepsilon(-x,T-t))\right\}\psi_\varepsilon^\lambda dx \\ &-\varepsilon\lambda(\lambda-1)\int_{x\in\mathbb{R}}\left(\partial_x\psi_\varepsilon\right)^2\psi_\varepsilon^{\lambda-2}dx. \end{aligned} \tag{3.20}$$

Using the (OSLC) inequality (3.15) and the Gronwall lemma in (3.20), we deduce $\forall t \leq T$

$$\|\psi_\varepsilon(\cdot,t)\|_{L^\lambda} \leq \|\psi_\varepsilon(\cdot,0)\|_{L^\lambda} \cdot \exp\left\{-\frac{\lambda-1}{\lambda}\int_t^T m(T-\tau)d\tau,\right\}. \tag{3.21}$$

We pass to the limit when λ tends to $+\infty$ in (3.21). By the definition of ψ_ε, we deduce (3.18). □

3.3. Convergence of the method. Now we prove that the artificial viscosity method converges in the sense that the sequences w^ε and p^ε converge, respectively, to the entropy solution of (1.1)—which is logical—and to the reversible solution of the adjoint equation. Moreover, the sequence of the derivatives of $\widetilde{J}^\varepsilon$ also converges toward an element of the subdifferential of $\widetilde{J}$. Using these BV and $W^{1,\infty}$ estimates we have the following convergence result concerning the adjoint state.

THEOREM 3.4. *We consider the solution $p^\varepsilon(x,t)$ of the linear parabolic problem* (3.13). *We suppose that the flux f satisfies $\beta \geq f'' \geq \alpha > 0$ and p^T is a function of $W^{1,\infty}(\mathbb{R}) \cap BV(\mathbb{R})$. Then*

$$p^\varepsilon \;\to\; p \qquad \textit{uniformly in } \bar{\Omega}, \tag{3.22}$$

where $\Omega = \omega \times (0,T)$, ω is a compact of $\mathbb{R}$, and p is the reversible solution of (2.10) *given by Theorem* 2.5.

Proof. The functions p^ε and $\partial_x p^\varepsilon$ are bounded in L^∞ (by (3.16) and (3.18)). We can extract a subsequence, still denoted by p^ε, and we have

$$p^\varepsilon(\cdot,t) \;\rightharpoonup\; p(\cdot,t) \qquad \text{in } W^{1,\infty}(\mathbb{R})\text{-weak}* . \tag{3.23}$$

By the Rellich–Kondrachov theorem (see Adams [1]) we deduce the strong convergence in $C^{0,\alpha}(\bar{\omega})$ for all $0 < \alpha < 1$, and ω compact set of $\mathbb{R}$, and using a classical diagonalization argument (see, for instance, [6]), we obtain the uniform convergence in $\bar{\Omega}$.

Now, in order to prove that p is a Lipschitz continuous solution of (2.10), we multiply the backward parabolic equation (3.13) by a test function $\varphi \in C_0^\infty$ $(\mathbb{R} \times (0,+\infty))$, and we integrate by parts. We have

$$\begin{aligned} 0 = &-\int_0^\infty \int_{\mathbb{R}} \{p^\varepsilon \partial_t \varphi - f'(w^\varepsilon)\partial_x p^\varepsilon \varphi - \varepsilon \partial_x p^\varepsilon \partial_x \varphi\}\, dxdt \\ &- \int_{\mathbb{R}} p^T(x)\varphi(x,0)dx. \end{aligned} \tag{3.24}$$

On the other hand, multiplying (3.13) by $p(\cdot,t)$, integrating by parts, and using Gronwall's lemma, we deduce that

$$\varepsilon^{\frac{1}{2}} \|\partial_x p^\varepsilon\|_{L^2(\mathbb{R}\times(0,T))} \leq C, \tag{3.25}$$

where C is a constant independent of ε. That implies

$$\varepsilon \partial_{xx}^2 p^\varepsilon \to 0 \quad \text{in } L^2(0,T,H^{-1}(\mathbb{R}))\text{-strong}. \tag{3.26}$$

Now, we know by Theorem 3.1 that $w^\varepsilon \to w$ in $L^1_{\text{loc}}(\mathbb{R} \times (0,+\infty))$-strong, and by Lebesgue's dominated convergence theorem we have $f'(w^\varepsilon) \;\to\; f'(w)$, in $L^1_{\text{loc}}(\mathbb{R} \times (0,+\infty))$-strong. Using the convergence $W^{1,\infty}$-weak$*$ of p^ε (3.23), we deduce

$$f'(w^\varepsilon)\partial_x p^\varepsilon \rightharpoonup f'(w)\partial_x p \quad \text{in } \mathcal{D}'(\mathbb{R} \times (0,+\infty)). \tag{3.27}$$

Using convergence results (3.26), (3.27), and the uniform convergence of p^ε, we let $\varepsilon \to 0$ in (3.24). We obtain

$$0 = -\int_0^\infty \int_{\mathbb{R}} \{p\partial_t\varphi - f'(w)\partial_x p\varphi\}\, dxdt - \int_{\mathbb{R}} p^T(x)\varphi(x,0)dx.$$

Thus the limit of p^ε is a solution of the backward transport equation (2.10) in the sense of distributions. In keeping with the $W^{1,\infty}$-weak$*$ convergence (3.23), we obtain that the limit p verifies $p(\cdot,t) \in W^{1,\infty}(\mathbb{R})\ \forall t \leq T$. On the other hand, by letting $\varepsilon \to 0$ in (3.18), we obtain inequality (2.16).

In order to prove the convergence of the whole sequence p^ε, we will prove that the limit of any converging subsequence is the unique reversible solution, which we denote p_r. For that, we suppose that $p^\varepsilon \to p$, and using the definition (2.15) of the support of exceptional solutions, at first we have

$$p = p_r \quad \text{almost everywhere (a.e.) } (x,t) \in \mathbb{R} \times (0,T) \setminus \mathcal{V}. \tag{3.28}$$

We set $\psi_\varepsilon(x,t) = \partial_x p^\varepsilon(-x, T-t)$ and $\psi_r(x,t) = \partial_x p_r(-x,T-t)$. Then, we substract (2.10) from (3.13) and we differentiate with respect to x. We obtain

$$\partial_t(\psi_\varepsilon - \psi_r) = -\partial_x\left(a_\varepsilon(\psi_\varepsilon - \psi_r)\right) - \partial_x\left((a_\varepsilon - a)\psi_r\right) + \varepsilon\partial_{xx}\psi_\varepsilon,$$

where $a_\varepsilon = f'(w^\varepsilon(-x,T-t))$ and $a = f'(w(-x,T-t))$. We multiply this last equation by $2(\psi_\varepsilon - \psi_r)$, and we integrate by parts. Using the (OSLC) inequality (3.15), we deduce

$$\begin{aligned}\frac{d}{dt}\left\{\|\psi_\varepsilon(\cdot,t) - \psi_r(\cdot,t)\|^2_{L^2(\mathbb{R})}\right\} &+ \varepsilon\|\partial_x\psi_\varepsilon(\cdot,t)\|^2_{L^2(\mathbb{R})} \\ &\leq m(T-t)\|\psi_\varepsilon(\cdot,t) - \psi_r(\cdot,t)\|^2_{L^2(\mathbb{R})} \\ &\quad - \int_{-\infty}^{+\infty}(a_\varepsilon - a)(\psi_\varepsilon - \psi_r)\partial_x\psi_r dx \\ &\quad - \int_{-\infty}^{+\infty}(\partial_x a_\varepsilon - \partial_x a)(\psi_\varepsilon - \psi_r)\psi_r dx \\ &\quad -\varepsilon\int_{-\infty}^{+\infty}\partial^2_{xx}\psi^\varepsilon\psi_r dx.\end{aligned} \tag{3.29}$$

From (3.23), (3.22), (3.26), and (3.27), we have that the second and fourth terms on the right-hand side converge to 0 when $\varepsilon \to 0$. On the other hand, from property (i) of Theorem 2.5 we have that $\psi_r = 0$ in $\mathcal{V}_t = \{x \mid (-x, T-t) \in \mathcal{V}\}$, and using the equality (3.28), we deduce

$$\int_{-\infty}^{+\infty}(\partial_x a_\varepsilon - \partial_x a)(\psi_\varepsilon - \psi_r)\psi_r dx = \int_{x\in\mathbb{R}\setminus\mathcal{V}_t}(\partial_x a_\varepsilon - \partial_x a)(\psi_\varepsilon - \psi_r)\psi_r dx \ \to\ 0,$$

when $\varepsilon \to 0$. Then, the third term on the right-hand side converges to 0. We can pass to the limit in (3.29) and using Gronwall's lemma, we obtain $\lim_{\varepsilon\to 0}\psi_\varepsilon = \psi_r$ and consequently $p = p_r$. □

Remark. Another way to prove this last result is to make use of the characterization (ii) of Theorem 2.5. Indeed, standard arguments allow us to prove that if $\partial_x p^T \geq 0$, then $\partial_x p^\varepsilon \geq 0$. Now, for any final data p we rewrite $p^T = p_1^T - p_2^T$, with $\partial_x p_1^T = (\partial_x p^T)_+$ and $\partial_x p_2^T = (\partial_x p^T)_-$. We denote by p_i^ε the solution of (3.13) with final data p_i^T. Therefore we have $p^\varepsilon = p_1^\varepsilon - p_2^\varepsilon$, with $\partial_x p_i^\varepsilon \geq 0$. We know that $p^\varepsilon \to p$,

$p_i^\varepsilon \to p_i$ with $\partial_x p_i \geq 0$, and p_i is solution of (2.10). Thus p is reversible. A similar monotonicity argument will be used for numerical schemes.

Now we turn to the convergence of the derivative. A fundamental consequence of the convergence of the solution of the adjoint problem with artificial viscosity is the convergence of $D\widetilde{J}^\varepsilon$ when $\varepsilon \to 0$. More precisely, we have the following theorem.

THEOREM 3.5. *Let $\widetilde{J}^\varepsilon : f \mapsto J(w_f^\varepsilon)$ be the cost function* (3.2) *defined for all Lipschitz continuous functions f, with w_f^ε the solution of the parabolic problem* (3.13), *and assume that w^0 has compact support. We suppose that the flux f satisfies $\beta \geq f'' \geq \alpha > 0$ and p^T satisfies a function of $W^{1,\infty}(\mathbb{R}) \cap BV(\mathbb{R})$. Then we have, if w_f is the entropy solution of* (1.1) *and p_f the reversible solution of* (2.10),

$$D\widetilde{J}^\varepsilon(f)\delta f \to \int_0^T \int_{-\infty}^{+\infty} \partial_x p_f \delta f(w_f) dxdt, \qquad \textit{when } \varepsilon \to 0 \tag{3.30}$$

for all Lipschitz direction δf.

Proof. From Theorem 3.4 and the continuity of δf, we obtain the result

$$\left\{ \begin{array}{rcl} \delta f(w_f^\varepsilon) & \to & \delta f(w_f) \quad \text{in } L^1_{\text{loc}}(\mathbb{R} \times (0, +\infty))\text{-strong}, \\ \|\delta f(w_f^\varepsilon)\|_{L^\infty(\mathbb{R}\times(0,+\infty))} & \leq & C \end{array} \right. \tag{3.31}$$

for some constant C independent of ε. Let $\theta \in C_0^\infty(\mathbb{R} \times (0, +\infty))$. In keeping with the convergence $W^{1,\infty}$-weak of p^ε (3.23), and using (3.31), we obtain

$$\int_0^T \int_{-\infty}^{+\infty} \theta \partial_x p_f^\varepsilon \delta f(w_f^\varepsilon) dxdt \to \int_0^T \int_{-\infty}^{+\infty} \theta \partial_x p_f \delta f(w_f) dxdt. \tag{3.32}$$

By hypothesis, w^0 has a compact support. It is known that for finite propagation velocity, the function $[\partial_x p_f]\delta f(w_f)$ stays in a compact support, for $T < \infty$ (see Kružkov [12]). We have

$$\begin{aligned} &\left| D\widetilde{J}^\varepsilon(f)\delta f - \int_0^T \int_{-\infty}^{+\infty} \partial_x \text{p} \delta f(w_f) \right| \\ &\qquad \leq \int_0^T \int_{-R}^{R} |\partial_x p_f^\varepsilon \delta f(w_f^\varepsilon) - \partial_x p_f \delta f(w_f)| \\ &\qquad + \int_0^T \int_{|x|>R} |\partial_x p_f^\varepsilon \delta f(w_f^\varepsilon)| \end{aligned} \tag{3.33}$$

for R large enough. Hence, by (3.32), the first term of the right-hand side in (3.33) converges to 0 when $\varepsilon \longrightarrow 0$ for all R large enough.

Let us prove the convergence of the second term. We deduce from the compact support of w^0 and from the maximum principle applied to the linear parabolic equation (3.13) that $|p^\varepsilon(x,t)| \leq C \exp(-r|x|)$, where C, r are constants which depend only on T and $\|w^0\|_{L^\infty}$ and thus are independent of ε. We deduce that the second term of the right-hand side of (3.33) uniformly converges to 0 in ε, when $R \to \infty$. This concludes the proof of (3.30). □

If the derivative of $\widetilde{J}(f)$ exists, then $D\widetilde{J}(f)$ is characterized by (2.12), and we deduce from Theorem 3.5 that $D\widetilde{J}^\varepsilon(f)\delta f \to D\widetilde{J}(f)\delta f$, when $\varepsilon \to 0$. The trouble is that we have no result concerning the differentiability of $\widetilde{J}$. Nevertheless, since we are interested in the behavior of $\widetilde{J}$ near a minimum, we can assume that $\widetilde{J}$ is convex in a neighborhood of this point. Therefore, we can define its subdifferential $\partial\widetilde{J}(f)$.

COROLLARY 3.6. *Assume that $\widetilde{J}$ is a minimum at f and that $\widetilde{J}$ and $\widetilde{J}^\varepsilon$ are convex in a neighborhood of f for all ε. Then, under the hypotheses of Theorem* 3.5*, $D\widetilde{J}^\varepsilon(f)\delta f$ converges to an element of the subdifferential $\partial\widetilde{J}(f)$ of $\widetilde{J}(f)$, when $\varepsilon \to 0$, that is,*

$$\partial\widetilde{J}(f) \ni \int_0^T \int_{-\infty}^{+\infty} \partial_x p_f \delta f(w_f) dx dt.$$

Proof. From the definition of convexity we have

$$D\widetilde{J}^\varepsilon(f)f - \widetilde{J}^\varepsilon(f) \geq D\widetilde{J}^\varepsilon(f)\nu - \widetilde{J}^\varepsilon(\nu)$$

for all Lipschitz continuous function ν. We apply Theorem 3.5, and we pass to the limit when $\varepsilon \to 0$. We obtain

$$\int_0^T \int_{-\infty}^{+\infty} [\partial_x p_f] \left[(f-\nu)(w_f)\right] dx dt \geq \widetilde{J}(f) - \widetilde{J}(\nu).$$

That is a characterization of the subgradient for convex functions (see Rockafellar [15]). Thus, the limit of sequence $D\widetilde{J}^\varepsilon(f)\delta f$ is an element of the set $\partial J(f)$. □

4. Numerical approximation. Now we shall give similar convergence results for discretization of the identification problem, and we will remark that at the limit, both approximations (artificial viscosity and discretization) reach the same element of the subdifferential $\partial J(f)$ characterized by the reversible solution of the adjoint problem. We shall prove these results for a large class of numerical schemes which contains the schemes used to resolve the identification problem in [9].

First we discretize the cost function (denoted J_Δ) and the direct problem (1.1). Next we compute a discrete Lagrangian, which will lead to a discrete adjoint state, and finally to a discrete gradient of J_Δ. This method of computing the exact gradient of the discretized problem seems to have better properties (concerning stability, for instance) than discretizing the exact adjoint state. Moreover, notice that we have no natural way to discretize it since the adjoint equation is ill-posed.

4.1. Discretization and convergence for the direct problem. Let Δz (respectively, Δt) be a positive space (respectively, time) step. These parameters will tend to 0, the ratio $\lambda = \Delta t/\Delta z$ remaining constant. For $n = 0, \dots, N$, $j = 0, \dots, J$, the sequence w_j^n is an approximation of solution w at the point $(z_j = j\Delta z, t_n = n\Delta t)$. In the same way, we discretize $w^0(z)$, $w^{\text{obs}}(z)$, by w_j^0, w_j^{obs}, respectively. We consider a conservative $(2K+1)$-points scheme for the hyperbolic equation (1.1)

$$w_j^{n+1} = w_j^n - \lambda \left\{ g_{j+\frac{1}{2}}^n(f) - g_{j-\frac{1}{2}}^n(f) \right\}, \tag{4.1}$$

where $g_{j+\frac{1}{2}}^n(f) = g_f(w_{j-K+1}^n, \dots, w_{j+K}^n)$, g_f being the numerical flux of the scheme, consistent with f: $g_f(w, \dots, w) = f(w)$. Then the discretized identification problem becomes the following minimization problem:

$$\min_f J\left(w_f^\Delta\right), \tag{4.2}$$

where w_f^Δ is the piecewise constant function defined by the sequence $\left\{w_j^n(f)\ ;\ j = 0, \dots, J-1,\ n = 0, \dots, N-1\right\} \in \mathbb{R}^M$, which was built out of (4.1).

To obtain the exact gradient of the discretized cost function, we follow along exactly the same lines as in the formal computation and viscous regularization, that is, we build up a discrete Lagrangian $\mathcal{L}_\Delta$ using a "discrete weak form" of the direct scheme (4.1), then we differentiate with respect to w_j^n, and we choose the sequence p_j^n in order to cancel $\partial\mathcal{L}_\Delta/\partial w_j^n$ for all j and n. This defines the adjoint scheme. Finally, for the discrete gradient, the computations give

$$D\widetilde{J}_\Delta(f)\delta f = \, -\Delta t \sum_{n,j} \left(p_j^{n+1} - p_{j+1}^{n+1}\right) Dg_{j+\frac{1}{2}}^n(f)\delta f, \tag{4.3}$$

where p_Δ is the solution of the adjoint scheme

$$\begin{cases} p_j^n = p_j^{n+1} - \lambda \displaystyle\sum_{k=-K}^{K-1} \frac{\partial}{\partial w_j^n} g_{j+k+\frac{1}{2}}^n (p_{j+k}^{n+1} - p_{j+k+1}^{n+1}), \\ \\ p_j^N = \dfrac{\partial J\left(w^\Delta\right)}{\partial w_j^N}. \end{cases} \tag{4.4}$$

The complete computations are rather tedious, and we refer to [9] or [17] for greater detail and for some examples as well.

First we show a convergence result of our discretized cost function, which allows us to say that the continuous identification problem can be approximated by the discretized identification problem. We consider a suitable set of functions $\mathcal{F}$, which we suppose are bounded and closed for the Lipschitz continuous norm $\| \, \|_{\mathrm{Lip}}$. Next we suppose that the numerical flux g introduced in (4.1) is Lipschitz continuous with respect to $w_{-k}, \dots, w_k$ and independent of $f \in \mathcal{F}$, i.e., that there exists a constant C_{Lip} independent of f, such that

$$\sup_{w,v\geq 0} \left\{ \frac{|g_f(w) - g_f(v)|}{|w-v|} \right\} \leq \; C_{\mathrm{Lip}} \qquad \forall f \in \mathcal{F}. \tag{4.5}$$

We also suppose that the scheme (4.1) satisfies

$$|w_j^n(f)| \leq C_\infty \qquad \forall\, j, n, \qquad \text{and for any } f \in \mathcal{F}, \tag{4.6}$$

where C_∞ is a constant independent of f. Condition (4.6) is verified for the Lax–Friedrichs, Godunov [7], and Van Leer [20] schemes, when the following CFL-condition is satisfied:

$$\lambda \sup_{\substack{w \,\geq\, 0 \\ f \,\in\, \mathcal{F}}} |f'(w)| < 1. \tag{4.7}$$

This leads us to our convergence result.

Proposition 4.1. *Let $w_{f_\Delta}^\Delta$ be built out of the conservative scheme* (4.1) *which we suppose consistent with* (1.1)*, and verifying hypotheses* (4.5) *and* (4.6)*, f_Δ being the solution of* (4.2)*. If the initial condition w_Δ^0 is bounded in $L^\infty(\mathbb{R}) \cap BV(\mathbb{R})$, then $w_{f_\Delta}^\Delta$ is bounded in $L^\infty(\mathbb{R}\times(0,+\infty))$ and in $L^\infty(0,T,\, BV\ (\mathbb{R}))$ for all $T>0$. Furthermore, if $f_{\Delta_k} \to \bar{f} \in \mathcal{F}$, for the Lipschitz continuous norm $\| \, \|_{\mathrm{Lip}}$, then*

$$w_{f_{\Delta_k}}^{\Delta_k} \to w_{\bar{f}} \quad \textit{in } L^\infty(0,T,L^1(0,L))\textit{-strong}, \tag{4.8}$$

$$w_{f_{\Delta_k}}^{\Delta_k} \rightharpoonup w_{\bar{f}} \quad \textit{in } L^\infty(\Omega)\textit{-weak}\,*\,. \tag{4.9}$$

Remarks. We can prove this proposition thanks to the continuity result of Lucier [13] and by copying the proof of a convergence result for schemes approximating scalar conservation laws in [6]. We omit the detail of this proof.

For instance, we consider the identification problem arising from the chromatographic model (see [9]), and we take a bounded subset of parameters $\mathcal{K} \in \mathbb{R}^N$. Then we deduce the CFL-condition (4.7), and we have at least one accumulation point of the sequence $\{f_\Delta\}_{\Delta x, \Delta t}$.

We suppose that w^{obs} and w^0 have a compact support, so that for $0 < t < T$, the support of the solution w_f to (1.1), with $f \in \mathcal{F}$, is in a compact set $\Omega = (0, L) \times (0, T)$. Proposition 4.1 and a convergence result of [9] imply the following.

COROLLARY 4.2. *Let $w^\Delta_{f_\Delta}$ be the solution of a conservative and TVD scheme* (4.1)*, consistent with* (1.1)*. Then any accumulation point $\bar{f} \in \mathcal{F}$ of the sequence $\{f_\Delta\}_{\Delta x, \Delta t}$, for the Lipschitz continuous norm is a solution of*

$$J(w_f) = \min_{g \in \mathcal{F}} J(w_g),$$

where w_f is the entropy solution of (1.1)*, with $f \in \mathcal{F}$.*

4.2. Monotone and TVD adjoint schemes. Here we study some properties of monotonicity and BV estimates for adjoint schemes associated with the schemes in conservative form (4.1). We notice that adjoint schemes cannot be put in conservative form, which is not surprising, since the adjoint equation is not conservative. However, we shall prove that a family of TVD difference schemes, including the Godunov and the Van Leer schemes, is associated with TVD adjoint schemes.

First we define the function $\delta_x p_\Delta$ by

$$\delta_x p_\Delta(x, t) = \frac{p_\Delta(x, t) - p_\Delta(x + \Delta x, t)}{\Delta x},$$

where $\delta_x p_\Delta$ is the piecewise constant function with value $\Delta p^n_{j+\frac{1}{2}} / \Delta x$ in the rectangle

$$\Pi^n_j = \left(\left(j - \frac{1}{2} \right) \Delta x, \left(j + \frac{1}{2} \right) \Delta x \right) \times (n\Delta t, (n+1)\Delta t).$$

On the other hand, we define the differences $\Delta p^n_{j+\frac{1}{2}} = p^n_j - p^n_{j+1}$ for all $j \in \mathbf{Z}$, $n \leq N$. Using the linearity of the scheme (4.4), we deduce the following scheme for $\Delta p^n_{j+\frac{1}{2}}$:

$$\Delta p^n_{j+\frac{1}{2}} = \sum_{k=-K}^{K} A^k_{j+k} \Delta p^{n+1}_{j+k+\frac{1}{2}}, \tag{4.10}$$

where the coefficients A^k_j are defined by

$$\begin{cases} A^k_j = -\lambda \left(\dfrac{\partial}{\partial w^n_{j-k}} - \dfrac{\partial}{\partial w^n_{j-k+1}} \right) g^n_{j+\frac{1}{2}} \quad \text{for } k \notin \{-K, 0, K\}, \\ A^{-K}_j = -\lambda \dfrac{\partial}{\partial w^n_{j+K}} g^n_{j+\frac{1}{2}}, \qquad A^K_j = \lambda \dfrac{\partial}{\partial w^n_{j-K+1}} g^n_{j+\frac{1}{2}}, \\ A^0_j = 1 - \lambda \left(\dfrac{\partial}{\partial w^n_j} - \dfrac{\partial}{\partial w^n_{j+1}} \right) g^n_{j+\frac{1}{2}}. \end{cases} \tag{4.11}$$

Notice that by construction the coefficients A^k_j satisfy

$$\sum_{k=-K}^{K} A^k_j = 1 \quad \forall j \in \mathbf{Z}. \tag{4.12}$$

Now we wish the adjoint scheme to have the same property of monotonicity preservation as the continuous equation. Thus, in view of (4.10), it is natural to impose

$$A_j^k \geq 0 \quad \text{for } -K \leq k \leq K,\ j \in \mathbf{Z}. \tag{4.13}$$

This is somehow the discrete version to the (OSLC) condition (3.15) used for the convergence of the viscous perturbation.

Example. In the model of chromatography, the Godunov scheme is very simple. (It is just an upwind scheme.) The adjoint scheme is given by (see [9])

$$p_j^n = p_j^{n+1} - \lambda f'(w_j^n)\left(p_j^{n+1} - p_{j+1}^{n+1}\right). \tag{4.14}$$

In this case, the coefficients A_j^k are

$$A_j^{-1} = \lambda f'(w_j^n), \qquad A_j^0 = 1 - \lambda f'(w_j^n), \qquad A_j^1 = 0.$$

The CFL-condition (4.7) implies that these coefficients are positive.

Similarly, the adjoint scheme associated with the Van Leer difference scheme verifies the hypothesis $A_j^k \geq 0$ when we have CFL-condition (4.7), and it is a monotone and TVD scheme, whereas we can see that the adjoint scheme associated with the Lax–Friedrichs difference scheme does not verify this hypothesis of positivity and that it is an unstable scheme in $BV(\mathbb{R})$.

We have the following a priori estimates.

PROPOSITION 4.3. *We consider a linear scheme in form* (4.4)*, with its coefficients defined by* (4.11) *and verifying the monotonicity property* (4.13)*. Then we have for all* $0 < s, t < T$ *the following estimates:*

$$TV(p_\Delta(.,t)) \leq TV(p_\Delta(.,T)), \tag{4.15}$$

$$\|p_\Delta(.,t)\|_{L^1(\mathbb{R})} \leq \|p_\Delta(.,T)\|_{L^1(\mathbb{R})} + CT\ TV(p_\Delta(.,T)), \tag{4.16}$$

$$\|\delta_x p_\Delta(\cdot,t)\|_{L^\infty(\mathbb{R})} \leq \|\delta_x p_\Delta(\cdot,T)\|_{L^\infty(\mathbb{R})}, \tag{4.17}$$

$$\|p_\Delta(.,t) - p_\Delta(.,s)\|_{L^p(\omega)} \leq C\frac{|t-s| + \Delta t}{|\Delta t|^{\frac{p-1}{p}}}\|\delta_x p_\Delta(\cdot,T)\|_{L^\infty(\mathbb{R})}, \tag{4.18}$$

where ω *is a compact set in* $\mathbb{R}$*,* $1 \leq p \leq +\infty$*, and* $C > 0$ *is a constant independent of the discretization.*

Remark. Estimates (4.15) and (4.16) are classical results of the theory of approximation of nonlinear conservation laws. They rely upon monotonicity properties of the schemes and are in particular independent of the conservative form. The estimate (4.18) is a technical result allowing us to use a diagonalization argument in order to pass to the limit, when $\Delta t, \Delta z \to 0$.

Notice that the result of estimate (4.17) bears some resemblance to that of the $W^{1,\infty}$ estimate proved in Theorem 3.3.

Proof of Proposition 4.3. Let $R \geq K$. Summing up (4.10) from $j = -R$ to $j = R$, and using the positivity of A_j^k and property (4.12), and passing to the limit $R \to +\infty$, we obtain the estimate (4.15). On the other hand, using this BV estimate, the L^1 estimate (4.16) is obtained by an analysis that is similar to the one made by Godlewski and Raviart [6, Chap. III] for the different schemes in conservative form. Now, using the condition $A_{j+k}^k \geq 0$, the estimate (4.17) results from the monotonicity of

$$(H_\Delta(\Delta p))_j = \sum_{k=-K}^{K} A_{j+k}^k \Delta p_{j+k+\frac{1}{2}}^{n+1}.$$

In order to obtain the next estimate (4.18), we write the scheme (4.4) in the form

$$p_j^n - p_j^{n+1} = -\lambda \sum_{k=-K}^{K-1} \frac{\partial}{\partial w_j^n} g_{j+k+\frac{1}{2}}^n \Delta p_{j+k+\frac{1}{2}}^{n+1},$$

and by the Lipschitz condition of g, we have

$$|p_j^n - p_j^{n+1}| \leq \lambda C \sup_{j\in\mathbf{Z}} |\Delta p_{j+\frac{1}{2}}^{n+1}|,$$

with a constant C independent of the discretization. That implies

$$\|p_\Delta(\cdot, n\Delta t) - p_\Delta(\cdot,(n+1)\Delta t)\|_{L^p(\omega)} \leq \lambda^{\frac{p-1}{p}} C|\omega|^{\frac{1}{p}}|\Delta t|^{\frac{1}{p}} \sup_{j\in\mathbf{Z}} |\Delta p_{j+\frac{1}{2}}^{n+1}|.$$

Let $m > n$. Applying successively this same result for $n, n+1, \ldots, m-1$, and using the triangular inequality of the norm $\|\cdot\|_{L^p}$, we obtain

$$\|p_\Delta(\cdot, n\Delta t) - p_\Delta(\cdot, m\Delta t)\|_{L^p(\omega)} \leq C|\Delta t|^{\frac{1}{p}}(n-m) \sup_{j\in\mathbf{Z}} |\Delta p_{j+\frac{1}{2}}^{N}|. \tag{4.19}$$

Let s, t such that $n\Delta t \leq s \leq (n+1)\Delta t$, $m\Delta t \leq t \leq (m+1)\Delta t$. We notice that $(m-n)\Delta t \leq |t-s| + \Delta t$. Inequality (4.19) gives the desired result (4.18). □

4.3. Convergence for adjoint schemes and derivatives. Now we shall provide evidence that the sequence of discrete gradients converges to the same element of the subdifferential of the cost function given by the limit of the viscous perturbation. First we have the following convergence result for the solution of the adjoint scheme.

THEOREM 4.4. *We consider the linear difference scheme in the form* (4.4)*, where the coefficients are defined in* (4.11) *and verify the monotonicity property* (4.13)*. We suppose* $p^T \in W^{1,\infty}(\mathbb{R}) \cap BV(\mathbb{R})$ *and the (OSLC)* (3.15)*. Then,*

$$p_\Delta \;\to\; p \quad \textit{in } L^\infty(0,T,L^q_{\mathrm{loc}}(\mathbb{R}))\textit{-strong}, \qquad 1 \leq q < +\infty,$$

where p *is the reversible solution of* (2.10)*.*

Proof. In accordance with the estimate (4.17) and in keeping with the hypothesis $p^T \in W^{1,\infty}(\mathbb{R})$, we have

$$\sup_x |\delta_x p_\Delta(x,t)| \leq C. \tag{4.20}$$

Applying the theorem of Riesz–Fréchet–Kolmogorov (see Adams [1]) in the last inequality, we deduce that we can take a subsequence $\Delta_k x, \Delta_k t \to 0$ which we still denote with $\Delta x, \Delta t$, such as

$$p_\Delta(\cdot,t) \;\to\; p(\cdot,t) \quad \text{in } L^q_{\mathrm{loc}}(\mathbb{R}))\text{-strong} \quad \forall t \in (0,T)$$

for all $1 \leq q < +\infty$. Using the estimate (4.18) and a classical diagonalization argument (see [6]), we deduce the convergence in $L^\infty(0,T,L^q_{\mathrm{loc}}(\mathbb{R}))$.

In order to prove that the limit p is a Lipschitz continuous solution of (2.10), we proceed by similar arguments to the proof of Theorem 3.4. Next we notice that the scheme in the form (4.4) preserves monotonicity as soon as $A_j^k \geq 0$ (which is satisfied, for instance, by the adjoints of Godunov and Van Leer schemes). Finally, we use the same arguments of the first remark following the proof of Theorem 3.4. That is,

$$\begin{cases} p_\Delta \to p \text{ in } L^\infty(0,T,L^q_{\mathrm{loc}}(\mathbb{R})), \\ p_\Delta = p_\Delta^1 - p_\Delta^2, \text{ with } (p_j^n)^i \leq (p_{j+1}^n)^i, \text{ and } p_\Delta^i \text{ solution of the scheme (4.4)} \end{cases}$$

implies that p is reversible. □

It is easy to verify that $\widetilde{J}_\Delta : f \mapsto J(w_f^\Delta)$ has a continuous derivative if the numerical flux g is of class C^1 with respect to $(w_{j-K+1}^n, \cdots, w_{j+K}^n)$. As a consequence of Theorem 4.4, we have the following.

COROLLARY 4.5. *Let $\widetilde{J}_\Delta : f \mapsto J(w_f^\Delta)$ be a discretized cost function defined for all Lipschitz continuous function f, with w_f^Δ solution of the scheme in the conservation form* (4.1). *We suppose that the coefficients A_j^k in* (4.11) *satisfy $A_j^k \geq 0$ for $-K \leq k \leq K$, $j \in \mathbf{Z}$, and that w^0 has a compact support. Furthermore, we suppose that $p^T \in W^{1,\infty}(\mathbb{R}) \cap BV(\mathbb{R})$. Let w_f be the entropy solution of* (1.1) *and p_f be the reversible solution of* (2.10). *Then we have*

$$D\widetilde{J}_\Delta(f)\delta f \to \int_0^T \int_{-\infty}^{+\infty} \partial_x p_f \delta f(w_f) dx dt, \qquad \text{when } \Delta x, \Delta t \to 0 \tag{4.21}$$

for all Lipschitz direction δf.

Remark. We note that Theorem 3.5 and Corollary 4.5 make it clear that a viscous perturbation or a numerical approximation of the gradient give the same result at the limit.

We clarify this result by another corollary, as follows.

COROLLARY 4.6. *Assume that $\widetilde{J}$ is a minimum at f and that $\widetilde{J}$ and $\widetilde{J}_\Delta$ are convex for all $(\Delta x, \Delta t)$ in a neighborhood of f. Then, under the hypotheses of Theorem* 4.5, *$D\widetilde{J}_\Delta(f)\delta f$ converges to the same element of the subdifferential $\partial J(f)$ of $J(f)$ which we have obtained by a viscous perturbation in Corollary* 3.6, *when $\Delta x, \Delta t \to 0$.*

5. Numerical results. In this section we illustrate our results by a numerical application on real experimental data. We consider the propagation of a single, pure compound in a column, which leads under several physical assumptions to a scalar conservation law of the form (1.1). More precisely, the experimental data are concentration profiles obtained from the adsorption of gaseous n-hexane on graphite carbon with helium vector gas. We refer to Rouchon et al. [16] for the complete description of the experiment, the discussion of the model, and the original results. A remarkable feature of this experiment is that we have an experimentally identified flux with which to compare.

The observation here is a profile of concentration versus time, at a fixed $L > 0$, where L is the length of the column. This is not quite the context of the previous analysis, but since in this kind of model the flux satisfies $f'(w) > 0$, only a slight modification is needed. The direct problem is discretized through a Godunov scheme, and we compute the exact gradient of the discrete functional as indicated in section 4.1 (see also [9] for details). At this discrete level, all the quantities are well defined, so we are able to apply a gradient-based minimization algorithm. Thus we are left with the problem of nonuniqueness for f which was mentioned before. To handle this, we specify an analytic form for f, based on physical arguments, which we do not detail here. We refer the interested reader to [10] and the references therein for more complete information. If c denotes the concentration of the compound, then a family of fluxes depending on a finite number of parameters is given by

$$f(c) = N \frac{KcP'(Kc)}{qP(Kc)}, \quad P(w) = \sum_{i=0}^{q} \alpha_i \exp(-\beta E_i) w^i, \tag{5.1}$$

where $q \geq 1$ is an integer and α_i and β are given constants involving temperature and other fixed parameters of the experiment. The relevant parameters to identify here are the so-called Langmuir coefficient K, N, and E_i, $2 \leq i \leq q$ ($E_0 = E_1 = 0$ by

TABLE 5.1
Initial and identified coefficients (1).

Parameters	Initial guess (Langmuir)	Identification result
K	1596.	1178.
E_2	0.	-806.
E_3	0.	-213.
E_4	0.	-16.

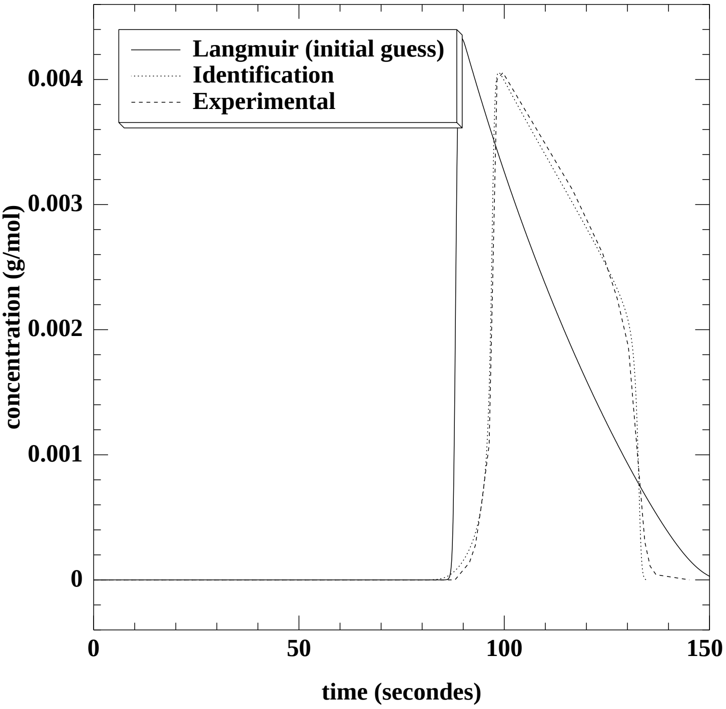


FIG. 5.1. *Comparison between experimental and identified profiles.*

construction). The model obtained with $q = 1$ or equivalently with $E_i = 0\ \forall i$ is the classical Langmuir model.

These parameters do not have the same influence on the concentration profiles. Roughly speaking, K and N act essentially on the position of the profile, while E_i modifies its shape. The minimization algorithm has therefore to be carefully modified to handle this problem (see [9]). In the first application, we chose as an initial guess the Langmuir model, with a given value of K, and we fixed the value of $N = 2.19\ 10^{-2}$.

The coefficients of the initial guess and the identification result are given in Table 5.1, and the comparison between the experimental flux, the initial guess, and the identified flux is shown in Figure 5.2. Finally, Figure 5.1 shows the experimental and identified concentration profiles.

These figures deserve a few remarks. The comparison between the profiles in Figure 5.1 proves that our identification is quite successful. Concerning the fluxes themselves, we notice in Figure 5.2 the good agreement with the experiment on the whole domain of measurement, even though the domain of identification is $0 \leq c \leq 0.004$ g./mol.

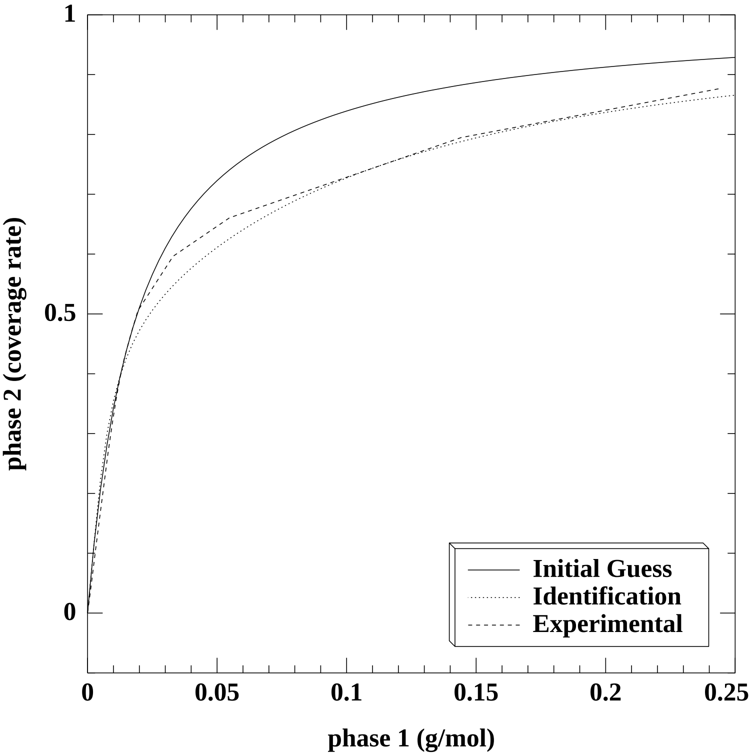


FIG. 5.2. *Comparison between experimental and identified isotherm.*

TABLE 5.2
Initial and identified coefficients (2).

Parameters	Initial guess	Identification result
N	$2.19\ 10^{-2}$	$1.42\ 10^{-2}$
K	1596.	2068.
E_2	214.	214.
E_3	-1252.	-1252.
E_4	1251.	1251.

In order to illustrate the nonuniqueness, we tried another identification. Starting from an initial guess which has a convenient shape, we tried to identify only the coefficients N and K, leaving the ratio K/N constant. The results are shown in Table 5.2. The resulting profile is identical to the one in Figure 5.1.

Figure 5.3 shows the preceding identification (labeled 1) and this one (labeled 2). We notice that, indeed, we obtain different functions but in the domain of identification, they are virtually indistinguishable.

We would like to emphasize that the choice of a physically relevant model for the flux is of great importance to obtain a good agreement with experiment. But even in this case the problem of uniqueness is not solved, and the correct choice of the flux therefore has to rely on physical arguments: The parameters here have a precise meaning, which has not been explored in these experiments. We refer to [11] for similar results on a binary mixture or, in terms of partial differential equations, on a system of two equations.

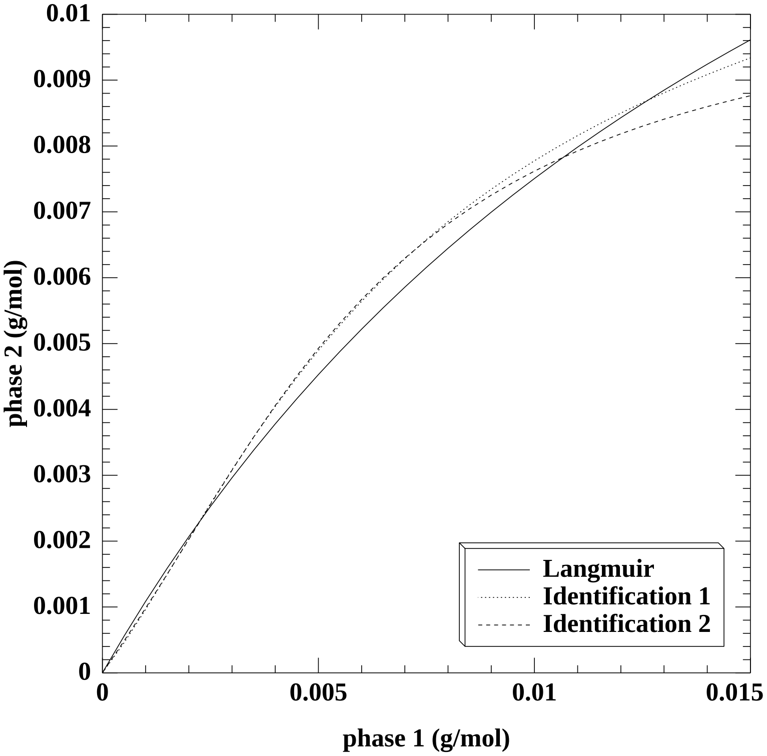


FIG. 5.3. *Comparison between two identified isotherms.*

REFERENCES

[1] R. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] F. BOUCHUT AND F. JAMES, *Équations de transport unidimensionnelles à coefficients discontinus*, C.R. Acad. Sci. Paris Sér. I Math., 320 (1995), pp. 1097–1102.

[3] F. BOUCHUT AND F. JAMES, *One-dimensional transport equations with discontinuous coefficients*, Nonlinear Anal., 32 (1998), pp. 891–933.

[4] G. CHAVENT, *Identification of distributed parameter systems: About the output least square method, its implementation, and identifiability*, in Proc. 5th IFAC Symposium on Identification and System Parameter Estimations, Pergamon Press, Elmsford, NY, Oxford, 1979, pp. 85–97.

[5] E.D. CONWAY, *Generalized solutions of linear differential equations with discontinuous coefficients and the uniqueness question for multidimensional quasilinear conservation laws*, J. Math. Anal. Appl., 18 (1967), pp. 238–251.

[6] E. GODLEWSKI AND P.-A. RAVIART, *Hyperbolic systems of conservation laws*, in Mathématiques et Applications, 3/4, Ellipses, Paris, 1991.

[7] S. K. GODUNOV, *A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics*, Mat. Sb., 47 (1959), pp. 271–290.

[8] D. HOFF, *The sharp form of Oleinik's entropy condition in several space variables*, Trans. Amer. Math. Soc., 276 (1983), pp. 707–714.

[9] F. JAMES, M. SEPÚLVEDA, AND P. VALENTIN, *Parameter identification for a model of chromatographic column*, Inverse Problems, 10 (1994), pp. 1299–1314.

[10] F. JAMES, M. SEPÚLVEDA, AND P. VALENTIN, *Statistical thermodynamic models for multicomponent diphasic isothermal equilibria*, Math. Models Methods Appl. Sci., 7 (1997), pp. 1–29.

[11] F. JAMES, M. SEPÚLVEDA, F. CHARTON, I. QUIÑONES, AND G. GUIOCHON, *Determination of Binary Competitive Equilibrium Isotherms from the Individual Chromatographic Band Profiles*, Technical Report 98-24 DIM, Universidad de Concepción, Concepción, Chile, 1998. Chem. Engrg. Sci., to appear.

[12] S.N. KRUŽKOV, *First order quasilinear equations with several independent variables*, Mat. Sb.

(N.B.), 81 (1970), pp. 228–255 (in Russian).

[13] B. Lucier, *A moving mesh numerical method for hyperbolic conservation laws*, Math. Comp., 46 (1986), pp. 59–69.

[14] O.A. Oleinik, *Discontinuous Solutions of Nonlinear Differential Equations*, Amer. Math. Soc. Transl. Ser. 2 26, AMS, Providence, RI, 1963, pp. 95–172.

[15] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[16] P. Rouchon, M. Schœnauer, P. Valentin, and G. Guiochon, *Numerical simulation of band propagation in nonlinear chromatography*, Separat. Sci. Tech., 22 (1987), pp. 1793–1833.

[17] M. Sepúlveda, *Identification de paramètres pour un système hyperbolique, Application à l'estimation des isothermes en chromatographie*, Thèse de l'Ecole Polytechnique, Palaiseau, France, 1992.

[18] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer–Verlag, New York, 1983.

[19] E. Tadmor, *Local error estimates for discontinuous solutions of nonlinear hyperbolic equations*, SIAM J. Numer. Anal., 28 (1991), pp. 891–906.

[20] B. Van Leer, *Towards the ultimate conservative difference scheme* V, *A second-order sequel to Godunov's method*, J. Comput. Phys., 32 (1979), pp. 101–136.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 3, pp. 892–910

AN AUGMENTED LAGRANGIAN METHOD FOR IDENTIFYING DISCONTINUOUS PARAMETERS IN ELLIPTIC SYSTEMS[*]

ZHIMING CHEN[†] AND JUN ZOU[‡]

Abstract. The identification of discontinuous parameters in elliptic systems is formulated as a constrained minimization problem combining the output least squares and the equation error method. The minimization problem is then proved to be equivalent to the saddle-point problem of an augmented Lagrangian. The finite element method is used to discretize the saddle-point problem, and the convergence of the discretization is also proved. Finally, an Uzawa algorithm is suggested for solving the discrete saddle-point problem and is shown to be globally convergent.

Key words. parameter identification, elliptic system, augmented Lagrangian, finite element method

AMS subject classifications. 65N30, 35R30

PII. S0363012997318602

1. Introduction. The main purpose of this paper is to propose a numerical approach and conduct convergence analyses on each approximation process in the identification of the unknown coefficient q in the elliptic problem

$$-\nabla \cdot (q\nabla u) = f \quad \text{in} \quad \Omega; \quad u = 0 \quad \text{on} \quad \Gamma.$$

The identifying process is carried out so that the solution u matches its observation data z optimally in a certain sense. Here Ω can be any bounded domain in R^d, $d = 1, 2$, or 3, with piecewise smooth boundary Γ and $f \in H^{-1}(\Omega)$ as given. The problem may describe the flow of a fluid (e.g., groundwater) through some medium with permeability $q(x)$, or the heat transfer in a material with conductivity $q(x)$; we refer to the books by Bank and Kunisch [1] and Engl, Hanke, and Neubauer [7]. Practically, it is often easier to measure the solution u at various points in the medium than to measure the parameter $q(x)$ itself [11]. Then the measured data of u (often the interpolated function of the data) are utilized to estimate the parameter $q(x)$ through the above boundary value problem. We study a hybrid method proposed in [13, 14] that combines the output least squares and the equation error formulation within the mathematical framework given by the augmented Lagrangian technique. The augmented Lagrangian methods have been widely used earlier in nonlinear constrained optimization problems and nonlinear boundary value problems to relax some complicated constraints or difficult couplings among some nonlinear and nonsmooth terms or to enhance convexities of the objective functions (cf. [10, 2]). Ito and Kunisch [13, 14] applied the augmented Lagrangian method for parameter identifying problems, incorporated with a regularization term of the H^2 seminorm of the parameters to be estimated. Their methods appear to be very efficient and successful in recovering the smooth parameters. The major novelty of this paper is to generalize the

[*] Received by the editors March 24, 1997; accepted for publication (in revised form) April 1, 1998; published electronically March 17, 1999.

http://www.siam.org/journals/sicon/37-3/31860.html

[†] Institute of Mathematics, Academia Sinica, Beijing 100080, People's Republic of China (zmchen@math03.math.ac.cn). The research of this author was partially supported by the China National Natural Science Foundation.

[‡] Department of Mathematics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (zou@math.cuhk.edu.hk). The research of this author was partially supported by Hong Kong RGC grant CUHK 338/96E.

aforementioned method so that we can identify even nonsmooth parameters. To this aim, we propose to search for the coefficients in the space of functions with bounded variation (BV), namely, in the space

$$BV(\Omega) = \Big\{ q \in L^1(\Omega); \quad \|q\|_{BV(\Omega)} < \infty \Big\}.$$

Here $\|q\|_{BV(\Omega)} = \|q\|_{L^1(\Omega)} + \int_\Omega |Dq|$ with the notation $\int_\Omega |Dq|$ defined by

$$\int_\Omega |Dq| = \sup \Big\{ \int_\Omega q \operatorname{div} g \, dx; \quad g \in \big(C_0^1(\Omega)\big)^d \quad \text{and} \quad |g(x)| \le 1 \quad \text{in} \quad \Omega \Big\},$$

which allows us to identify the discontinuous parameters in elliptic systems.

Because of the involvement of the $BV(\Omega)$ norm in the cost function and because there is not as much regularity as in [13, 14], we cannot apply the techniques of Ito and Kunisch to show the existence of the saddle-points of the augmented Lagrangian and the convergence of the discrete saddle-points to the continuous ones. Instead, our crucial tool for the convergence analyses will be an appropriate application of the Hahn–Banach convex separating theorem. This enables us to have a clear and simple convergence theory without making any a priori assumptions on cost functional or constraint functionals. We note that quite a different approach was used in [12] for the identification of discontinuous parameters.

We now formulate the aforementioned parameter identifying problem as the following constrained minimization problem:

$$\text{(1.1)} \qquad \text{minimize} \quad J(q,v) = \frac{1}{2} \int_\Omega q |\nabla v - \nabla z|^2 dx + \beta \int_\Omega |Dq|$$

$$\text{(1.2)} \qquad \text{subject to} \quad (q,v) \in K \times V \quad \text{and}$$

$$\text{(1.3)} \qquad e(q,v) = (-\Delta)^{-1}(\nabla \cdot (q \nabla v) + f) = 0,$$

where $V = H_0^1(\Omega)$ and K is a subset of the function space $BV(\Omega)$ of BVs defined by

$$K = \{ q \in BV(\Omega); \quad \alpha_1 \le q(x) \le \alpha_2 \quad \text{almost everywhere (a.e.)} \quad \text{in } \Omega \}.$$

Here α_1 and α_2 are two positive constants and $\beta > 0$ is a regularization parameter. $-\Delta$ is the Laplace operator from $H_0^1(\Omega)$ to its dual space $H^{-1}(\Omega)$, so $e(q,v)$ is understood as an operator from $K \times V$ into V defined by

$$\text{(1.4)} \qquad (\nabla e(q,v), \nabla \phi) = (q \nabla v, \nabla \phi) - (f, \phi) \quad \forall\, (q,v) \in K \times V, \qquad \phi \in V,$$

where $(\cdot,\cdot)$ denotes the duality pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$, which is the extension of the inner product in $L^2(\Omega)$. It is useful to remark that $e(q,v)$ is convex with respect to each variable.

The problem (1.1)–(1.3) will be solved by the augmented Lagrangian method. Thus we introduce the augmented Lagrangian functional $\mathcal{L}_r : K \times V \times V \to R$ by

$$\text{(1.5)} \qquad \mathcal{L}_r(q,v;\mu) = J(q,v) + (\nabla \mu, \nabla e(q,v)) + \frac{r}{2} \|\nabla e(q,v)\|^2_{L^2(\Omega)},$$

where $r \ge 0$ is some given constant. The first main result of the paper states that the minimization problem (1.1)–(1.3) is equivalent to the saddle-point problem associated with the Lagrangian functional $\mathcal{L}_r$ in (1.5). To solve the saddle-point problem, we propose a finite element discretization of the problem and show that the saddle-points of the discrete problem converge to those of the continuous problem. Finally, we propose an Uzawa algorithm to solve the discrete saddle-point problem and prove the global convergence of the algorithm. We note that recently Chan and Tai have performed many numerical experiments on a local convergent Uzawa algorithm and its combination with domain decomposition and multigrid methods [4, 16].

Throughout the paper, the constant C is a generic constant that might be different at each occurrence but is independent of the mesh parameter h and of the various functions involved.

2. The continuous saddle-point problem. We start this section with the existence of the solutions to the minimization problem (1.1)–(1.3) and then prove that the minimization problem is equivalent to the saddle-point problem of the augmented Lagrangian $\mathcal{L}_r$ defined in (1.5).

LEMMA 2.1. *There exists at least one solution to the minimization problem* (1.1)–(1.3).

Proof. Let

$$A = \Big\{(q,v) \in K \times V; \quad e(q,v) = 0\Big\}$$

be the admissible set of the minimization problem (1.1)–(1.3). It is clear that $A \neq \emptyset$ and $J(q,v) \geq 0$ on A. Thus there exists a minimizing sequence $(q_n, v_n) \in A$ such that

$$\lim_{n\to\infty} J(q_n, v_n) = \inf_{(q,v)\in A} J(q,v). \tag{2.1}$$

Hence $J(q_n, v_n) \leq C$ for each $n > 0$, which implies by definition of J and K that

$$\|v_n\|_{H^1(\Omega)} + \|q_n\|_{BV(\Omega)} \leq C.$$

Therefore, by possibly extracting a subsequence, there exists a pair $(q^*, v^*) \in BV(\Omega) \times V$ satisfying

$$v_n \rightharpoonup v^* \quad \text{in} \quad H_0^1(\Omega), \qquad q_n \to q^* \quad \text{in} \quad L^1(\Omega). \tag{2.2}$$

Since $q_n \in K$, we also have $q^* \in K$. To show that $e(q^*, v^*) = 0$, we first note that $e(q_n, v_n) = 0$ as $(q_n, v_n) \in A$; therefore,

$$(q_n \nabla v_n, \nabla \phi) = (f, \phi) \quad \forall\, \phi \in V. \tag{2.3}$$

However,

$$\begin{aligned}
&\Big|(q_n \nabla v_n, \nabla\phi) - (q^* \nabla v^*, \nabla\phi)\Big| \\
&\leq \Big|((q_n - q^*)\nabla v_n, \nabla\phi)\Big| + \Big|(q^*\nabla(v_n - v^*), \nabla\phi)\Big| \\
&\leq \Big\{\int_\Omega |q_n - q^*|\,|\nabla\phi|^2 dx\Big\}^{1/2} \Big\{\int_\Omega |q_n - q^*|\,|\nabla v_n|^2 dx\Big\}^{1/2} + \Big|(q^*\nabla(v_n - v^*), \nabla\phi)\Big| \\
&\leq C\Big\{\int_\Omega |q_n - q^*|\,|\nabla\phi|^2 dx\Big\}^{1/2} + \Big|(q^*\nabla(v_n - v^*), \nabla\phi)\Big|,
\end{aligned} \tag{2.4}$$

where we have used the fact that $\alpha_1 \leq q_n, q \leq \alpha_2$ and $\|v_n\|_{H^1(\Omega)} \leq C$. Now letting $n \to \infty$ in (2.4), we obtain

$$(q_n \nabla v_n, \nabla\phi) \to (q^* \nabla v^*, \nabla\phi) \quad \forall\, \phi \in V$$

by means of the Lebesgue dominant convergence theorem and the weak convergence in (2.2). Thus we see that $e(q^*, v^*) = 0$ by (2.3) and the definition of $e(\cdot, \cdot)$. Now using (2.2), we have (cf. [9])

$$\int_\Omega |Dq^*| \le \liminf_{n\to\infty} \int_\Omega |Dq_n|.$$

On the other hand, by $e(q_n, v_n) = 0$, we have

$$(q_n \nabla(v_n - z),\ \nabla\phi) = (f, \phi) - (q_n \nabla z,\ \nabla\phi) \quad \forall\, \phi \in V.$$

Taking $\phi = v_n - z$ gives

$$\int_\Omega q_n |\nabla(v_n - z)|^2 dx = (f, v_n - z) - (q_n \nabla z,\ \nabla(v_n - z)).$$

Similarly, using $e(q^*, v^*) = 0$, we get

$$\int_\Omega q^* |\nabla(v^* - z)|^2 dx = (f, v^* - z) - (q^* \nabla z,\ \nabla(v^* - z)).$$

Then using the last two relations, (2.2), and the Lebesgue dominant convergence theorem, we can immediately derive

$$\int_\Omega q^* |\nabla(v^* - z)|^2 dx = \lim_{n\to\infty} \int_\Omega q_n |\nabla(v_n - z)|^2 dx, \tag{2.5}$$

which with (2.1) yields

$$\begin{aligned} J(q^*, v^*) &\le \liminf_{n\to\infty}\ \frac{1}{2}\int_\Omega q_n |\nabla v_n - \nabla z|^2 dx + \liminf_{n\to\infty} \int_\Omega |Dq_n| \\ &\le \liminf_{n\to\infty} J(q_n, v_n) = \inf_{(q,v)\in A} J(q, v). \end{aligned}$$

This completes the proof of Lemma 2.1 as $(q^*, v^*) \in A$. □

The following theorem is the main result of this section.

THEOREM 2.2. *$(q^*, v^*) \in K \times V$ is a solution of the minimization problem (1.1)–(1.3) if and only if there exists a $\lambda^* \in V$ such that $(q^*, v^*, \lambda^*) \in K \times V \times V$ is a saddle-point of the augmented Lagrangian $\mathcal{L}_r : K \times V \times V \to R$, namely,*

$$\mathcal{L}_r(q^*, v^*; \mu) \le \mathcal{L}_r(q^*, v^*; \lambda^*) \le \mathcal{L}_r(q, v; \lambda^*) \quad \forall\, (q, v, \mu) \in K \times V \times V. \tag{2.6}$$

The key step in proving Theorem 2.2 is an appropriate application of the Hahn–Banach convex set separating theorem. To do so, we introduce two subsets in $R \times V$:

$$S = \Big\{ (J(q, v) - J(q^*, v^*) + s,\ e(q, v)) \in R \times V; \quad (q, v) \in K \times V,\ s \ge 0 \Big\}, \tag{2.7}$$

$$T = \Big\{ (-t, 0) \in R \times V; \quad t > 0 \Big\}, \tag{2.8}$$

where $(q^*, v^*) \in K \times V$ is some minimal point of the problem (1.1)–(1.3). The following three lemmas provide the properties of two subsets required by the Hahn–Banach theorem.

LEMMA 2.3. *S and T are two convex subsets in $R \times V$.*

Proof. It is obvious that T is a convex subset in $R \times V$. To see that S is also a convex subset, we let

$$P_i = (J(q_i, v_i) - J(q^*, v^*) + s_i,\ e(q_i, v_i)), \qquad i = 1, 2,$$

be two points in S, where $(q_i, v_i) \in K \times V$ and $s_i \geq 0$. We let $0 < \alpha < 1$, and we have to show that

$$P_\alpha = \alpha P_1 + (1-\alpha)P_2 \equiv (p_\alpha, w_\alpha)$$

with

$$\begin{aligned} p_\alpha &= \alpha J(q_1, v_1) + (1-\alpha)J(q_2, v_2) - J(q^*, v^*) + \alpha s_1 + (1-\alpha)s_2, \\ w_\alpha &= \alpha e(q_1, v_1) + (1-\alpha)e(q_2, v_2) \end{aligned}$$

is also a point in S. Let us now define $q_\alpha \in K$ as

$$q_\alpha = \alpha q_1 + (1-\alpha)q_2$$

and $v_\alpha \in V$ as the solution of the variational problem

$$(q_\alpha \nabla v_\alpha, \nabla \phi) = (\alpha q_1 \nabla v_1 + (1-\alpha)q_2 \nabla v_2,\ \nabla \phi) \quad \forall\, \phi \in V. \tag{2.9}$$

Clearly, $(q_\alpha, v_\alpha) \in K \times V$ is well defined. By (2.9) and the definition of $e(\cdot, \cdot)$, we have

$$\begin{aligned} (\nabla e(q_\alpha, v_\alpha),\ \nabla\phi) &= (q_\alpha \nabla v_\alpha, \nabla\phi) - (f, \phi) \\ &= (\alpha q_1 \nabla v_1 + (1-\alpha)q_2\nabla v_2,\ \nabla\phi) - (f,\ \phi) \\ &= \alpha\{(q_1\nabla v_1,\ \nabla\phi) - (f,\ \phi)\} + (1-\alpha)\{(q_2\nabla v_2,\ \nabla\phi) - (f,\ \phi)\} \\ &= (\alpha\nabla e(q_1, v_1) + (1-\alpha)\nabla e(q_2, v_2),\ \nabla\phi) \quad \forall\, \phi \in V, \end{aligned}$$

which implies that

$$e(q_\alpha, v_\alpha) = \alpha e(q_1, v_1) + (1-\alpha)e(q_2, v_2). \tag{2.10}$$

On the other hand, by the convexity of the BV-seminorm we have

$$\int_\Omega |Dq_\alpha| \leq \alpha \int_\Omega |Dq_1| + (1-\alpha)\int_\Omega |Dq_2|, \tag{2.11}$$

and we know from (2.9) that

$$(q_\alpha \nabla(v_\alpha - z),\ \nabla z) = (\alpha q_1 \nabla(v_1 - z) + (1-\alpha)q_2\nabla(v_2 - z),\ \nabla\phi).$$

Then letting $\phi = v_\alpha - z$ and using Schwarz's inequality give

$$\begin{aligned} &\int_\Omega q_\alpha |\nabla(v_\alpha - z)|^2 dx \\ &\leq \int_\Omega q_\alpha^{-1} |\alpha q_1 \nabla(v_1 - z) + (1-\alpha)q_2\nabla(v_2 - z)|^2 dx \\ &\leq \int_\Omega q_\alpha \left| \frac{\alpha q_1}{q_\alpha}\nabla(v_1 - z) + \frac{(1-\alpha)q_2}{q_\alpha}\nabla(v_2 - z)\right|^2 dx \\ &\leq \int_\Omega q_\alpha \left\{ \frac{\alpha q_1}{q_\alpha}|\nabla(v_1 - z)|^2 + \frac{(1-\alpha)q_2}{q_\alpha}|\nabla(v_2 - z)|^2 \right\} dx \\ &= \alpha\int_\Omega q_1|\nabla(v_1 - z)|^2 dx + (1-\alpha)\int_\Omega q_2|\nabla(q_2 - z)|^2 dx, \end{aligned}$$

where we have used the fact that $(\alpha q_1 + (1-\alpha)q_2)/q_\alpha = 1$ and the convexity of the function $|\cdot|^2$. Now combining this bound with (2.11) we obtain

$$J(q_\alpha, v_\alpha) \le \alpha J(q_1, v_1) + (1-\alpha) J(q_2, v_2), \tag{2.12}$$

and so (2.10) and (2.12) imply that

$$P_\alpha = (J(q_\alpha, v_\alpha) - J(q^*, v^*) + s_\alpha,\ e(q_\alpha, v_\alpha)\ ) \in S$$

since $(q_\alpha, v_\alpha) \in K \times V$ and

$$s_\alpha = \alpha s_1 + (1-\alpha)s_2 + \alpha J(q_1, v_1) + (1-\alpha)J(q_2, v_2) - J(q_\alpha, v_\alpha) \ge 0.$$

This completes the proof of Lemma 2.3. □

LEMMA 2.4. *We have $S \cap T = \emptyset$.*

Proof. Assume that $(a, w) \in S \cap T$; then there exists $(q, v) \in K \times V$ and $s \ge 0$ such that

$$a = J(q, v) - J(q^*, v^*) + s, \qquad w = e(q, v).$$

But $(a, w) \in T$ implies that $a < 0$ and $w = e(q, v) = 0$. Thus

$$J(q, v) + s < J(q^*, v^*),$$

which contradicts the assumption that (q^*, v^*) is a minimal point of the problem (1.1)–(1.3). □

LEMMA 2.5. *The subset S has at least one interior point.*

Proof. It is easy to see that for any $s_0 > 0$, $(s_0, 0) = (J(q^*, v^*) - J(q^*, v^*) + s_0,\ e(q^*, v^*))$ is a point in S. We will show that $(s_0, 0) \in R \times V$ is also an interior point of S. For any $\varepsilon \in (0, 1)$, let (s, w) belong to the ε-neighborhood of $(s_0, 0)$ in $R \times V$, that is,

$$|s - s_0| + \|w\|_{H^1(\Omega)} \le \varepsilon. \tag{2.13}$$

Let $q = q^*$ and $v \in V$ be the solution to the equation

$$(q\nabla v,\ \nabla\phi) = (f, \phi) + (\nabla w,\ \nabla\phi) \quad \forall\, \phi \in V. \tag{2.14}$$

Then we have $e(q, v) = w$. Let

$$\begin{aligned} s' &= s + J(q^*, v^*) - J(q, v) \\ &= s + \frac{1}{2}\int_\Omega q^*|\nabla(v^* - z)|^2 dx - \frac{1}{2}\int_\Omega q^*|\nabla(v - z)|^2 dx \\ &= s - \frac{1}{2}\int_\Omega q^*\nabla(v - v^*)\cdot\nabla(v + v^* - 2z)dx. \end{aligned} \tag{2.15}$$

From (2.14) and $e(q^*, v^*) = 0$, we derive that $\|\nabla v^*\|_{L^2(\Omega)} \le \|f\|_{H^{-1}(\Omega)}/\alpha_1$ and

$$(q^*\nabla(v - v^*),\ \nabla\phi) = (\nabla w,\ \nabla\phi) \quad \forall\, \phi \in V,$$

which yields $\|\nabla(v - v^*)\|_{L^2(\Omega)} \le \varepsilon/\alpha_1$ by (2.13). Also, (2.14) implies that $\|\nabla v\|_{L^2(\Omega)} \le (\|f\|_{H^{-1}(\Omega)} + \varepsilon)/\alpha_1$, thus we deduce from (2.15) that

$$\begin{aligned} s' &\ge s_0 - \varepsilon - \frac{1}{2}\alpha_2\|\nabla(v - v^*)\|_{L^2(\Omega)}\,\|\nabla(v + v^* - 2z)\|_{L^2(\Omega)} \\ &\ge s_0 - \varepsilon - \frac{\alpha_2}{2\alpha_1^2}\varepsilon\ \{\varepsilon + 2\|f\|_{H^{-1}(\Omega)} + 2\alpha_1\|\nabla z\|_{L^2(\Omega)}\}. \end{aligned}$$

Now if ε is sufficiently small, then $s' \geq 0$. Therefore

$$(s, w) = (J(q, v) - J(q^*, v^*) + s',\ e(q, v)\ ) \in K \times V$$

for any (s, w) in the ε-neighborhood of $(s_0, 0)$. This completes the proof. □

Now we are ready to prove Theorem 2.2.

Proof of Theorem 2.2. First, assume that $(q^*, v^*, \lambda^*) \in K \times V \times V$ is a saddle-point of $\mathcal{L}_r$. Then the first inequality in (2.6) immediately gives $e(q^*, v^*) = 0$, and the fact that (q^*, v^*) is a minimal point of the problem (1.1)–(1.3) follows readily from the second inequality in (2.6).

Next we prove the remaining part of the theorem. Let (q^*, v^*) be a minimal point of the problem (1.1)–(1.3), so we have

$$J(q^*, v^*) \leq J(q, v) \quad \forall\, (q, v) \in K \times V \quad \text{satisfying} \quad e(q, v) = 0. \tag{2.16}$$

By Lemmas 2.3–2.5, we can apply the Hahn–Banach theorem (see, e.g., [3, 4, 5, 6]) to separate the two convex subsets S and T defined in (2.7) and (2.8). Thus there exists a pair $(\alpha_0, \lambda_0) \in R \times V$, but $(\alpha_0, \lambda_0) \neq (0, 0) \in R \times V$ such that

$$\alpha_0(J(q, v) - J(q^*, v^*) + s) + (\nabla\lambda_0,\ \nabla e(q, v)) \geq \alpha_0(-t)$$

for any $(q, v) \in K \times V$, $s \geq 0$, and $t > 0$. Taking $(q, v) = (q^*, v^*)$, $s = t = 1$, we get $\alpha_0 \geq 0$, while taking $s = 0$ and letting $t \to 0^+$, we obtain

$$\alpha_0(J(q, v) - J(q^*, v^*)) + (\nabla\lambda_0,\ \nabla e(q, v)) \geq 0 \quad \forall\, (q, v) \in K \times V. \tag{2.17}$$

We now claim that $\alpha_0 > 0$. Otherwise, if $\alpha_0 = 0$ we have from (2.17) that

$$(\nabla\lambda_0,\ \nabla e(q, v)) = (q\nabla v,\ \nabla\lambda_0) - (f, \lambda_0) \geq 0 \quad \forall\, (q, v) \in K \times V, \tag{2.18}$$

which implies that $\lambda_0 = 0$. In fact, taking $q = q^* \in K$ and $v \in V$ to be the solution of the equation

$$(q^*\nabla v,\ \nabla\phi) = (f - \lambda_0,\ \phi) \quad \forall\, \phi \in V, \tag{2.19}$$

we know from (2.18), (2.19) that $-\|\lambda_0\|^2_{L^2(\Omega)} \geq 0$. Thus we have $(\alpha_0, \lambda_0) = (0, 0)$, which is a contradiction. Therefore $\alpha_0 > 0$. Then taking $\lambda^* = \lambda_0/\alpha_0$ and dividing both sides of (2.17) by α_0, we get

$$J(q^*, v^*) \leq J(q, v) + (\nabla\lambda^*,\ \nabla e(q, v)) \quad \forall\, (q, v) \in K \times V,$$

which, combined with (2.16) indicates that $(q^*, v^*, \lambda^*) \in K \times V \times V$ is a saddle-point of the augmented Lagrangian $\mathcal{L}_r$. So we have proved Theorem 2.2. □

3. The discrete saddle-point problem.

Theorem 2.2 tells us that the minimization problem (1.1)–(1.3) is equivalent to finding the saddle-points of the functional $\mathcal{L}_r$ defined in (1.5). In this section, we will consider how to discretize the augmented Lagrangian $\mathcal{L}_r$ and derive a discrete saddle-point problem.

Let Ω be a polyhedral domain in R^d, $d = 1, 2$, or 3, and $\{\mathcal{T}^h\}_{h>0}$ be a family of regular triangulations (cf. Ciarlet [5]) of the domain Ω, with simplicial elements. Denote by V_h the standard piecewise linear finite element space over the triangulation $\mathcal{T}^h$ and

$$\overset{\circ}{V}_h = V_h \cap H^1_0(\Omega), \qquad K_h = K \cap V_h.$$

We now introduce a discrete version of the operator $e(q,v) : K \times V \to V$ defined in (1.4): for any $(q_h, v_h) \in K_h \times \overset{\circ}{V}_h$, $e_h(q_h, v_h) \in \overset{\circ}{V}_h$ is the solution of the system

$$(\nabla e_h(q_h, v_h),\ \nabla\phi) = (q_h \nabla v_h,\ \nabla\phi) - (f, \phi) \quad \forall\, \phi \in \overset{\circ}{V}_h\ . \tag{3.1}$$

It is clear that the operator $e_h : K_h \times \overset{\circ}{V}_h \to \overset{\circ}{V}_h$ is well defined. Moreover, the following estimate holds:

$$\|\nabla e_h(q_h, v_h)\|_{L^2(\Omega)} \le \{\alpha_2 \|\nabla v_h\|_{L^2(\Omega)} + C\|f\|_{H^{-1}(\Omega)}\} \quad \forall\, (q_h, v_h) \in K_h \times V_h, \tag{3.2}$$

where the constant C comes from the Poincaré inequality.

Now for any given $r \ge 0$, we define the discrete augmented Lagrangian L_r: $K_h \times \overset{\circ}{V}_h \times \overset{\circ}{V}_h \to R$ as follows:

$$L_r(q_h, v_h; \mu_h) = J_h(q_h, v_h) + (\nabla \mu_h,\ \nabla e_h(q_h, v_h)) + \frac{r}{2}\|\nabla e_h(q_h, v_h)\|^2_{L^2(\Omega)} \tag{3.3}$$

with

$$J_h(q_h, v_h) = \frac{1}{2}\int_\Omega q_h |\nabla(v_h - z)|^2 dx + \beta \int_\Omega \sqrt{|\nabla q_h|^2 + \delta(h)}\ dx,$$

where $\delta(h)$ above is any given positive function satisfying $\lim_{h\to 0} \delta(h) = \delta(0) = 0$.

With the above preparations, we can state the following theorem.

THEOREM 3.1. *For any $r \ge 0$, there exists at least one saddle-point for the discrete augmented Lagrangian $L_r : K_h \times \overset{\circ}{V}_h \times \overset{\circ}{V}_h \to R$. Moreover, each saddle-point $(q_h^*, v_h^*, \lambda_h^*)$ of L_0 is a saddle-point of L_r for any $r > 0$.*

Proof. It is obvious that each saddle-point of L_0 is a saddle-point of L_r for any $r > 0$. Then it suffices to show that $L_0 : K_h \times \overset{\circ}{V}_h \times \overset{\circ}{V}_h \to R$ has a saddle-point, which we can argue in exactly the same way as in the proof for the continuous saddle-point problem of the last section by showing first the existence of the solutions to the discrete minimization problem

$$\min_{(q_h, v_h) \in A_h} J_h(q_h, v_h) \tag{3.4}$$

with

$$A_h = \{(q_h, v_h) \in K_h \times \overset{\circ}{V}_h; \quad e_h(q_h, v_h) = 0\},$$

and then the existence of the Lagrangian multiplier $\lambda_h^* \in \overset{\circ}{V}_h$ satisfying

$$J_h(q_h^*, v_h^*) \le J_h(q_h, v_h) + (\nabla \lambda_h^*,\ \nabla e_h(q_h, v_h)) \quad \forall\, (q_h, v_h) \in K_h \times \overset{\circ}{V}_h$$

for some minimal point (q_h^*, v_h^*) of the problem (3.4). We omit the details. □

The following theorem is the main result of this section.

THEOREM 3.2. *Each subsequence of the saddle-points $\{(q_h^*, v_h^*; \lambda_h^*)\}_{h>0}$ of the discrete augmented Lagrangian $L_r : K_h \times \overset{\circ}{V}_h \times \overset{\circ}{V}_h \to R$ defined in* (3.3) *has a subsequence that converges to some saddle-point $(q^*, v^*; \lambda^*)$ of the augmented Lagrangian $\mathcal{L}_r : K \times V \times V \to R$ defined in* (1.5) *strongly in $L^1(\Omega) \times L^2(\Omega) \times L^2(\Omega)$.*

The proof of Theorem 3.2 depends on the following three lemmas.

LEMMA 3.3. *Let $g \in BV(\Omega)$. Then for any $\varepsilon > 0$, there exists a function $g_\varepsilon \in C^\infty(\bar{\Omega})$ such that*

$$\int_\Omega |g - g_\varepsilon| dx < \varepsilon, \qquad \Big| \int_\Omega |\nabla g_\varepsilon| dx - \int_\Omega |Dg| \Big| < \varepsilon.$$

Proof. By the approximation property of functions with BVs (cf. p. 172 of [8]), there exists $\tilde{g}_\varepsilon \in C^\infty(\Omega) \cap W^{1,1}(\Omega)$ satisfying

$$\int_\Omega |g - \tilde{g}_\varepsilon| dx < \varepsilon/2, \qquad \Big| \int_\Omega |\nabla \tilde{g}_\varepsilon| dx - \int_\Omega |Dg| \Big| < \varepsilon/2.$$

Then the lemma follows from the density of $C^\infty(\bar{\Omega})$ in $W^{1,1}(\Omega)$ as $\partial\Omega$ is Lipschitz continuous (cf. page 127 of [8]). □

In what follows we will make use of the standard nodal value interpolant $I_h : C(\bar{\Omega}) \to V_h$ and the projection operator $R_h : V \to \overset{\circ}{V}_h$ defined by

$$(\nabla R_h v,\ \nabla \phi) = (\nabla v,\ \nabla \phi) \quad \forall\, v \in V, \qquad \phi \in \overset{\circ}{V}_h\ . \tag{3.5}$$

It is well known (cf. [5]) that for any $p > d = \dim(\Omega)$,

$$\lim_{h \to 0} \|v - I_h v\|_{W^{1,p}(\Omega)} = 0 \quad \forall\, v \in W^{1,p}(\Omega), \tag{3.6}$$

$$\lim_{h \to 0} \|v - R_h v\|_{H_0^1(\Omega)} = 0 \quad \forall\, v \in V. \tag{3.7}$$

LEMMA 3.4. *Assume that $(q, v) \in K \times V$ and $(q_h, v_h) \in K_h \times \overset{\circ}{V}_h$. Then $\lim_{h\to 0} q_h = q$ in $L^1(\Omega)$ and $\lim_{h \to 0} v_h = v$ in $H_0^1(\Omega)$ imply $\lim_{h\to 0} e_h(q_h, v_h) = e(q, v)$ in $H_0^1(\Omega)$.*

Proof. By the definitions of $e(\cdot, \cdot)$ and $e_h(\cdot, \cdot)$ we have

$$(\nabla\{e_h(q_h, v_h) - e(q, v)\},\ \nabla\phi) = ((q_h - q)\nabla v,\ \nabla \phi) + (q_h \nabla(v_h - v), \nabla\phi) \quad \forall\, \phi \in \overset{\circ}{V}_h\ .$$

By taking $\phi = e_h(q_h, v_h) - R_h e(q, v) \in \overset{\circ}{V}_h$ above and using (3.5) we obtain

$$\begin{aligned} \|\nabla\{e_h(q_h, v_h) - R_h e(q, v)\}\|^2_{L^2(\Omega)} &\le 2 \int_\Omega |q_h - q|^2\, |\nabla v|^2 dx + 2 \int_\Omega q_h^2 |\nabla(v_h - v)|^2 dx \\ &\le 2 \int_\Omega |q_h - q|^2\, |\nabla v|^2 dx + 2(\alpha_2)^2 \int_\Omega |\nabla(v_h - v)|^2 dx. \end{aligned}$$

Then the Lebesgue dominant convergence theorem and the fact that $\lim_{h\to 0} v_h = v$ in $H_0^1(\Omega)$ show that

$$\lim_{h\to 0} \|\nabla\{e_h(q_h, v_h) - R_h e(q, v)\}\|_{L^2(\Omega)} = 0.$$

Lemma 3.4 now follows from (3.7). □

LEMMA 3.5. *Assume that $(q, v) \in K \times V$ and $(q_h, v_h) \in K_h \times \overset{\circ}{V}_h$. Then $\lim_{h \to 0} q_h = q$ in $L^1(\Omega)$ and $\lim_{h\to 0} v_h = v$ weakly in $H_0^1(\Omega)$ imply that $\lim_{h\to 0} e_h(q_h, v_h) = e(q, v)$ weakly in $H_0^1(\Omega)$.*

Proof. For any $\phi \in V$, let $\phi_h = R_h\phi$. By the definition of R_h and $e_h(\cdot,\cdot)$ we have

$$\begin{aligned}
&(\nabla e_h(q_h, v_h),\ \nabla\phi)\\
&= (\nabla e_h(q_h, v_h),\ \nabla\phi_h)\\
&= (q\nabla v_h,\ \nabla\phi_h) + ((q_h - q)\nabla v_h,\ \nabla\phi_h) - (f, \phi_h)\\
&= (q\nabla v_h,\ \nabla\phi_h) + ((q_h - q)\nabla v_h,\ \nabla\phi) - (f, \phi_h)\\
&\quad +((q_h - q)\nabla v_h,\ \nabla(\phi_h - \phi)).
\end{aligned} \tag{3.8}$$

Then using the assumed convergence on v_h, we know that $\{\|\nabla v_h\|_{L^2(\Omega)}\}$ is bounded; combining this with the Lebesgue dominant convergence theorem we derive

$$\Big|((q_h - q)\nabla v_h,\ \nabla\phi)\Big| \le \|\nabla v_h\|_{L^2(\Omega)}\Big\{\int_\Omega |q_h - q|^2|\nabla\phi|^2 dx\Big\}^{1/2} \to 0 \quad \text{as} \quad h \to 0.$$

Similarly, we can show that all other terms in (3.8) converge; we then take the limit in (3.8) and use the definition of $e(\cdot,\cdot)$ to yield

$$\lim_{h\to 0}(\nabla e_h(q_h, v_h),\ \nabla\phi) = (q\nabla v,\ \nabla\phi) - (f, \phi) = (\nabla e(q, v),\ \nabla\phi) \quad \forall\,\phi \in V.$$

Thus we have proved Lemma 3.5. □

Now we are ready to prove Theorem 3.2.

Proof of Theorem 3.2. Let $(q_h^*, v_h^*, \lambda_h^*) \in K_h\times \overset{\circ}{V}_h \times \overset{\circ}{V}_h$ be the saddle-point of L_r, that is,

$$L_r(q_h^*, v_h^*; \mu_h) \le L_r(q_h^*, v_h^*; \lambda_h^*) \le L_r(q_h, v_h; \lambda_h^*) \quad \forall\,(q_h, v_h, \mu_h) \in K_h\times \overset{\circ}{V}_h \times \overset{\circ}{V}_h\ .$$

The first inequality implies immediately that $e_h(q_h^*, v_h^*) = 0$, and the second inequality gives us

$$\begin{aligned}
J_h(q_h^*, v_h^*) \le J_h(q_h, v_h) + (\nabla\lambda_h^*,\ \nabla e_h(q_h, v_h)) + \frac{r}{2}\|\nabla e_h(q_h, v_h)\|^2_{L^2(\Omega)}\\
\forall\,(q_h, v_h) \in K_h\times \overset{\circ}{V}_h\ .
\end{aligned} \tag{3.9}$$

By letting $q_h = \alpha_1$, a constant, and $v_h \in\overset{\circ}{V}_h$ be the unique solution of the equation

$$(\nabla v_h,\ \nabla\phi) = \left(\frac{1}{\alpha_1}f,\ \phi\right) \quad \forall\,\phi \in\overset{\circ}{V}_h$$

and hence $e_h(q_h, v_h) = 0$, we deduce from (3.9) that $\|q_h^*\|_{BV(\Omega)} + \|v_h^*\|_{H^1(\Omega)} \le C$. But taking $q_h = q_h^*$ in (3.9) and using (3.2) and the definition of $e_h(\cdot,\cdot)$ we get for any $v_h \in\overset{\circ}{V}_h$ that

$$\begin{aligned}
&\frac{1}{2}\int_\Omega q_h^*|\nabla(v_h^* - z)|^2 dx\\
&\le \frac{1}{2}\int_\Omega q_h^*\,|\nabla(v_h - z)|^2 dx + (q_h^*\nabla v_h, \nabla\lambda_h^*) - (f, \lambda_h^*) + \frac{r}{2}\|\nabla e_h(q_h^*, v_h)\|^2_{L^2(\Omega)}\\
&\le (q_h^*\nabla v_h, \nabla\lambda_h^*) + \eta\|\nabla\lambda_h^*\|^2_{L^2(\Omega)} + \frac{C}{\eta}\|f\|^2_{H^{-1}(\Omega)} + C\{\|\nabla v_h\|^2_{L^2(\Omega)} + \|\nabla z\|^2_{L^2(\Omega)}\}
\end{aligned}$$

for any $\eta > 0$. Now we take $v_h = -\varepsilon\lambda_h^*$ for some constant $\varepsilon > 0$ and $\eta = \frac{1}{2}\alpha_1\varepsilon$ and we derive

$$\frac{1}{2}\alpha_1\varepsilon\|\nabla\lambda_h^*\|^2_{L^2(\Omega)} \le C\Big\{\varepsilon^2\|\nabla\lambda_h^*\|^2_{L^2(\Omega)} + \frac{1}{\varepsilon}\|f\|^2_{H^{-1}(\Omega)} + \|\nabla z\|^2_{L^2(\Omega)}\Big\}.$$

Then choosing $\varepsilon = \alpha_1/(4C)$ above gives $\|\nabla\lambda_h^*\|_{L^2(\Omega)} \le C$. Hence each subsequence of $\{(q_h^*, v_h^*, \lambda_h^*)\}_{h>0}$ has a subsequence, still denoted by $\{(q_h^*, v_h^*, \lambda_h^*)\}$, satisfying

$$q_h^* \to q^* \text{ in } L^1(\Omega), \qquad v_h^* \to v^* \text{ weakly in } H_0^1(\Omega), \qquad \lambda_h^* \to \lambda^* \text{ weakly in } H_0^1(\Omega) \tag{3.10}$$

or

$$q_h^* \to q^* \text{ in } L^1(\Omega), \qquad v_h^* \to v^* \text{ in } L^2(\Omega), \qquad \lambda_h^* \to \lambda^* \text{ in } L^2(\Omega)$$

for some $(q^*, v^*, \lambda^*) \in K \times V \times V$. By Lemma 3.5 we have $e_h(q_h^*, v_h^*) \to e(q^*, v^*)$ weakly in $H_0^1(\Omega)$. Thus $e_h(q_h^*, v_h^*) = 0$ also implies that $e(q^*, v^*) = 0$, and the following holds:

$$\mathcal{L}_r(q^*, v^*; \mu) \le \mathcal{L}_r(q^*, v^*; \lambda^*) \quad \forall\, \mu \in V. \tag{3.11}$$

On the other hand, for any $(q, v) \in K \times V$ and any $\varepsilon > 0$, by Lemma 3.3 we can find a function $q_\varepsilon \in C^\infty(\bar{\Omega})$ satisfying

$$\int_\Omega |q_\varepsilon - q| dx < \varepsilon, \qquad \Big| \int_\Omega |\nabla q_\varepsilon| dx - \int_\Omega |Dq| \Big| < \varepsilon. \tag{3.12}$$

Now we define

$$\tilde{q}_\varepsilon = \begin{cases} \alpha_1 & \text{if } \; q_\varepsilon < \alpha_1, \\ q_\varepsilon & \text{if } \; \alpha_1 \le q_\varepsilon \le \alpha_2, \\ \alpha_2 & \text{if } \; q_\varepsilon > \alpha_2. \end{cases} \tag{3.13}$$

Then $\tilde{q}_\varepsilon \in K \cap W^{1,\infty}(\Omega)$ since

$$\nabla\tilde{q}_\varepsilon = \begin{cases} \nabla q_\varepsilon & \text{on } \; A_\varepsilon = \{x \in \Omega : \alpha_1 \le q_\varepsilon \le \alpha_2\}, \\ 0 & \text{on } \; \Omega \setminus A_\varepsilon. \end{cases} \tag{3.14}$$

Now we take $(q_h, v_h) = (I_h\tilde{q}_\varepsilon, R_h v) \in K_h \times \overset{\circ}{V}_h$ in (3.9) and get

$$J_h(q_h^*, v_h^*) \le J_h(I_h\tilde{q}_\varepsilon, R_h v) + (\nabla\lambda_h^*, \nabla e_h(I_h\tilde{q}_\varepsilon, R_h v)) + \frac{r}{2}\|\nabla e_h(I_h\tilde{q}_\varepsilon, R_h v)\|_{L^2(\Omega)}^2. \tag{3.15}$$

Then by the lower semicontinuity of the BV-norm (cf. [9]) we derive

$$\begin{aligned} &\liminf_{h\to 0} J_h(q_h^*, v_h^*) \\ &\ge \liminf_{h\to 0} \Big\{ \frac{1}{2}\int_\Omega q_h^* |\nabla(v_h^* - z)|^2 dx + \beta \int_\Omega |Dq_h^*| \Big\} \\ &\ge \liminf_{h\to 0} \frac{1}{2}\int_\Omega q_h^* |\nabla(v_h^* - z)|^2 dx + \liminf_{h\to 0} \beta \int_\Omega |Dq_h^*| \\ &\ge \frac{1}{2}\int_\Omega q^* \, |\nabla(v^* - z)|^2 dx + \beta \int_\Omega |Dq^*| = J(q^*, v^*), \end{aligned} \tag{3.16}$$

where we have used the following result:

$$\lim_{h\to 0} \frac{1}{2}\int_\Omega q_h^* |\nabla(v_h^* - z)|^2 dx = \frac{1}{2}\int_\Omega q^* \, |\nabla(v^* - z)|^2 dx,$$

which can be proved in exactly the same way as for (2.5).

Now by (3.6) and (3.7) we know that

$$\lim_{h\to 0} I_h\tilde{q}_\varepsilon = \tilde{q}_\varepsilon \text{ in } W^{1,1}(\Omega), \qquad \lim_{h\to 0} R_h v = v \text{ in } H_0^1(\Omega);$$

combining this with Lemma 3.4 gives

$$\lim_{h\to 0} e_h(I_h\tilde{q}_\varepsilon, R_h v) = e(\tilde{q}_\varepsilon, v) \text{ in } H_0^1\Omega).$$

Then letting $h \to 0$ in (3.15) and using (3.16) we obtain

$$J(q^*, v^*) \le J(\tilde{q}_\varepsilon, v) + (\nabla\lambda^*, \nabla e(\tilde{q}_\varepsilon, v)) + \frac{r}{2}\|\nabla e(\tilde{q}_\varepsilon, v)\|^2_{L^2(\Omega)}. \tag{3.17}$$

Since $q \in K$, we have from (3.12) and (3.13) that

$$\|\tilde{q}_\varepsilon - q\|_{L^1(\Omega)} \le \|q_\varepsilon - q\|_{L^1(\Omega)} < \varepsilon.$$

Thus $\lim_{\varepsilon\to 0}\tilde{q}_\varepsilon = q$ in $L^1(\Omega)$, which implies that $\lim_{\varepsilon\to 0} e(\tilde{q}_\varepsilon, v) = e(q, v)$ in $H_0^1(\Omega)$. Hence as $\varepsilon \to 0$, we derive

(3.18)
$$(\nabla\lambda^*, \nabla e(\tilde{q}_\varepsilon, v)) + \frac{r}{2}\|\nabla e(\tilde{q}_\varepsilon, v)\|^2_{L^2(\Omega)} \to (\nabla\lambda^*, \nabla e(q, v)) + \frac{r}{2}\|\nabla e(q, v)\|^2_{L^2(\Omega)}.$$

But by (3.14) and (3.12) we obtain

$$\int_\Omega |\nabla\tilde{q}_\varepsilon| dx = \int_{A_\varepsilon} |\nabla q_\varepsilon| dx \le \int_\Omega |\nabla q_\varepsilon| dx \le \int_\Omega |Dq| + \varepsilon;$$

therefore

$$\liminf_{\varepsilon\to 0} J(\tilde{q}_\varepsilon, v) \le \frac{1}{2}\int_\Omega q|\nabla(v - z)|^2 dx + \beta\int_\Omega |Dq| = J(q, v). \tag{3.19}$$

By substituting (3.18), (3.19) into (3.17) and passing to the limit $\varepsilon \to 0$ we finally get

$$\mathcal{L}_r(q^*, v^*; \lambda^*) = J(q^*, v^*) \le \mathcal{L}_r(q, v; \lambda^*) \quad \forall\,(q, v) \in K \times V.$$

This, together with (3.11), indicates that $(q^*, v^*; \lambda^*)$ is a saddle-point of $\mathcal{L}_r$. □

4. An Uzawa algorithm.

In this section, we study an algorithm of the Uzawa type to find the saddle-points of the discrete augmented Lagrangian $L_r :\ K_h\times \overset{\circ}{V}_h \times \overset{\circ}{V}_h\to R$ defined in (3.3). We consider the following algorithm.

Uzawa Algorithm 1. We are given $\lambda^0 \in\overset{\circ}{V}_h$. Then for $n \ge 0$, with λ^n known, determine the pair $\{p^n, u^n\} \in K_h\times \overset{\circ}{V}_h$ such that

$$L_r(p^n, u^n; \lambda^n) \le L_r(q, v; \lambda^n) \quad \forall\,(q, v) \in K_h\times \overset{\circ}{V}_h; \tag{4.1}$$

then compute λ^{n+1} by

$$\lambda^{n+1} = \lambda^n + \rho_n e_h(p^n, u^n). \tag{4.2}$$

THEOREM 4.1. *Assume that* $0 < \beta_0 \le \rho_n \le \beta_1 < r$ *for any* $n = 1, 2, \ldots.$ *Then any subsequence of* $\{p^n, u^n; \lambda^n\}$ *computed in the Uzawa algorithm* (4.1), (4.2) *has a subsequence (still denoted by)* $\{p^n, u^n; \lambda^n\}$ *such that*

$$p^n \to p \quad in \quad L^1(\Omega), \qquad u^n \to u \quad in \quad L^2(\Omega), \qquad \lambda^n \to \lambda \quad in \quad L^2(\Omega),$$

and $J_h(p^n, u^n) \to J_h(p, u)$ *as* $n \to \infty$. *Furthermore,* $\{p, u; \lambda\} \in K_h \times \overset{\circ}{V}_h \times \overset{\circ}{V}_h$ *is a saddle-point of* $L_r : K_h \times \overset{\circ}{V}_h \times \overset{\circ}{V}_h \to R$.

Proof. First, by Theorem 3.1 there exists a saddle-point $(q_h^*, v_h^*; \lambda_h^*)$ of $L_0 : K_h \times \overset{\circ}{V}_h \times \overset{\circ}{V}_h$, namely,

$$L_0(q_h^*, v_h^*; \mu_h) \le L_0(q_h^*, v_h^*; \lambda^*) \le L_0(q_h, v_h; \lambda^*) \quad \forall\, (q_h, v_h, \mu_h) \in K_h \times \overset{\circ}{V}_h \times \overset{\circ}{V}_h \,.$$

The first inequality immediately gives $e_h(q_h^*, v_h^*) = 0$, and the second implies that

$$(4.3) \quad J_h(q_h^*, v_h^*) \le J_h(q_h, v_h) + (\nabla \lambda_h^*,\ \nabla e_h(q_h, v_h)), \qquad \forall\, (q_h, v_h) \in K_h \times \overset{\circ}{V}_h \,.$$

Then taking $(q, v) = (q_h^*, v_h^*)$ in (4.1) and using (4.3) we obtain

$$\begin{aligned} &J_h(p^n, u^n) + (\nabla \lambda^n,\ \nabla e_h(p^n, u^n)) + \frac{r}{2}\|\nabla e_h(p^n, u^n)\|^2_{L^2(\Omega)} \\ \le\ &L_r(q_h^*, v_h^*; \lambda^n) = J_h(q_h^*, v_h^*) \\ \le\ &J_h(p^n, u^n) + (\nabla \lambda_h^*,\ \nabla e_h(p^n, u^n)). \end{aligned}$$

Hence

$$(4.4) \qquad (\nabla(\lambda^n - \lambda_h^*),\ \nabla e_h(p^n, u^n)) + \frac{r}{2}\|\nabla e_h(p^n, u^n)\|^2_{L^2(\Omega)} \le 0.$$

Now let $\bar\lambda^n = \lambda^n - \lambda_h^*$; then we have

$$\bar\lambda^{n+1} = \bar\lambda^n + \rho_n e_h(p^n, u^n)$$

and thus

$$\begin{aligned} (\nabla \bar\lambda^n,\ \nabla e_h(p^n, u^n)) &= \frac{1}{\rho_n}(\nabla \bar\lambda_h,\ \nabla(\bar\lambda^{n+1} - \bar\lambda^n)) \\ &= \frac{1}{2\rho_n}\left\{\|\nabla \bar\lambda^{n+1}\|^2_{L^2(\Omega)} - \|\nabla \bar\lambda^n\|^2_{L^2(\Omega)} - \|\nabla(\bar\lambda^{n+1} - \bar\lambda^n)\|^2_{L^2(\Omega)}\right\} \\ &= \frac{1}{2\rho_n}\left\{\|\nabla \bar\lambda^{n+1}\|^2_{L^2(\Omega)} - \|\nabla \bar\lambda^n\|^2_{L^2(\Omega)} - \rho_n^2\|\nabla e_h(p^n, u^n)\|^2_{L^2(\Omega)}\right\}. \end{aligned}$$

Substituting this into (4.4), we get

$$\frac{1}{2\rho_n}\left\{\|\nabla \bar\lambda^{n+1}\|^2_{L^2(\Omega)} - \|\nabla \bar\lambda^n\|^2_{L^2(\Omega)}\right\} + \frac{1}{2}(r - \rho_n)\|\nabla e_h(p^n, u^n)\|^2_{L^2(\Omega)} \le 0.$$

Thus if $0 < \rho_n < r$, the sequence $\{\|\nabla \bar\lambda^n\|^2_{L^2(\Omega)}\}$ is monotonically decreasing and $\|\nabla e_h(p^n, u^n)\|_{L^2(\Omega)} \to 0$ as $n \to \infty$. Now letting $(q, v) = (q_h^*, v_h^*)$ in (4.1) we derive

$$J_h(p^n, u^n) \le J_h(q_h^*, v_h^*) - (\nabla \lambda^n,\ \nabla e_h(p^n, u^n)) \le C$$

with constant C independent of n. Therefore

$$\|p^n\|_{BV(\Omega)} + \|\nabla u^n\|_{L^2(\Omega)} \le C,$$

which implies that each subsequence of $\{p^n, u^n, \lambda^n\}$ has a subsequence (still denoted by) $\{p^n, u^n, \lambda^n\}$ such that

$$(p^n, u^n, \lambda^n) \to (p, u, \lambda) \text{ in } L^1(\Omega) \times L^2(\Omega) \times L^2(\Omega) \quad \text{as} \quad n \to \infty$$

for some $(p, u, \lambda) \in K_h\times \overset{\circ}{V}_h \times \overset{\circ}{V}_h$. Note that in a finite-dimensional space all the convergences are equivalent. Thus $e_h(p, u) = 0$ by means of

$$\|\nabla e_h(p^n, u^n)\|_{L^2(\Omega)} \to 0 \quad \text{and} \quad e_h(p^n, u^n) \to e_h(p, u) \quad \text{as} \quad n \to \infty.$$

Now letting $n \to \infty$ in (4.1) we easily obtain

$$L_r(p, u; \lambda) \le L_r(q, v; \lambda) \quad \forall\, (q, v) \in K_h\times \overset{\circ}{V}_h\ .$$

Therefore $(p, u, \lambda) \in K_h\times \overset{\circ}{V}_h \times \overset{\circ}{V}_h$ is a saddle-point of L_r. □

Remark. To reduce the size of the minimization problem in (4.1), one may further divide the problem into two minimization subproblems with each seeking only one of the first two variables of the discrete augmented Lagrangian $L_r(\cdot, \cdot; \cdot)$. See Uzawa Algorithm 2 in the next section and [10, 3] for more algorithms of the same kind.

5. Numerical experiments. We now show some numerical experiments on the proposed method for parameter identification. We first describe how to implement the optimization step in (4.1). In order to solve the system (4.1) for the pair $\{p^n, u^n\}$, we use the following alternative iteration.

Uzawa Algorithm 2. We are given $\lambda^0 \in\overset{\circ}{V}_h$ and $q^0 \in K_h$. Set $n = 1$.

1. Set $k = 1$ and $q^{n,0} = q^{n-1}$.

2. Compute $u^{n,k}$ by solving

$$L_r(q^{n,k-1}, u^{n,k}; \lambda^{n-1}) = \min_{v_h \in V_h^0} L_r(q^{n,k-1}, v_h; \lambda^{n-1}), \tag{5.1}$$

and then compute $q^{n,k}$ by solving

$$L_r(q^{n,k}, u^{n,k}; \lambda^{n-1}) = \min_{p_h \in V_h} L_r(p_h, u^{n,k}; \lambda^{n-1}). \tag{5.2}$$

Compute $q^{n,k} = \max\{\alpha_1, \min\{q^{n,k}, \alpha_2\}\}$.
If $\|q^{n,k} - q^{n,k-1}\| \le$ tolerance, set $u^n = u^{n,k}$ and $q^n = q^{n,k}$, GOTO **3**;
Otherwise set $k = k + 1$, GOTO **2.**

3. Compute λ^n by

$$\lambda^n = \lambda^{n-1} + \frac{3}{4}\, r\, e_{_h}(p^n, u^n). \tag{5.3}$$

Set $n = n + 1$, GOTO **1.**

We use the Armijo algorithm (cf. Keung and Zou [15]) to solve problem (5.2). As the problem corresponds to a nonlinear algebraic system of equations, one may also use some other more efficient iterative methods. Problem (5.1), combining with the equation for $e_{_h}(q^{n,k-1}, u^{n,k})$, corresponds to two linear algebraic systems of equations (both are positive definite), which are solved here by the conjugate gradient method.

We apply Uzawa Algorithm 2 to identify the discontinuous coefficients in the following test problem:

$$-\frac{d}{dx}\Big(q(x)\frac{d}{dx}u(x)\Big) = f(x), \qquad x \in (0, 1) \quad \text{with} \quad u(0) = u(1) = 0. \tag{5.4}$$

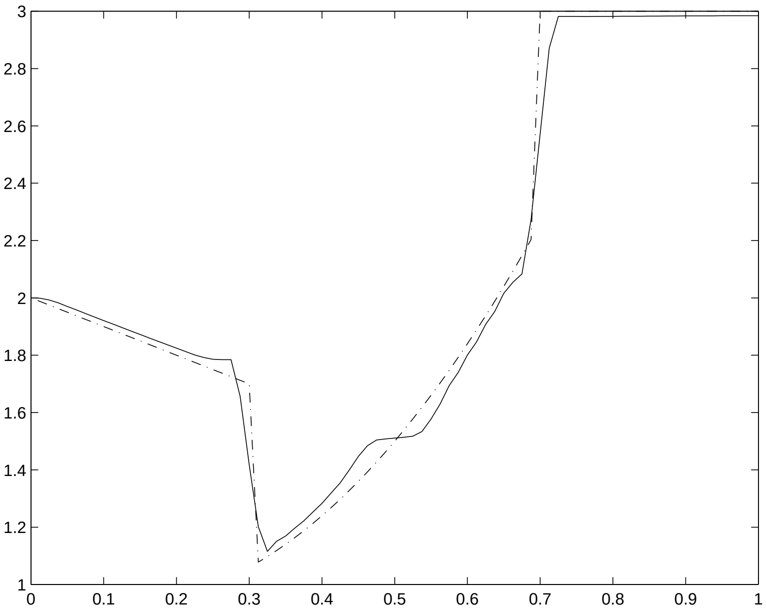


FIG. 5.1. $q_h^0 = 5.0$, $\beta = 10^{-3}$, *error* $= 0.032$, *iter* $= 5$.

Most parameters used in the algorithm are given below each figure. The error shown is the relative L^2-norm error between the exact parameter $q(x)$ to be identified and the computed parameter q_h. The regularization and smoothing parameters β and $\delta(h)$ (see (3.3)) are chosen to be 10^{-3} and 0.01. The augmented Lagrangian coefficient r is taken to be 1, and the finite element mesh size h to be 1/80. The lower and upper bounds α_1 and α_2 in the constrained set K are taken to be 0.5 and 20.0, respectively.

Example 1. We take the following discontinuous coefficient:

$$q(x) = \begin{cases} 2 - x, & x \in [0,\ 0.3], \\ 1 - x + 4x^2, & x \in (0.3,\ 0.7), \\ 3, & x \in [0.7,\ 1], \end{cases}$$

and compare it with the numerically identified solution q_h obtained by using Uzawa Algorithm 2. The exact observation data z is taken as $z(x) = u(q)(x) = \sin(\pi x)$, and the function $f(x)$ is then computed by (5.4) using $u(x)$ and $q(x)$. Figure 5.1 shows the exact solution $q(x)$ (the dotted line) and the numerically identified solution $q_h(x)$ (the solid line). The initial guesses λ^0 and q_h^0 are taken to be the constants 0 and 5.0, respectively. $q_h^0 = 5.0$ is not a good initial guess at all, but the numerical method converges very stably and fast; Figure 5.1 gives the result of the 5th iteration ($n = 5$).

We now add some random noise to the gradient of the true solution u. (Recall that we used the energy-norm in the output least squares formulation. If the L^2-norm is used, one should consider the noised observation data z of the true solution u directly, instead of the gradient.) Namely, we replace the gradient ∇z in the cost functional L_r with the noised data

$$\nabla z^\delta(x) = \nabla z(x) + \delta\, \mathrm{rand}\,(x),$$

where rand (x) is a uniformly distributed random function in $[-1, 1]$ and δ is the noise level parameter. The numerical result of the 5th iteration is shown in Figure 5.2 with

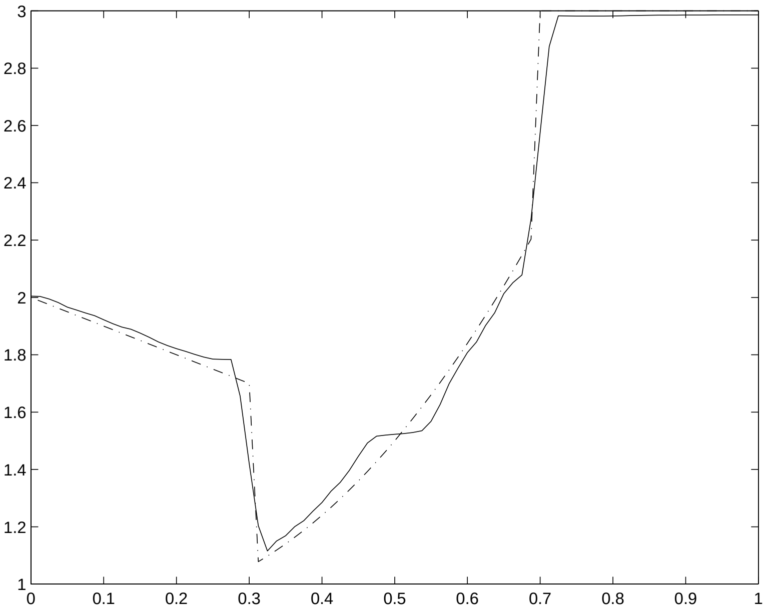


FIG. 5.2. $q_h^0 = 5.0$, $\beta = 10^{-3}$, *noise level* $\delta = 1\%$, *error* $= 0.033$, *iter* $= 5$.

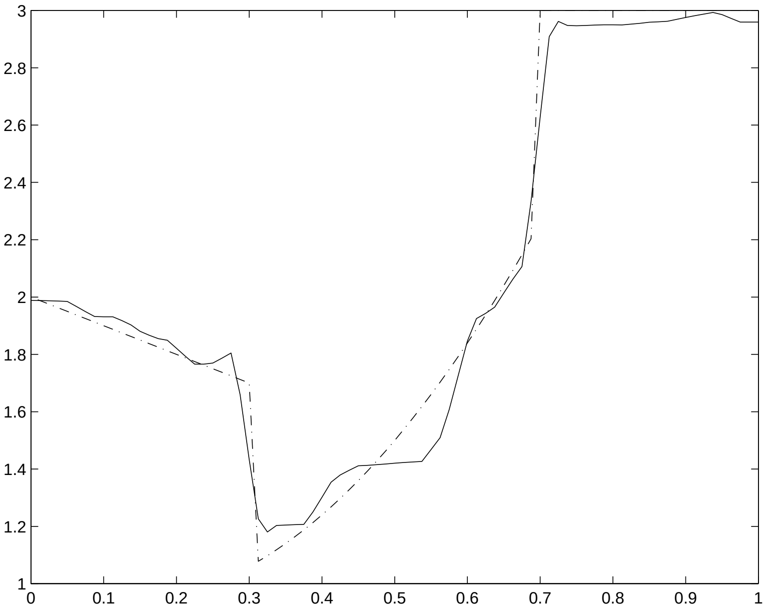


FIG. 5.3. $q_h^0 = 5.0$, $\beta = 10^{-3}$, *noise level* $\delta = 10\%$, *error* $= 0.038$, *iter* $= 5$.

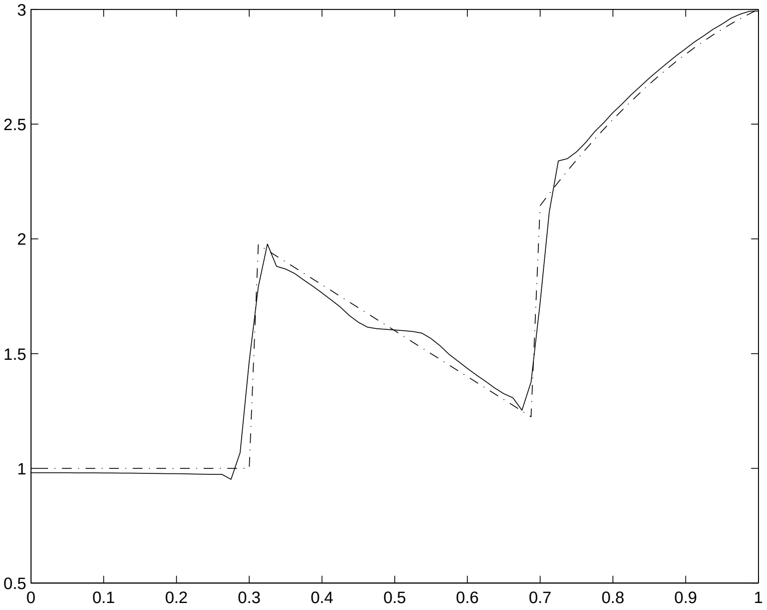


FIG. 5.4. $q_h^0 = 5.0$, $\beta = 10^{-3}$, *error* $= 0.043$, *iter* $= 5$.

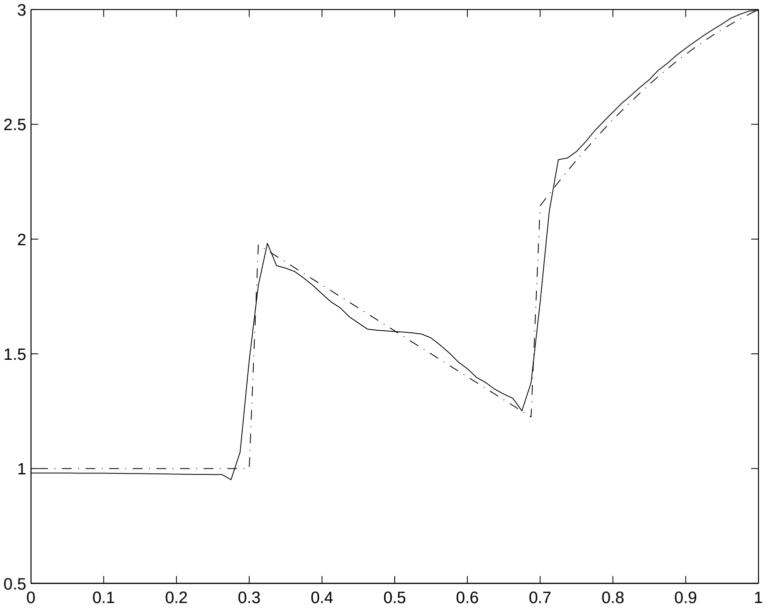


FIG. 5.5. $q_h^0 = 5.0$, $\beta = 10^{-3}$, *noise level* $\delta = 1\%$, *error* $= 0.045$, *iter* $= 5$.

the noise level parameter $\delta = 1\,\%$. We do not see much difference compared with the noise-free case (Figure 5.1). When the noise increases to 10%, the numerical identified solution is still very satisfactory; see Figure 5.3. This indicates that the numerical

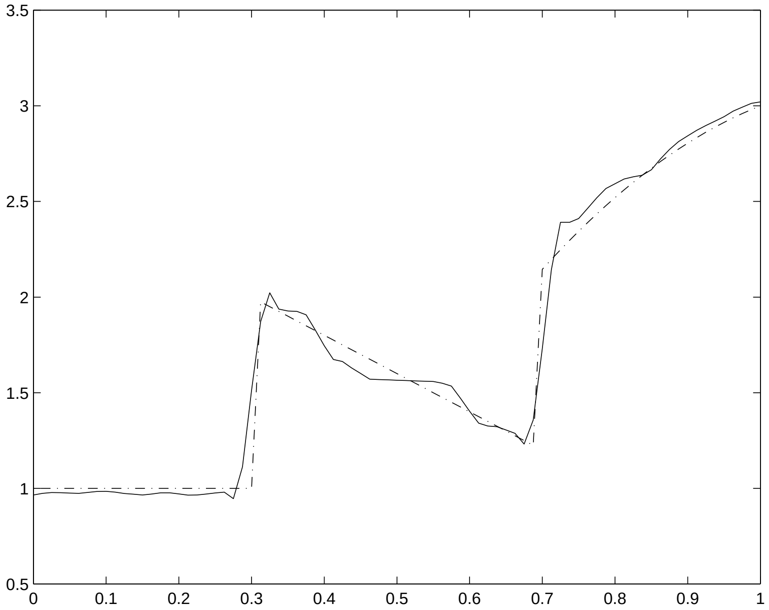


FIG. 5.6. $q_h^0 = 5.0$, $\beta = 10^{-3}$, *noise level* $\delta = 10\%$, *error* $= 0.051$, *iter* $= 5$.

method is not very sensitive to the noise.

Example 2. We take the discontinuous coefficient:

$$q(x) = \begin{cases} 1, & x \in [0,\ 0.3], \\ 2.6 - 2x, & x \in (0.3,\ 0.7), \\ -9\,x^2/2 + 21\,x/2 - 3, & x \in [0.7,\ 1], \end{cases}$$

and compare it again with the numerical solution q_h recovered by Uzawa Algorithm 2. Figure 5.4 shows the exact solution $q(x)$ (the dotted line) and the numerically identified solution $q_h(x)$ (the solid line), where we have taken the initial guesses $\lambda^0 = 0$ and $q_h^0 = 5.0$. We see again that the numerical method converges very stably and fast. Figure 5.4 is the result of the 5th iteration ($n = 5$).

Again, we add some random noise to the gradient of the true solution u; namely, we assume that the available data are the following noised data:

$$\nabla z^\delta(x) = \nabla z(x) + \delta\,\mathrm{rand}\,(x).$$

Figure 5.5 gives the numerical result of the 5th iteration with the noise level parameter $\delta = 1\,\%$. We can see that noise of this level has very little effect on the accuracy and stability of the numerical method. When the noise increases to 10%, the numerical identified solution is still very satisfactory; see Figure 5.6.

Our numerical experiences show that the numerical method proposed in the paper converges very fast (5 iterations for the considered examples) and globally, which is consistent with our theory. In fact one can take much worse initial guesses than the preceding ones ($q_h^0 = 5.0$). More importantly, the method seems to be not very sensitive to the noise.

REFERENCES

[1] H.T. Bank and K. Kunisch, *Estimation Techniques for Distributed Parameter System*, Birkhäuser, Boston, MA, 1989.

[2] D. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[3] J. Cea, *Lectures on Optimization: Theory and Algorithms*, Springer-Verlag, New York, 1978.

[4] T. Chan and X.-C. Tai, *Identification of Discontinuous Coefficients from Elliptic Problems Using Total Variation Regularization*, Technical report CAM 97-35, Dept. of Mathematics, University of California, Los Angeles, CA, 1997.

[5] P. Ciarlet, *Basic error estimates for elliptic problems,* in Handbook of Numerical Analysis, Vol. II, P. G. Ciarlet and J.-L. Lions, eds., North-Holland, Amsterdam, 1991, pp. 17–352.

[6] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[7] H.W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, Dordrecht, 1996.

[8] L. Evans and R. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[9] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser, Boston, MA, 1984.

[10] R. Glowinski and P. Le Tallec, *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*, SIAM, Philadelphia, PA, 1989.

[11] R. Guenther, R. Hudspeth, W. McDougal, and J. Gerlach, *Remarks on parameter identification* I. Numer. Math., 47 (1985), pp. 355–361.

[12] S. Gutman, *Identification of discontinuous parameters in flow equations*, SIAM J. Control Optim., 28 (1990), pp. 1049–1060.

[13] K. Ito and K. Kunisch, *The augmented Lagrangian method for parameter estimation in elliptic systems*, SIAM J. Control Optim., 28 (1990), pp. 113–136.

[14] K. Ito and K. Kunisch, *Augmented Lagrangian-SQP-methods in Hilbert spaces and applications to control in the coefficients problems*, SIAM J. Optim., 6 (1996), pp. 96–125.

[15] Y. Keung and J. Zou, *Numerical identifications of parameters in parabolic systems*, Inverse Problems, 14 (1998), pp. 83–100.

[16] X.-C. Tai, J. Froyen, M. Espedal, and T. Chan, *Overlapping Domain Decomposition and Multigrid Methods for Inverse Problems*, Technical report 113, Dept. of Mathematics, University of Bergen, Bergen, Norway, 1997.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 3, pp. 911–933

PIECEWISE-CONSTANT STABILIZATION*

SERGEY NIKITIN†

Abstract. With the help of topological necessary conditions for continuous stabilization it is shown that, in general, in order to stabilize continuous- and discrete-time systems one has to use time-dependent or discontinuous feedback controls. On the other hand, the criterion of stabilization in the class of piecewise-constant feedbacks is established. In the context of this paper a piecewise-constant feedback is associated with a piecewise-constant function of the form $u = u(x)$, where $x \in \mathrm{R}^n_x$. The piecewise-constant feedback synthesis outlined here has several attractive features. First, it can be effectively applied to design feedback stabilizers subjected to control constraints. Second, the designed feedback laws do not cause sliding mode or chattering behavior in the closed loop system; i.e., on a finite interval of time the control in the closed loop system may have only a finite number of jump discontinuities.

Key words. nonlinear system, degree of functions, feedback stabilization, discrete-time system

AMS subject classifications. 93D15, 93D20, 93C10, 93C55

PII. S0363012997321930

1. Introduction.

Stabilization of dynamical systems is one of the basic problems in systems theory. In [9], [10], and [11] it is shown that many nonlinear systems are not stabilizable by any continuous feedback of the form $u = u(x)$. For the purpose of illustration we give a geometrical interpretation of the results presented in [9], [10], and [11]. In particular, one can see from these geometrical illustrations that on a compact, simply connected manifold a nonlinear system cannot be globally stabilized at any of its equilibria by a continuous feedback of the form $u = u(x)$. We use this fact as a motivation for our work on the main contribution of this paper, the criterion of stabilization by means of piecewise-constant feedbacks that do not cause sliding mode or chattering behavior. In other words, on a finite interval of time the control in the closed loop system may have only a finite number of jump discontinuities.

This paper deals with a dynamical nonlinear system having either the form

$$\dot{x} = f(x, u) \tag{1}$$

or

$$x_{k+1} = f(x_k, u_k), \tag{2}$$

where $x \in \mathrm{R}^n$ (R^n denotes n-dimensional Euclidean space) and $u \in \mathrm{U} \subset \mathrm{R}^m$.

Let pr_x denote the projection of $\mathrm{R}^n_x \times \mathrm{R}^n_u$ onto R^n_x, i.e., $pr_x(x, u) = x$. It will be shown that (1) (or (2)) is stabilizable at an equilibrium

$$(x^*, u^*) \in f^{-1}(0) = \{(x, u) \in \mathrm{R}^n \times \mathrm{U} : \quad f(x, u) = 0\} \text{ for system (1),}$$

or

$$(x^*, u^*) \in (pr_x - f)^{-1}(0) = \{(x, u) \in \mathrm{R}^n \times \mathrm{U} : \quad f(x, u) = x\} \text{ for system (2),}$$

*Received by the editors May 28, 1997; accepted for publication (in revised form) June 25, 1998; published electronically March 17, 1999.

http://www.siam.org/journals/sicon/37-3/32193.html

†Department of Mathematics, Arizona State University, Tempe, AZ 85287-1804 (nikitin@asu.edu).

over a compact set $\mathrm{K} \subset \mathrm{R}^n$ if and only if one can steer the system from any initial point $x \in \mathrm{K}$ into x^* with the help of piecewise-constant control inputs and system (1) (or (2)) satisfies the following condition (of some sort of stability) at $(x^*, u^*) \in f^{-1}(0)$: For any neighborhood W of x^* (open connected set containing x^*) there exists another neighborhood V of x^*, such that one can move the system (1) (or (2)) from any point in V into x^* without leaving the neighborhood W. Although we prove all our results for systems defined on R^n, their reformulation for systems on a smooth manifold is straightforward and seems not to bring anything new except some changes in phraseology.

The use of continuous stationary feedbacks of the form $u = u(x)$ does not allow solution of the problem of stabilization for many nonlinear systems. This fact was understood by many system researchers (see, e.g., [1], [2], [7]). Thus, to stabilize a nonlinear system in many practical and theoretical situations one needs to design time-dependent or discontinuous feedbacks instead of continuous feedbacks of the form $u = u(x)$. The use of time-dependent continuous feedback laws was considered, for instance, in [4], [12], [13], [14] and discontinuous feedback controls have been discussed in, e.g., [9], [10].

The idea of using discontinuous instead of continuous stabilizers is not new and was broadly discussed in the framework of variable structure systems (see, e.g., [5]). The mathematical foundation of variable structure systems was developed in [15].

Necessary conditions of smooth stabilization underline the fact that the use of nonstationary and discontinuous feedback stabilizers is unavoidable in many applications of control theory. The commonly accepted starting point for the discussion on necessary conditions of smooth stabilization is the classical result of [2].

Theorem 1 (see Brockett [2]). *If the system $\dot{x} = f(x, u)$ is continuously stabilizable at (x^*, u^*), then:*

(bi) *all the modes of its linearization $\dot{x} = Ax + Bu$ with positive real parts are controllable;*

(bii) *there exists some neighborhood Q of $x^* \in \mathrm{R}$ such that for each $y \in Q$ one can find a control*

$$u_y(t) : [0, \infty) \to \mathrm{R}^{\mathrm{m}},$$

which steers the system from y at $t = 0$ to x^* at $t = \infty$;

(biii) *the mapping $f(x, u) : \mathrm{R}^{\mathrm{n}}_{\mathrm{x}} \times \mathrm{R}^{\mathrm{m}}_{\mathrm{u}} \to \mathrm{R}^{\mathrm{n}}$ maps every neighborhood of (x^*, u^*) onto a neighborhood of zero.*

Theorem 1 was one of the first necessary conditions for smooth stabilizability of nonlinear systems. It played an outstanding role in the development of nonlinear control theory. For example, Theorem 1 was successfully used to establish that many drift-free nonholonomic nonlinear control systems cannot be continuously stabilized. Nevertheless, necessary condition (biii) suffers deficiency of being generic, i.e., (biii) is almost always satisfied. The result obtained in [3] strengthens (biii), but nevertheless (biii) remains generic and therefore fulfilled for almost every smooth nonlinear system (1).

In [10], [11] the first nongeneric topological necessary conditions have been obtained. Here we give geometrical illustration of these conditions and generalize them to the class of discrete-time systems having form (2). These necessary conditions imply that on a smooth compact manifold neither any continuous nor any discrete-time system is globally stabilized by a continuous feedback of the form $u = u(x)$. We also prove the criterion of stabilization in the class of piecewise-constant feedbacks and discuss synthesis procedures for this class of feedback control laws. The motivation for studying piecewise-constant feedbacks is threefold. First, a piecewise-constant feedback is meant to be implemented on digital processors and does not require quantification of input-output signals. Second, the control theory ideology and, in particular, piecewise-constant feedback stabilization may serve as a framework for software development projects in which a software package is treated as a feedback control. Third, as one can see from the results of this paper, any system that is stabilized by a continuous feedback $u = u(t, x)$ can be stabilized by a piecewise-constant feedback.

We hope that we have convinced our reader that, in some situations where stabilization of a nonlinear system is concerned, a piecewise-constant feedback could be preferable to a continuous one.

The structure of the paper is as follows. The first section is introductory. The second section presents geometrical interpretation of the necessary conditions obtained in [11] and generalizes them to the class of discrete-time systems of the form (2). The third section contains the criterion of stabilization in the class of piecewise-constant feedbacks. The fourth section presents conclusions.

2. Topological necessary conditions of continuous stabilization. This section presents topological necessary conditions for continuous stabilization. It also contains geometrical interpretations of the results obtained in [11], as well as new formulation and proof of these results for discrete-time systems.

2.1. Continuous-time systems. Consider the system

$$\dot{x} = f(x, u),$$

where $x \in \mathbf{R}^n,\ u \in \mathbf{R}^m$. $f(x, u)$ is a complete C^∞ vector field on $\mathbf{R}^{\mathrm{n}}$ for every $u \in \mathbf{R}^{\mathrm{m}}$ fixed.

The set

$$f^{-1}(0) = \{(x, u) \in \mathbf{R}^{\mathrm{n+m}};\ \ f(x, u) = 0\}$$

is called the *equilibrium set* of the control system.

A system is said to be *continuously stabilizable* at $(x^*, u^*) \in f^{-1}(0)$ over a set $K \subset \mathrm{R}^n$ if there exists a continuous function $u = u(x)$, such that $u(x^*) = u^*$ and x^* is an asymptotically stable equilibrium of the closed loop system

$$\dot{x} = f(x, u(x))$$

and the domain of x^*-attraction contains K, i.e.,

$$\lim_{t \to +\infty} e^{tf} x = x^* \quad \forall\, x \in K \subset \mathbf{R}^n,$$

where e^{tf} is the flow generated by the vector field $f(x, u(x))$. We say that x^* is stable if for any neighborhood W of x^* (open connected set containing x^*) there is a neighborhood V of x^*, such that

$$\forall\ \ t \geq 0 \quad e^{tf} V \subset W,$$

where $e^{tf} V = \{e^{tf} x;\ \ x \in V\}$.

A system which is continuously stabilizable at $(x^*, u^*) \in f^{-1}(0)$ over R^n is said to be *continuously stabilizable in the large.*

2.1.1. Basics of function degree. For the sake of completeness and reader convenience, we briefly recall some facts about the degree of continuous functions. Additional details can be found in [6], [8]. We use the following notation:

(i) M is a compact, n-dimensional, oriented C^∞ manifold. Its interior is denoted by $\mathrm{Int}(M)$, its boundary by ∂M.

(ii) $x = (x_1, \ldots, x_n) \in \mathrm{R}^\mathrm{n}$; $|x|^2 = \langle x, x\rangle$, where

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i \quad \forall\, x,\ y \in \mathrm{R}^\mathrm{n}.$$

x is also used for local coordinates on M with a fixed orientation. If local coordinates are given, then $\frac{\partial}{\partial x} f(x)$ is the Jacobian matrix and $\det(\frac{\partial}{\partial x} f(x))$ is the Jacobian determinant of f at x.

(iii) $f \in C^1(M)$ if $f : M \to \mathrm{R}^\mathrm{n}$ and f has continuous first-order partial derivatives in M. $f \in C(M)$ if f is a continuous function from M into R^n.

(iv) Given a real positive number r and $y \in \mathrm{R}^\mathrm{n}, B_r(y)$ is the closed ball center y, radius r:

$$B_r(y) = \{x \in \mathrm{R}^\mathrm{n};\ \mid x - y \mid \leq r\ \}.$$

(v) An immersion $i:\ M \ \to\ \mathrm{R}^\mathrm{n}_\mathrm{x} \times \mathrm{R}^\mathrm{m}_\mathrm{u}$ which maps M homeomorphically into its image $i(M) \subset \mathrm{R}^\mathrm{n}_\mathrm{x} \times \mathrm{R}^\mathrm{m}_\mathrm{u}$ with topology induced by $\mathrm{R}^\mathrm{n}_\mathrm{x} \times \mathrm{R}^\mathrm{m}_\mathrm{u}$ is called a natural embedding. $i_x:\ M\ \to\ \mathrm{R}^\mathrm{n}_\mathrm{x}, i_\mathrm{u}:\ \mathrm{M}\ \to\ \mathrm{R}^\mathrm{m}_\mathrm{u}$ are used for $P_x \circ i$ and $P_u \circ i$, respectively, where $P_x,\ P_u$ are the projections: $P_x(x, u) = x,\ \ P_u(x, u) = u$.

Definition 1. *Suppose $\phi \in C^1(M)$, $p \notin \phi(\partial M)$ and p is not a critical value of ϕ on M, i.e.,*

$$\det\left(\frac{\partial}{\partial x}\phi(x)\right) \neq 0 \quad \forall\, x \in \phi^{-1}(p) = \{x \in M;\quad \phi(x) = p\}.$$

Define the degree of ϕ at p relative to M to be $d(\phi, M, p)$, where

$$d(\phi, M, p) = \sum_{x\in\phi^{-1}(p)} sign\left[\det\left(\frac{\partial}{\partial x}\phi(x)\right)\right].$$

If $\phi \in C(M)$, then the degree of ϕ can be defined as the degree of a sufficiently good C^1 approximation of ϕ (for details, see [6], [8]).

Definition 2. *Suppose that $\phi \in C(M)$ and $p \notin \phi(\partial M)$. Define $d(\phi, M, p)$ to be $d(\psi, M, p)$, where ψ is any function in $C^1(M)$ satisfying*

$$\mid \phi(x) - \psi(x) \mid < \rho(p, \phi(\partial M)) \quad \forall\, x \in M,$$

where $\rho(x, \phi(\partial M)) = \inf_{y\in\phi(\partial M)} \mid x - y \mid$ and p is not a critical value of ψ on M.

Recall that if X and Y are topological spaces, two continuous functions f and g are said to be homotopic ($f \sim g$) if there is a continuous function (homotopy)

$$H:\ \ [0, 1] \times X\ \to\ Y$$

such that

$$H(0, x) = f(x), \quad H(1, x) = g(x) \quad (x \in X).$$

We will need the following properties of degree.

THEOREM 2.

(1) *If $H(t,x) \equiv h_t(x)$ is a homotopy and $p \notin h_t(\partial M)$ for $0 \leq t \leq 1$, then $d(h_t, M, p)$ is independent of $t \in [0,1]$.*

(2) *If a continuous feedback $u = u(x)$ stabilizes the system $\dot{x} = f(x,u)$ at $p \in \mathrm{Int}(B_r(0)) \subset \mathrm{R}^{\mathrm{n}}$, then*

$$d(f(\cdot, u(\cdot)), B_r(0), 0) = (-1)^n.$$

(3) *Suppose $\phi \in C(M)$. If $d(\phi, M, p)$ is defined and nonzero, then there is $q \in M$ such that $\phi(q) = p$.*

(4) *If a continuous feedback $u = u(x)$ stabilizes the system $x_{k+1} = f(x_k, u)$ at $p \in \mathrm{Int}(B_r(0)) \subset \mathrm{R}^{\mathrm{n}}$, then*

$$d(Id_n - f(\cdot, u(\cdot)), B_r(0), 0) = 1,$$

where Id_n is the identity $n \times n$ matrix.

For a proof, see [6], [8].

LEMMA 1. *Let $f : \mathrm{R}^{\mathrm{n}}_{\mathrm{x}} \times \mathrm{R}^{\mathrm{m}}_{\mathrm{u}} \rightarrow \mathrm{R}^{\mathrm{n}}$ be a smooth function, and let ω be a bounded connected component of $f^{-1}(0)$ such that*

$$\mathrm{rank}\left(\frac{\partial}{\partial x} f(x,u), \frac{\partial}{\partial u} f(x,u)\right) = n \quad \forall \ (x,u) \in \omega.$$

If $u = v(x) : \mathrm{R}^n_x \rightarrow \mathrm{R}^m_u$ is a continuous function such that

$$f^{-1}(0) \cap \{(x,u) \in \mathrm{R}^n_x \times \mathrm{R}^m_u \ ; \ u = v(x)\} = \omega \cap \{(x,u) \in \mathrm{R}^n_x \times \mathrm{R}^m_u ; u = v(x)\},$$

then

$$d((f, P_u - v \circ P_x), B_R(0), 0) = 0,$$

where $R > 0$ such that $\omega \subset \mathrm{Int}(B_R(0))$ and P_x, P_u are the projections: $P_x(x,u) = x, P_u(x,u) = u$.

This lemma is proved in [10], [11].

2.1.2. Geometrical illustrations of necessary conditions. We start with the following necessary condition of continuous stabilization.

THEOREM 3. *If $f : \mathrm{R}^{\mathrm{n}}_{\mathrm{x}} \times \mathrm{R}^{\mathrm{m}}_{\mathrm{u}} \rightarrow \mathrm{R}^{\mathrm{n}}$ is a smooth function and $f^{-1}(0)$ is bounded, then the system $\dot{x} = f(x,u)$ is not continuously stabilizable in the large at any point $(x^*, u^*) \in f^{-1}(0)$.*

For a proof of this theorem see [10], [11].

For geometrical illustrations consider a nonlinear single-input system of the form

$$\dot{x} = f(x,u),$$

where $(x,u) \in \mathrm{R}^2_x \times \mathrm{R}_u$.

If $f^{-1}(0)$ is bounded, then Theorem 3 asserts that the system cannot be continuously stabilizable. The reason is that the graph of any continuous feedback $u = u(x)$ stabilizing the system has at least two points of intersection with the equilibrium set $f^{-1}(0)$. Figure 1 depicts a geometrical illustration of this fact.

Example 1. Consider the system

$$\dot{x}_1 = \sin(x_1^2 + x_2^2),$$
$$\dot{x}_2 = u.$$

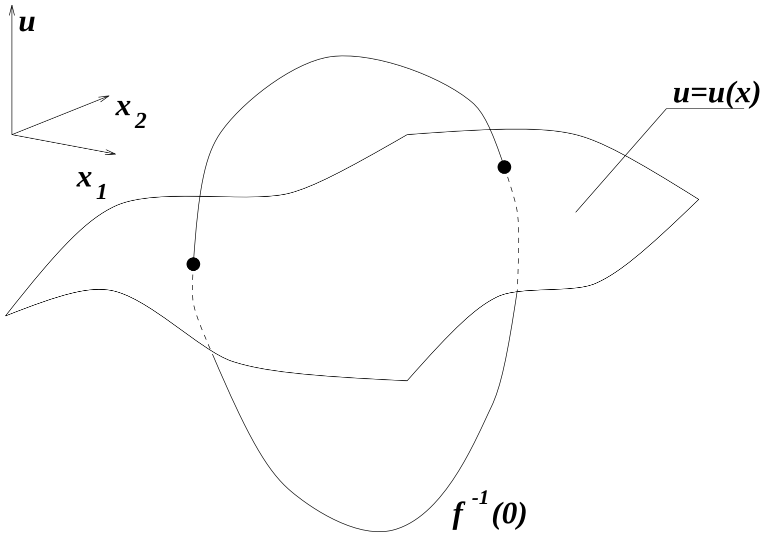


Fig. 1. *If $f^{-1}(0)$ is a closed curve, then the graph of a continuous feedback $u = u(x)$ intersects $f^{-1}(0)$ more than once.*

The equilibrium set for this system is defined by

$$\{(x_1, x_2, u) : \quad u = 0, \;\; x_1^2 + x_2^2 = \pi \cdot n, \quad \text{where} \;\; n = 0, 1, 2, \ldots\}.$$

The system is locally continuously stabilizable at any point of its equilibrium set with $x_2 \neq 0$ (see [10] for further details). But it is not continuously stabilizable at $(x_1^*, x_2^*, 0)$ (with $(x_1^*)^2 + (x_2^*)^2 = \pi \cdot n$) over any compact set containing the entire circle $(x_1^*)^2 + (x_2^*)^2 = \pi \cdot n$. It happens because on a compact set containing the circle the graph of any continuous feedback locally stabilizing the system at (x_1^*, x_2^*) has at least two points of intersection with this circle (Figure 2).

As illustrated by Example 1, system (1) is not continuously stabilizable at any point of a connected bounded component of its equilibrium set. This leads us to the following theorem.

Theorem 4. *Let $f : \mathrm{R}^{\mathrm{n}}_{\mathrm{x}} \times \mathrm{R}^{\mathrm{m}}_{\mathrm{u}} \rightarrow \mathrm{R}^{\mathrm{n}}$ be a smooth function, and let ω be a connected component of $f^{-1}(0)$ such that*

$$\operatorname{rank}\left(\frac{\partial}{\partial x} f(x, u), \frac{\partial}{\partial u} f(x, u)\right) = n \quad \forall \;\; (x, u) \in \omega.$$

If the system $\dot{x} = f(x, u)$ is smoothly stabilizable in the large at a point $(x^*, u^*) \in \omega$, then ω is unbounded.

The proof of this theorem is given in [10], [11].

Theorems 3 and 4 give us the topological necessary conditions that are stable under perturbations which are small in fine Whitney topology. Moreover, these conditions remain valid for a system on a compact smooth simply connected manifold, and since on the compact manifold the equilibrium set $f^{-1}(0)$ is evidently bounded, we conclude that the system cannot be continuously stabilizable over this manifold. Thus a smooth generic system is never globally continuously stabilizable on a compact simply connected manifold.

Theorems 3 and 4 may give an impression that the topological barrier for continuous stabilization is provided only by the fact that either the equilibrium set or

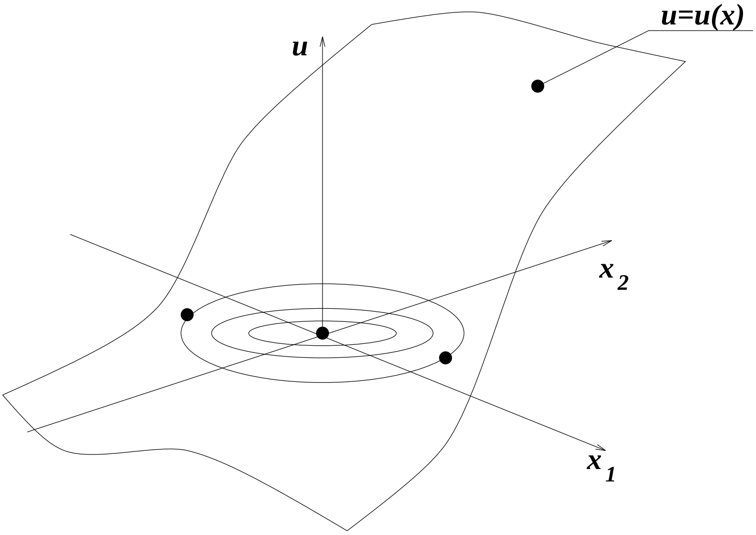


FIG. 2. *Over a compact set containing the circle* $(x_1^*)^2 + (x_2^*)^2 = \pi \cdot n$ *a graph of a continuous feedback stabilizing the system at* (x_1^*, x_2^*) *has at least two points of intersection with the circle.*

one of its connected components is bounded. However, stabilization also depends upon the way in which one connected component of the equilibrium set loops around another. Let us consider a single-input nonlinear system. We introduce the normal parameterization of $f^{-1}(0)$ as follows.

DEFINITION 3. *Let* $\zeta \subset \mathbf{R}_{\mathrm{x}}^{\mathrm{n}} \times \mathbf{R}_{\mathrm{u}}$ *be a segment of a connected component of* $f^{-1}(0)$, *and let*

$$\operatorname{rank}\left(\frac{\partial}{\partial x} f(x,u), \frac{\partial}{\partial u} f(x,u)\right) = n \quad \forall \ (x,u) \in \zeta.$$

Then a parameterization

$$\zeta = \{(x_\zeta(\tau), u_\zeta(\tau)) \in \mathbf{R}_{\mathrm{x}}^{\mathrm{n}} \times \mathbf{R}_{\mathrm{u}}; \ \ \tau \in \mathbf{R}\}$$

of the segment will be called normal if

$$\det\begin{pmatrix} \frac{dx_\zeta(\tau)}{d\tau} & \frac{du_\zeta(\tau)}{d\tau} \\ \frac{\partial}{\partial x} f(x_\zeta(\tau), u_\zeta(\tau)) & \frac{\partial}{\partial u} f(x_\zeta(\tau), u_\zeta(\tau)) \end{pmatrix} > 0 \qquad \forall \tau \in \mathbf{R}.$$

We can formulate the following necessary condition of smooth stabilization in the large.

THEOREM 5. *Let* $\dot{x} = f(x,u)$ *be a smooth single-input system such that*

$$\operatorname{rank}\left(\frac{\partial}{\partial x} f(x,u), \frac{\partial}{\partial u} f(x,u)\right) = n \quad \forall \ (x,u) \in \zeta,$$

where $\zeta \subset f^{-1}(0)$ *is a segment with the normal parameterization*

$$\zeta = \{(x_\zeta(\tau), u_\zeta(\tau)) \in \mathbf{R}_{\mathrm{x}}^{\mathrm{n}} \times \mathbf{R}_{\mathrm{u}}; \ \ \tau \in \mathbf{R}\}.$$

Then the system is not continuously stabilizable in the large at a point $(x^*, u^*) \in \zeta$ *whenever there is a connected component* $\tilde{\omega} \subset f^{-1}(0) \setminus \zeta$ *such that one can find points* $(\tilde{x}, \tilde{u}),\ \ (\hat{x}, \hat{u}) \in \tilde{\omega}$ *such that*

$$x_\zeta(\tau_1) = \hat{x}, \qquad \tau_1 < \tau^*,$$

$$x_\zeta(\tau_2) = \tilde{x}, \qquad \tau_2 > \tau^*,$$

and

$$\begin{aligned} u_\zeta(\tau_1) &\geq \hat{u}, \\ u_\zeta(\tau_2) &\leq \tilde{u}, \end{aligned} \tag{3}$$

where $\tau^* \in \mathrm{R}$ *and* $x_\zeta(\tau^*) = x^*, \quad u_\zeta(\tau^*) = u^*.$

The proof of this theorem is given in [10], [11].

If the system $\dot{x} = f(x,u)$ is continuously stabilizable at (x^*, u^*), then without loss of generality [8] we can assume that $f(x, u(x))$ is differentiable at x^* and

$$sign\left(\det\left(\frac{\partial}{\partial x} f(x, u(x))\right)|_{x=x^*}\right) = (-1)^n,$$

where $(x^*, u^*) \in f^{-1}(0)$ and $u^* = u(x^*)$.

Let $\zeta \subset f^{-1}(0)$ be a segment with the normal parameterization

$$\zeta = \{(x_\zeta(\tau), u_\zeta(\tau)) \in \mathrm{R}^{\mathrm{n}}_{\mathrm{x}} \times \mathrm{R}_{\mathrm{u}};\ \ \tau \in \mathrm{R}\}$$

and $\tau^* \in \mathrm{R},\ \ x_\zeta(\tau^*) = x^*,\ \ u_\zeta(\tau^*) = u^*.$

Then it follows from the definition of natural parameterization that there exists a positive real number α such that

$$\frac{d}{d\tau}(u_\zeta(\tau) - u(x_\zeta(\tau)))|_{\tau=\tau^*} = \left(\frac{du}{d\tau} - \frac{\partial u}{\partial x}\frac{dx}{d\tau}\right)|_{\tau=\tau^*}$$

$$= \alpha \cdot \det\begin{pmatrix} -\frac{\partial u}{\partial x} & 1 \\ \frac{\partial}{\partial x} f(x,u) & \frac{\partial}{\partial u} f(x,u) \end{pmatrix}|_{x=x^*,\ u=u^*}$$

$$= \alpha \cdot \det\begin{pmatrix} 0 & 1 \\ \frac{\partial}{\partial x}(f(x,u(x))) & \frac{\partial}{\partial u} f(x,u) \end{pmatrix}|_{x=x^*,\ u=u^*}$$

$$= \alpha \cdot (-1)^{n+2} \cdot \det\left(\frac{\partial}{\partial x}(f(x,u(x)))\right)|_{x=x^*} > 0,$$

since

$$sign\left(\det\left(\frac{\partial}{\partial x}(f(x,u(x)))\right)|_{x=x^*}\right) = (-1)^n.$$

Thus

$$u_\zeta(\tau) \leq u(x_\zeta(\tau)) \quad \text{for} \quad \tau \leq \tau^*$$

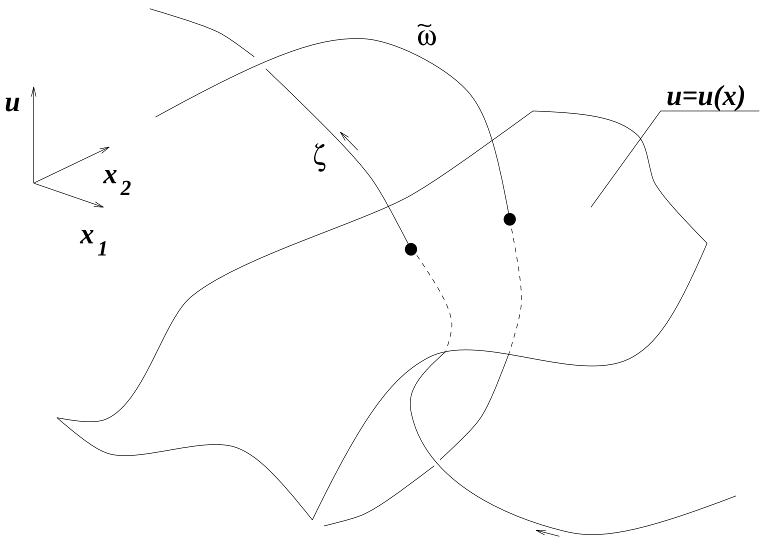


FIG. 3. *If $f^{-1}(0)$ has two connected components knotted as shown, then the graph of a continuous feedback $u = u(x)$ intersects $f^{-1}(0)$ more than once. Arrows show the normal parameterization of ζ.*

and

$$u_\zeta(\tau) \geq u(x_\zeta(\tau)) \quad \text{for } \tau \geq \tau^*.$$

Taking these inequalities into account and using the fact that $\tilde{\omega}$ is a connected component of the equilibrium set we conclude that under the conditions of Theorem 5 the graph of the feedback $u = u(x)$ has at least two points of intersection with the equilibrium set $f^{-1}(0)$. One of these points of intersection is (x^*, u^*) and another is on the connected component $\tilde{\omega}$ defined in Theorem 5. A geometrical illustration of this fact is Figure 3.

Example 2. Consider the system

$$\dot{x}_1 = (x_1 - \cos(u)) \cdot u,$$
$$\dot{x}_2 = (x_2 - \sin(u)).$$

The equilibrium set of this system is defined by

$$(x_1 - \cos(u)) \cdot u = 0,$$
$$x_2 - \sin(u) = 0.$$

It is easy to show that the system is locally stabilizable at the origin by a linear feedback law. On the other hand, our topological considerations, in particular Theorem 5, show that this system is not continuously stabilizable over any set containing the unit disk centered at the origin (Figure 4).

The discussion presented in this subsection shows that many systems are not continuously stabilizable. Hence it is reasonable and even sometimes necessary to design piecewise-constant stabilizers.

2.2. Discrete-time systems. This subsection shows that many discrete-time systems are not continuously stabilizable. Consider a discrete-time system of the form

$$x_{k+1} = f(x_k, u), \tag{4}$$

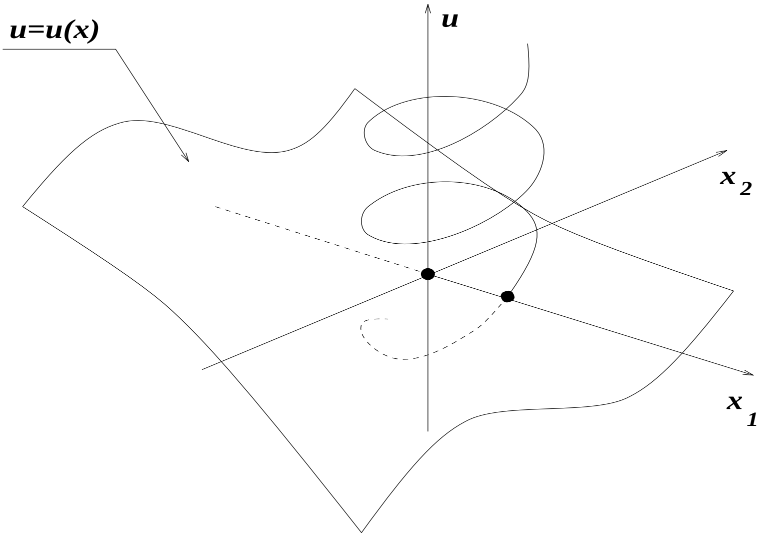


FIG. 4. *Due to the topological structure of the equilibrium set the continuous stabilization at the origin over a set containing the unit disk* $x_1^2 + x_2^2 \leq 1$ *is impossible.*

where $u \in \mathrm{R}^m$, $k \in \mathbf{N} = \{0, 1, 2, \ldots\}$, and for any $k \in \mathrm{N}$, $x_k \in \mathrm{R}^n$.

Let pr_x denote the projection of $\mathbf{R}_x^n \times \mathbf{R}_u^m$ onto $\mathbf{R}_x^n$, that is, $pr_x(x, u) = x$. $(pr_x - f)^{-1}(0)$ denotes the equilibrium set of system (4),

$$(pr_x - f)^{-1}(0) = \{(x, u) \in \mathbf{R}_x^n \times \mathbf{R}_u^m; \ \ x = f(x, u)\}.$$

System (4) is said to be continuously stabilizable at $(x^*, u^*) \in (pr_x - f)^{-1}(0)$ over a set $K \subset \mathrm{R}^n$ if there exists a continuous function $u = u(x)$, such that $u(x^*) = u^*$ and x^* is an asymptotically stable singular point of the closed loop system

$$x_{k+1} = f(x_k, u(x_k)), \tag{5}$$

and the domain of x^*-attraction contains K, i.e.,

$$\lim_{k \to +\infty} e^{kf} x = x^* \quad \forall \, x \in K \subset \mathbf{R}^n,$$

where $e^{kf}x$ is the solution of the closed loop system (5) with the initial condition $x_0 = x$. We say that x^* is stable if for any neighborhood W of x^* (open connected set containing x^*) there is a neighborhood V of x^*, such that

$$\forall \ \ k \in \mathbf{N} \quad e^{kf} V \subset W,$$

where $e^{kf}V = \{e^{kf}x; \ \ x \in V\}$.

A system that is continuously stabilizable at $(x^*, u^*) \in (pr_x - f)^{-1}(0)$ over $\mathbf{R}^n$ is said to be continuously stabilizable in the large.

The analogue of Theorem 3 for discrete-time systems is as follows.

THEOREM 6. *If* $f : \mathrm{R}_\mathrm{x}^\mathrm{n} \times \mathrm{R}_\mathrm{u}^\mathrm{m} \ \to \ \mathrm{R}^\mathrm{n}$ *is a smooth function and* $(pr_x - f)^{-1}(0)$ *is bounded, then the system* $x_{k+1} = f(x_k, u)$ *is not continuously stabilizable in the large at any point* $(x^*, u^*) \in (pr_x - f)^{-1}(0)$.

Proof. If $(pr_x - f)^{-1}(0)$ is bounded, then there is a positive real number r, such that

$$(pr_x - f)^{-1}(0) \subset \mathrm{Int}(B_r(0)).$$

Hence

$$f(x,u) \neq x \quad \forall u \in \mathrm{R}_u^m, \quad \mid x \mid = r,$$

and properties (1), (3) (Theorem 2) imply

$$d(P_x(\cdot) - f(\cdot, u(\cdot)), P_x(B_r(0)), 0) = d(P_x(\cdot) - f(x, \bar{u}), P_x(B_r(0)), 0) = 0,$$

where $\mid \bar{u} \mid = r$ and $u = u(x)$ is a continuous function, while

$$d(x - f(x, u(x)), P_x(B_r(0)), 0) = 1$$

whenever $u = u(x)$ is a continuous feedback stabilizing the system in the large. Thus the system cannot be continuously stabilized in the large at any point $(x^*, u^*) \in (pr_x - f)^{-1}(0)$; hence the theorem is proved. □

The analogue of Theorem 4 for discrete-time systems is as follows.

THEOREM 7. *Let* $f : \mathrm{R_x^n} \times \mathrm{R_u^m} \rightarrow \mathrm{R^n}$ *be a smooth function, and let* ω *be a connected component of* $(pr_x - f)^{-1}(0)$ *such that*

$$\mathrm{rank}\left(\mathrm{Id}_n - \frac{\partial}{\partial x} f(x,u), \frac{\partial}{\partial u} f(x,u)\right) = n \quad \forall \ (x,u) \in \omega,$$

where Id_n *is the identity matrix with* n *columns and* n *rows.*

If the system $x_{k+1} = f(x_k, u)$ *is continuously stabilizable in the large at* $(x^*, u^*) \in \omega$*, then* ω *is unbounded.*

Proof. If $u = v(x)$ is a continuous feedback stabilizing in the large the system $x_{k+1} = f(x_k, u)$ at a point $(x^*, u^*) \in \omega$ and ω is bounded, then there is $B_r(0)$ such that

$$\omega \subset \mathrm{Int}(B_r(0))$$

and

$$d(P_x(\cdot) - f(\cdot, v(\cdot)), P_x(B_r(0)), 0) = 1.$$

Note that

$$(pr_x - f)^{-1}(0) \cap \{(x,u) \in \mathrm{R}_x^n \times \mathrm{R}_u^m \ ; \ u = v(x)\} = \omega \cap \{(x,u) \in \mathrm{R}_x^n \times \mathrm{R}_u^m \ ; \ u = v(x)\}$$

and

$$\mid d(P_x(\cdot) - f(\cdot, v(\cdot)), P_x(B_r(0)), 0) \mid = \mid d((P_x - f, P_u - v \circ P_x), B_r(0), 0) \mid .$$

Making use of Lemma 1, we obtain the contradiction that proves the theorem. □

Theorems 6 and 7 have the same geometrical sense as Theorems 3 and 4 (Figure 1).

In order to formulate the analog of Theorem 5 for discrete-time systems we need to change the definition of natural parameterization.

Definition 4. *Let ζ be a segment of a connected component of $(pr_x - f)^{-1}(0)$, and let*

$$\operatorname{rank}\left(\mathrm{Id}_n - \frac{\partial}{\partial x}f(x,u), \frac{\partial}{\partial u}f(x,u)\right) = n \quad \forall \ (x,u) \in \zeta.$$

Then a parameterization

$$\zeta = \{(x_\zeta(\tau), u_\zeta(\tau)) \in \mathrm{R}^{\mathrm{n}}_{\mathrm{x}} \times \mathrm{R}_{\mathrm{u}} \quad ; \tau \in \mathrm{R}\}$$

of the segment will be called normal if

$$\det\begin{pmatrix} \frac{dx_\zeta(\tau)}{d\tau} & \frac{du_\zeta(\tau)}{d\tau} \\ \frac{\partial}{\partial x}f(x_\zeta(\tau), u_\zeta(\tau)) - \mathrm{Id}_n & \frac{\partial}{\partial u}f(x_\zeta(\tau), u_\zeta(\tau)) \end{pmatrix} > 0 \qquad \forall \tau \in \mathrm{R}.$$

The analogue of Theorem 5 for discrete-time systems is as follows.

Theorem 8. *Let $f(x,u)$ be such that*

$$\operatorname{rank}\left(\mathrm{Id}_n - \frac{\partial}{\partial x}f(x,u), \frac{\partial}{\partial u}f(x,u)\right) = n \quad \forall \ (x,u) \in \zeta,$$

where $\zeta \subset (pr_x - f)^{-1}(0)$ is a segment with a normal parameterization

$$\zeta = \{(x_\zeta(\tau), u_\zeta(\tau)) \in \mathrm{R}^{\mathrm{n}}_{\mathrm{x}} \times \mathrm{R}_{\mathrm{u}}; \ \ \tau \in \mathrm{R}\}.$$

Then the system $x_{k+1} = f(x_k, u)$ is not smoothly stabilizable in the large at a point $(x^*, u^*) \in \zeta$ whenever there is a connected component $\tilde{\omega} \subset (pr_x - f)^{-1}(0) \setminus \zeta$ such that one can find points $(\tilde{x}_2, \tilde{u}_2), (\tilde{x}_1, \tilde{u}_1) \in \tilde{\omega}$ such that

$$x_\zeta(\tau_1) = \tilde{x}_1, \qquad \tau_1 < \tau^*,$$

$$x_\zeta(\tau_2) = \tilde{x}_2, \qquad \tau_2 > \tau^*,$$

and

$$\begin{aligned} u_\zeta(\tau_1) &\geq \tilde{u}_1, \\ u_\zeta(\tau_2) &\leq \tilde{u}_2, \end{aligned} \tag{6}$$

where $\tau^* \in \mathrm{R}$ and $x_\zeta(\tau^*) = x^*, \quad u_\zeta(\tau^*) = u^*$.

The proof and geometrical illustration for this theorem coincide with that of Theorem 5 (Figure 3).

After a few technical modifications all topological necessary conditions of continuous stabilization also remain valid for discrete-time systems. Thus many discrete-time systems are not continuously stabilizable, and therefore we need to consider either piecewise-continuous or nonstationary stabilizing feedbacks. This fact serves as one of the main motivations for establishing the result presented in the next section. Moreover, the approach outlined in the next section can be effectively applied to design feedbacks subjected to state and control constraints.

3. Piecewise-constant stabilization criterion.

The goal of this section is to give the criterion of piecewise-constant stabilization. The proof of this criterion is constructive and can be used for feedback synthesis. Moreover, it is based on the general topology and therefore is valid for both discrete- and continuous-time systems. For the sake of brevity, only the criterion for continuous-time systems is presented.

Consider the system

$$\dot{x} = f(x,u), \tag{7}$$

where $x \in \mathrm{R}^{\mathrm{n}},\ u \in U\ \subset\ \mathrm{R}^{\mathrm{m}},\ U$ is a subset in R^{m}. $f(x,u)$ is a complete C^∞ vector field on R^{n} for every $u \in\ U\ \subset\ \mathrm{R}^{\mathrm{m}}$ fixed.

$PC(U)$ is the set of all piecewise-constant mappings form R^{n} into $U\ \subset\ \mathrm{R}^{\mathrm{m}}$. A function $u = u(x)$ is called piecewise-constant on some set Q if there exists a covering $Q \subseteq \bigcup_i V_i$ such that

(i) $V_i \cap V_j = \emptyset$ when $i \neq j$;

(ii) Int $V_i \neq \emptyset\ \forall i$;

(iii) the closure $\overline{\text{Int } V_i}$ of the interior Int V_i coincides with $\overline{V_i}$;

(iv) the restriction $u|_{V_i}$ of u to V_i is a constant from U.

Let $u(x) \in PC(U)$. Then we define the solution for the initial value problem

$$\dot{x} = f(x,u(x)), \quad x(0) = x_0 \tag{8}$$

as follows.

Definition 5. *A function of time $x_u(t,x_0)$ $(t > 0,\ \ t \in \mathrm{R})$ is called a solution for the initial value problem* (8) *if the following conditions hold:*

(i) *$x_u(t,x_0)$ is a continuous function of time,*

(ii) *$\forall\ T > 0$ the derivative $\frac{d}{dt}x_u(t,x_0)$ exists, and*

(iii)

$$\frac{d}{dt}x_u(t,x_0) = f(x_u(t,x_0), u(x_u(t,x_0)))$$

either $\forall t \in [0,T]$ or $\forall t \in [0,T]$ excluding a finite number of points.

Having fixed the feedback $u = u(x)$ such that $u(x) \in PC(U)$ and the solution for (7) exists for all $x_0 \in \mathrm{R}^{\mathrm{n}}$ we obtain the flow e^{tf} generated by the closed loop system

$$\dot{x} = f(x,u(x))$$

evolving over R^{n}. $e^{tf}x_0$ denotes the point into which the flow e^{tf} steers x_0 and $e^{tf}(V) = \{e^{tf}x_0; x_0 \in V\}$. V is called an invariant set of the system if and only if $e^{tf}V \subseteq V\ \forall t \geq 0$.

Definition 6. *A system $\dot{x} = f(x,u)$ is said to be piecewise-constantly stabilizable at $(x^*,u^*) \in f^{-1}(0)$ over a domain $K \subseteq \mathrm{R}^{\mathrm{n}}$, if there exists a piecewise-constant feedback $u = u(x)$ from $PC(U)$, such that $u(x^*) = u^*, x^*$ is an asymptotically stable equilibrium of* (8) *and $K \subseteq D(x^*)$, where $D(x^*)$ is the domain of x^*-attraction, i.e., for every $x_0 \in D(x^*)$ the solution $e^{tf}x_0$ of the closed loop system exists $\forall t \geq 0$ and $\lim_{t\to\infty} e^{tf}x_0 = x^*$.*

If $K = \mathrm{R}^{\mathrm{n}}$, then $\dot{x} = f(x,u)$ is called (completely) piecewise-constantly stabilizable at $(x^*,u^*) \in f^{-1}(0)$ (over R^{n} or in the large). If there exists a neighborhood $O(x^*)$ and $\dot{x} = f(x,u)$ is piecewise-constantly stabilizable in $(x^*,u^*) \in f^{-1}(0)$ over $O(x^*)$, then $\dot{x} = f(x,u)$ is said to be locally piecewise-constantly stabilizable at (x^*,u^*).

A control $u : [0,T] \to U$ is said to be piecewise constant if there exist time points $0 = t_0 < t_1 < t_2 < \cdots < t_N = T$ and $u_1, \ldots, u_N \in U$, such that $u(t) = u_i$ for $t_{i-1} \le t < t_i$ $(i < N)$ and $u(t) = u_N$ for $t_{N-1} \le t \le t_N$.

Definition 7. *We say that a point p can be piecewise-constantly steered into a point q, if there exist $0 < T < \infty$ and a piecewise-constant control $u : [0,T] \to U$, such that the solution $x_u(t,p)$ of the initial value problem*

$$\begin{aligned} \dot{x} &= f(x, u(t)), \\ x(0) &= p \end{aligned}$$

exists on the time interval $[0,T]$, is unique, and $x_u(T,p) = q$. If for every point $p \in K \subset \mathrm{R}^{\mathrm{n}}$ there exists $q \in V \subset \mathrm{R}^{\mathrm{n}}$, such that p is piecewise-constantly steered into q, then the set V is called piecewise-constantly accessible from the set K.

The piecewise-constant accessibility has a very important property formulated in the following lemma.

Lemma 2. *Let K be a compact subset in R^{n}. Let $V \subset \mathrm{R}^n$ be an open subset. Then if V is piecewise-constantly accessible from K there exist a natural number N, real number $T > 0$, and*

$$\{u_1, \ldots, u_N\} \in U^N = \underbrace{U \times \cdots \times U}_{N}$$

such that for each point $p \in K$ there exist a point $q \in V$ and $t_1 \ge 0, \ldots, t_N \ge 0$, $\sum_{i=1}^N t_i \le T$, such that

$$e^{-t_1 f(*,u_1)} \circ e^{-t_2 f(*,u_2)} \circ \cdots \circ e^{-t_N f(*,u_N)} q = p,$$

where $f(*, u_i)$ denotes the vector field $f(x, u_i)$ $(i = 1, 2, \ldots, N)$.

Proof. The set

$$V(n,u,T) = \left\{ e^{-t_1 f(*,u_1)} \circ e^{-t_2 f(*,u_2)} \circ \cdots \circ e^{-t_n f(*,u_n)} z : \right. \\ \left. t_i \ge 0 \ (i = 1, 2, \ldots, n) \sum_{i=1}^n t_i \le T, z \in V \right\}$$

is open $\forall n = 1, 2, \ldots,\ \ T > 0,\ u \in U^n$. The piecewise-constant accessibility of V from K implies

$$K \subset \bigcup_{n=1}^{\infty} \bigcup_{\substack{u \in U^n \\ T > 0}} V(n,u,T).$$

Thus it follows from the compactness of K that there exists a natural number μ such that

$$K \subset \bigcup_{i=1}^{\mu} V(n_i, u_i, T_i),$$

where $u_i = (u_{i1}, u_{i2}, \ldots, u_{in_i})$ $(i = 1, 2, \ldots, \mu)$. Therefore we can take $N = \sum_{i=1}^{\mu} n_i$,

$$\begin{aligned}&\{u_1, \dots, u_N\}\\&= \{u_{11}, u_{12}, \dots, u_{1n_1}, u_{21}, u_{22}, \dots, u_{2n_2}, \dots, u_{\mu n_\mu}\}\end{aligned}$$

and $T = \sum_{i=1}^{\mu} T_i$. □

To formulate the main result of this paper we need the following analogue of the Liapunov asymptotic stability.

DEFINITION 8. *An equilibrium $(x^*, u^*) \in f^{-1}(0)$ of $\dot{x} = f(x, u)$ is said to be Liapunov asymptotically stable if there exists $\Delta > 0$ such that for any $0 \le \varepsilon \le \Delta$ one can find $\delta > 0$ such that x^* is piecewise-constantly accessible from $B_\delta(x^*)$ without leaving $B_\varepsilon(x^*)$; i.e., for any $p \in B_\delta(x^*)$ there exist $0 < T \le \infty$ (T can be ∞) and a piecewise-constant control $u : [0, T) \to U$, such that the solution $x_u(t, p)$ of the initial value problem*

$$\begin{aligned}\dot{x} &= f(x, u(t)),\\ x(0) &= p\end{aligned}$$

exists on $[0, T)$, $x_u(t, p) \in B_\varepsilon(x^*)$ for any $t \in [0, T)$, and

$$\lim_{t \to T} x_u(t, p) = x^*.$$

THEOREM 9. *A system $\dot{x} = f(x, u)$ is piecewise-constantly stabilizable at its equilibrium $(x^*, u^*) \in f^{-1}(0)$ over a compact set K if and only if (x^*, u^*) is Liapunov asymptotically stable and x^* is piecewise-constantly accessible from K.*

Proof.

Necessity. If $u(x) \in PC(U)$ stabilizes $\dot{x} = f(x, u)$ at $(x^*, u^*) \in f^{-1}(0)$ over K, then (x^*, u^*) is evidently Liapunov asymptotically stable and x^* is piecewise-constantly accessible from K.

Sufficiency. Since $(x^*, u^*) \in f^{-1}(0)$ is Liapunov asymptotically stable, then there exists a sequence of positive real numbers $\{\varepsilon_n\}_{n=0}^{\infty}$ such that

$$\begin{gathered}\varepsilon_0 > \varepsilon_1 > \cdots > \varepsilon_n > \varepsilon_{n+1} > \cdots,\\ \lim_{n \to \infty} \varepsilon_n = 0\end{gathered}$$

and for any $n = 1, 2, 3, \dots$, $B_{\varepsilon_{n+1}}(x^*)$ is accessible from $B_{\varepsilon_n}(x^*)$ without leaving $B_{\varepsilon_{n-1}}(x^*)$.

Roughly speaking, the main idea of the proof is to design piecewise-constant feedbacks $u_0(x)$, $u_1(x)$, $u_2(x), \dots$ which steer the systems in accordance with the arrows marked in Figure 5.

The point x^* is piecewise-constantly accessible from K and therefore, for any $n = 0, 1, 2, \dots$, $B_{\varepsilon_n}(x^*)$ is piecewise-constantly accessible from K. Lemma 2 with V being the interior $\mathrm{Int}(B_{\varepsilon_2}(x^*))$ of the ball $B_{\varepsilon_2}(x^*)$ gives us the existence of a natural number N, real number $T > 0$, and

$$\{u_1, \dots, u_N\} \in U^N = \underbrace{U \times \cdots \times U}_{N}$$

such that for each point $p \in K$ there exist a point $q \in B_{\varepsilon_2}(x^*)$ and

$$t_1 \ge 0, \dots, t_N \ge 0, \qquad \sum_{i=1}^{N} t_i \le T$$

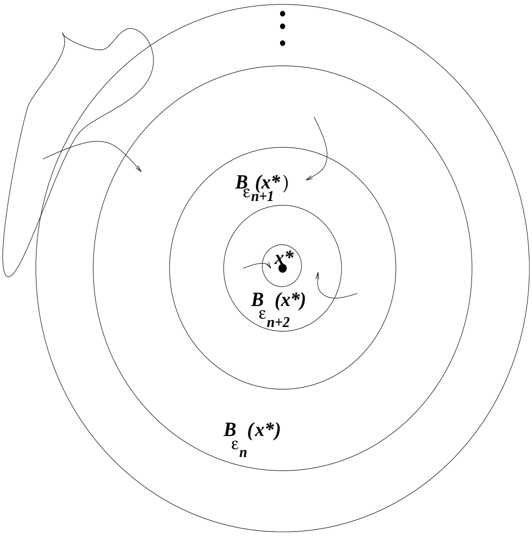


FIG. 5. *The closed loop system is supposed to move with respect to the sets* K, $\{B_{\varepsilon_n}(x^*)\}_{n=0}^{\infty}$ *in accordance with the arrows.*

such that

$$e^{-t_1 f(*,u_1)} \circ e^{-t_2 f(*,u_2)} \circ \cdots \circ e^{-t_N f(*,u_N)} q = p,$$

where $f(*, u_i)$ denotes the vector field $f(x, u_i)$ $(i = 1, 2, \ldots, N)$.

Consider the family of open sets generated by

$$V(t_1, t_2, \ldots, t_N) = e^{-t_1 f(*,u_1)} \circ e^{-t_2 f(*,u_2)} \circ \cdots \circ e^{-t_N f(*,u_N)} \mathrm{Int}(B_{\varepsilon_2}(x^*)),$$

where $t_1 \geq 0, \ldots, t_N \geq 0, \sum_{i=1}^{N} t_i \leq T$.

Due to Lemma 2

$$K \subset \bigcup_{t_1 \geq 0, \ldots, t_N \geq 0, \sum_{i=1}^{N} t_i \leq T} V(t_1, t_2, \ldots, t_N).$$

But K is a compact set, and hence one can choose a finite number of N-tuples $\{(t_{1j}, t_{2j}, \ldots, t_{Nj})\}_{j=1}^{\nu}$ such that

$$K \subset \bigcup_{j=1}^{\nu} V(t_{1j}, t_{2j}, \ldots, t_{Nj}). \tag{9}$$

The feedback $u_0(x) \in PC(U)$ which steers any point of $K \setminus B_{\varepsilon_2}(x^*)$ into the interior of the set $B_{\varepsilon_1}(x^*) \setminus B_{\varepsilon_2}(x^*)$ can be designed in the following way:

$$u_0(x) = u_N \text{ for } x \in \overline{\bigcup_{0<\tau\leq t_{N1}} V(0, \ldots, \tau)} \setminus B_{\varepsilon_2}(x^*),$$

$$u_0(x) = u_{N-1} \text{ for } x \in \overline{\bigcup_{0<\tau\leq t_{N-1\,1}} V(0, \ldots, \tau, t_{N1})} \setminus \left\{ \overline{\bigcup_{0<\tau\leq t_{N1}} V(0, \ldots, \tau)} \cup B_{\varepsilon_2}(x^*) \right\}$$

and so on until

$$\begin{aligned} u_0(x) = u_1 \text{ for } x \in {} & \overline{\bigcup_{0<\tau\le t_{11}} V(\tau, t_{21}, \dots, t_{N1})} \setminus \Bigg\{ \overline{\bigcup_{0<\tau\le t_{21}} V(0, \tau, \dots, t_{N1})} \\ & \bigcup \left(\overline{\bigcup_{0<\tau\le t_{31}} V(0, 0, \tau, \dots, t_{N1})} \right) \cup \cdots \cup \left(\overline{\bigcup_{0<\tau\le t_{N1}} V(0, \dots, \tau)} \right) \cup B_{\varepsilon_2}(x^*) \Bigg\}. \end{aligned}$$

Thus $u_0(x)$ is defined on the set

$$\begin{aligned} \Xi_1 = \Bigg[& \overline{\left(\bigcup_{0<\tau\le t_{11}} V(\tau, t_{21}, \dots, t_{N1}) \right)} \cup \overline{\left(\bigcup_{0<\tau\le t_{21}} V(0, \tau, \dots, t_{N1}) \right)} \\ & \cup \cdots \cup \overline{\left(\bigcup_{0<\tau\le t_{N1}} V(0, \dots, \tau) \right)} \Bigg] \setminus B_{\varepsilon_2}(x^*). \end{aligned}$$

If $K \setminus B_{\varepsilon_2}(x^*) \subseteq \Xi_1$, then the construction of $u_0(x)$ is completed. Otherwise, there exists $1 \le i \le \nu$ for which

$$V(t_{1i}, t_{2i}, \dots, t_{Ni}) \setminus B_{\varepsilon_2}(x^*) \not\subseteq \Xi_1.$$

We put

$$\begin{aligned} & u_0(x) = u_N \text{ for } x \in \overline{\bigcup_{0<\tau\le t_{Ni}} V(0, \dots, \tau,)} \setminus (\Xi_1 \cup B_{\varepsilon_2}(x^*)), \\ & u_0(x) = u_{N-1} \text{ for } x \in \overline{\bigcup_{0<\tau\le t_{N-1i}} V(0, 0, \dots, \tau, t_{Ni})} \setminus \Bigg\{ \Xi_1 \cup \overline{\left(\bigcup_{0<\tau\le t_{Ni}} V(0, \dots, \tau,) \right)} \\ & \qquad \qquad \cup B_{\varepsilon_2}(x^*) \Bigg\} \\ & \quad \vdots \\ & u_0(x) = u_1 \text{ for } x \in \overline{\bigcup_{0<\tau\le t_{1i}} V(\tau, t_{2i}, \dots, t_{Ni})} \setminus \Bigg\{ \Xi_1 \cup \overline{\left(\bigcup_{0<\tau\le t_{Ni}} V(0, \dots, \tau,) \right)} \\ & \qquad \cup \cdots \cup \overline{\left(\bigcup_{0<\tau\le t_{2i}} V(0, \tau, \dots, t_{Ni}) \right)} \cup B_{\varepsilon_2}(x^*) \Bigg\}. \end{aligned}$$

Thus we have defined $u_0(x)$ on the set

$$\Xi_2 = \left[\overline{\left(\bigcup_{0<\tau\le t_{1i}} V(\tau, \dots, t_{Ni}) \right)} \cup \overline{\left(\bigcup_{0<\tau\le t_{2i}} V(0, \tau, \dots, t_{Ni}) \right)} \cup \cdots \cup \Xi_1 \right] \setminus B_{\varepsilon_2}(x^*).$$

If $K \setminus B_{\varepsilon_2}(x^*) \subseteq \Xi_2$, then $u_0(x) \in PC(U)$ is constructed. Otherwise there exists j such that $j \ne 1, j \ne i$, and $V(t_{1j}, \dots, t_{Nj}) \setminus B_{\varepsilon_2}(x^*) \not\subseteq \Xi_2$. We can define the feedback

$u_0(x)$ on the set

$$\Xi_3 = \left[\overline{\left(\bigcup_{0<\tau\le t_{1j}} V(\tau,\dots,t_{Nj}) \right)} \cup \overline{\left(\bigcup_{0<\tau\le t_{2j}} V(0,\tau,\dots,t_{Nj}) \right)} \cup \dots \cup \Xi_2 \right] \setminus B_{\varepsilon_2}(x^*)$$

in the same way as it has been done on the set Ξ_2.

The existence of a natural number μ such that $K \setminus B_{\varepsilon_2}(x^*) \subseteq \Xi_\mu$ follows from (9). Thus the construction of $u_0(x)$ is completed after a finite number of steps. Let Λ_0 be the domain where $u_0(x)$ is defined. By construction, if a trajectory of the system closed by $u_0(x)$ starts at a point of Λ_0, then it will reach the set $B_{\varepsilon_1}(x^*) \setminus \mathrm{Int}(B_{\varepsilon_2}(x^*))$ in a finite time.

The set $B_{\varepsilon_1}(x^*) \setminus \mathrm{Int}(B_{\varepsilon_2}(x^*))$ is compact, and therefore we can employ the method used to construct $u_0(x)$ for designing $u_1(x) \in PC(U)$ which steers any point from $B_{\varepsilon_1}(x^*) \setminus \mathrm{Int}(B_{\varepsilon_2}x^*))$ into $\mathrm{Int}(B_{\varepsilon_2}(x^*)) \setminus B_{\varepsilon_3}(x^*)$. We denote by Λ_1 the domain where $u_1(x)$ is defined. Since (x^*, u^*) is Liapunov asymptotically stable, one can construct $u_1(x) \in PC(U)$ so that $\Lambda_1 \subset B_{\varepsilon_0}(x^*)$. Moreover, if a trajectory of the system closed by $u_1(x)$ starts at a point of Λ_1, then it will reach the set $B_{\varepsilon_2}(x^*) \setminus \mathrm{Int}(B_{\varepsilon_3}(x^*))$ in finite time.

We proceed in this way and obtain for each $n = 1, 2, 3, \dots$ the feedback $u_n(x) \in PC(U)$ defined on $\Lambda_n \subset B_{\varepsilon_{n-1}}(x^*)$ and $B_{\varepsilon_n}(x^*) \setminus \mathrm{Int}(B_{\varepsilon_{n+1}}(x^*)) \subset \Lambda_n$. The feedback $u_n(x)$ steers the system from any state in Λ_n into $\mathrm{Int}(B_{\varepsilon_{n+1}}(x^*)) \setminus B_{\varepsilon_{n+2}}(x^*)$.

Hence the function $u(x) \in PC(U)$ which stabilizes $\dot{x} = f(x, u)$ at x^* over the compact set K can be defined as

$$\lim_{n\to\infty} w_n(x) = u(x),$$

where $w_n(x)$ is given by

$$\begin{aligned} w_n(x) &= u_n(x) \quad \text{for } x \in \Lambda_n, \\ w_n(x) &= u_{n-1}(x) \quad \text{for } x \in \Lambda_{n-1} \setminus \Lambda_n, \\ &\vdots \\ w_n(x) &= u_0(x) \quad \text{for } x \in \Lambda_0 \setminus \bigcup_{j=1}^{n} \Lambda_j, \\ w_n(x) &= u^* \quad \text{for } x \notin \bigcup_{j=0}^{n} \Lambda_j. \end{aligned}$$

The proof is completed. □

Definitions 6, 7, and 8 and Lemma 2 admit natural formulations for discrete-time systems. Moreover, the main ideas in the proof of Theorem 9 are topological and, after minor technical adjustments, they lead us to the following analogue of Theorem 9 for discrete-time systems.

THEOREM 10. *A system $x_{k+1} = f(x_k, u)$ is piecewise-constantly stabilizable at its equilibrium $(x^*, u^*) \in (pr_x - f)^{-1}(0)$ over a compact set K if and only if (x^*, u^*) is Liapunov asymptotically stable and x^* is piecewise-constantly accessible from K.*

The proof of Theorem 9 remains valid for much more general objects than discrete-time systems. The method of the proof outlines a general framework which leads to the synthesis of control systems defined over a finite algebra, in particular Boolean algebra. Such systems very often are called algorithms and arise in numerous software engineering problems. To demonstrate the use of this framework in software engineering the author developed a simple game called Corners. (The Java version of Corners is posted at http://lagrange.la.asu.edu/docs/Corners.) The analysis of algorithms from the point of view of a control theorist is beyond the scope of this paper. Here we present only some examples of the synthesis of piecewise constant stabilizers.

Although in the next example the construction of the piecewise-constant feedback seems completely ad hoc, it contains an effective recipe which, if combined with topological methods developed in [10], leads us to an effective stabilizer synthesis procedure for two-dimensional affine nonlinear systems.

Example 3. Consider the linear system

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= u, \end{aligned} \tag{10}$$

where $(x_1, x_2) \in \mathbf{R}^2$ and $u \in \mathbf{R}$. System (10) is continuously stabilizable at the origin by a linear feedback. Hence by Theorem 9 it is stabilizable by a piecewise-constant feedback as well. We look for a piecewise-constant feedback stabilizer having the following form:

$$u(x) = -k_1 \cdot \theta(x_1) - k_2 \cdot \theta(x_2) \quad \text{for } x \neq 0 \tag{11}$$

and

$$u(0) = 0,$$

where $\theta(\tau)$ is the function defined as

$$\theta(\tau) = \begin{cases} -1 & : \quad \tau < 0, \\ 1 & : \quad \tau \geq 0. \end{cases} \tag{12}$$

If the real numbers $k_1,\ k_2$ satisfy the inequality

$$k_1 > k_2 > 0,$$

then the feedback (11) globally stabilizes the system (10) at the origin. Indeed, one can show that the function

$$V(x_1, x_2) = k_1 \cdot \mid x_1 \mid + \frac{x_2^2}{2}$$

is monotonically decreasing along the trajectories of the system (10) closed by the feedback (11).

Example 3 combined with the topological methods from [10] allows us to construct a piecewise-constant stabilizer for any controllable affine nonlinear system

$$\dot{x}_1 = f_1(x_1, x_2) + b_1(x_1, x_2) \cdot u,$$
$$\dot{x}_2 = f_2(x_1, x_2) + b_2(x_1, x_2) \cdot u$$

satisfying the following conditions:

(i) $f_1(x_1, x_2), f_2(x_1, x_2), b_1(x_1, x_2), b_2(x_1, x_2)$ are continuously differentiable functions.

(ii)

$$\left(b_1(x_1,x_2)\cdot\frac{\partial}{\partial x_1}\varphi(x)+b_2(x_1,x_2)\cdot\frac{\partial}{\partial x_2}\varphi(x)\right)|_{x\in\varphi^{-1}(0)}\neq 0,$$

where

$$\varphi(x)=f_1(x_1,x_2)b_2(x_1,x_2)-f_2(x_1,x_2)b_1(x_1,x_2)$$

and

$$\varphi^{-1}(0)=\{x\in\mathbf{R}^2:\ \ \varphi(x)=0\}.$$

Example 4. In this example the system

$$\begin{array}{rcl}\dot{x}_1 & = & x_2,\\ \dot{x}_2 & = & a(x)+b(x)\cdot u\end{array}\tag{13}$$

is globally stabilized at the origin by means of a piecewise-constant feedback. We assume that $\forall x\in\mathbf{R}^2$ the functions $a(x)$ and $b(x)$ satisfy the following inequalities:

$$\mid a(x)\mid\leq a_+,$$
$$0<b_-\leq b(x)\leq b_+.$$

It is known [10] that system (13) can be stabilized by a time-dependent continuous feedback. Hence by Theorem 9 there exists a piecewise-constant stabilizer for (13). Choose positive real numbers α, β so that

$$\alpha\cdot b_->\beta\cdot b_+>0$$

and

$$\alpha\cdot b_->2\cdot a_++(2\cdot b_++a_+)\cdot\beta.$$

Then the piecewise-constant feedback

$$u(x)=\left\{\begin{array}{rcl}-k_1\theta(x_1)-k_2\theta(x_2) & : & x\neq 0,\\ -\frac{a(0)}{b(0)} & : & x=0,\end{array}\right.$$

with $\theta(\tau)$ being defined in (12) and

$$k_1=\frac{\alpha+\beta}{2},\qquad k_2=\frac{\alpha-\beta}{2},$$

globally stabilizes the system at the origin.

Indeed, the function

$$V(x)=\frac{\alpha\cdot b_--a_+\cdot\beta}{2}\mid x_1\mid+\frac{x_2^2}{2}$$

is monotonically decreasing along the trajectories of the closed loop system, and therefore the origin is globally asymptotically stable.

Using the topological methods based on so-called stable covering (see [10]) one can construct piecewise-constant stabilizers for a generic nonlinear affine system with an equilibrium set having several connected components.

The next example illustrates the use of Theorem 9 in the stabilizability analysis of a nonholonomic system.

Example 5. Consider the system

$$\begin{aligned} \dot{x}_1 &= u, \\ \dot{x}_2 &= v, \\ \dot{x}_3 &= x_1 v - x_2 u. \end{aligned} \tag{14}$$

It is stated in [2] that system (14) is not continuously stabilizable at the origin. Our goal is to analyze its piecewise-constant stabilizability.

Let us show that the origin is Liapunov asymptotically stable. Indeed, for any point $P = (x_1, x_2, x_3) \in \mathrm{R}^3$ the piecewise-constant control

$$\begin{aligned}
(u(t), v(t)) &= (-x_1, -x_2) \quad \text{for } t \in [0,1], \\
(u(t), v(t)) &= \left(\sqrt{\frac{|x_3|}{2}}, 0\right) \quad \text{for } t \in [1,2], \\
(u(t), v(t)) &= \left(0, -sign(x_3) \cdot \sqrt{\frac{|x_3|}{2}}\right) \quad \text{for } t \in [2,3], \\
(u(t), v(t)) &= \left(-\sqrt{\frac{|x_3|}{2}}, 0\right) \quad \text{for } t \in [3,4], \\
(u(t), v(t)) &= \left(0, sign(x_3) \cdot \sqrt{\frac{|x_3|}{2}}\right) \quad \text{for } t \in [4,5]
\end{aligned}$$

steers system (14) from the point P into the origin. First, this control moves the system from P into $(0, 0, x_3)$ along the straight line. Then $(x_1(t), x_2(t))$ traces the boundary of the square with the side of length $\sqrt{|x_3|/2}$ and moves system (14) from $(0, 0, x_3)$ into the origin.

Thus the origin is Liapunov asymptotically stable for system (14) and all conditions of Theorem 9 are satisfied. Hence system (14) is piecewise-constantly stabilizable at the origin over any compact set from R^3.

The next example illustrates the use of Theorem 10.

Example 6. Consider a discrete-time system of the form

$$\begin{aligned} x_1(k+1) &= \tfrac{1}{2}(1 + \sin(x_1^2(k) + u^2(k)))x_1(k), \\ x_2(k+1) &= 2 \cdot x_2(k) - u(k). \end{aligned} \tag{15}$$

Let us analyze stabilizability of system (15) at the origin. By Theorem 8 system (15) is not continuously stabilizable at the origin. On the other hand, for any x_1 one can find $\bar{u}$ such that $\sin(x_1^2 + \bar{u}^2) = -1$. Then on the next step $u = 2x_2$ will steer system (15) to the origin. Thus all conditions of Theorem 10 are satisfied and system (15) is piecewise-constantly stabilizable at the origin over any compact set from R^2.

4. Conclusion. This paper presents geometrical illustrations of the topological necessary conditions of continuous stabilization obtained in [10], [11]. It follows from these conditions that many systems are not stabilizable by continuous stationary feedbacks. Since in the majority of control applications one needs to implement a control system with the help of digital processors, it is natural to design piecewise-constant stabilizing feedbacks. We have proved the criterion of stabilization in the class of piecewise-constant feedbacks and discussed their synthesis for affine nonlinear two-dimensional systems.

It follows from Theorem 9 that any system that is stabilizable by a continuous feedback $u = u(t, x)$ also can be stabilized by a piecewise-constant feedback of the form $u = u(x)$. Moreover, the proof of Theorem 9 leads to the synthesis procedure for stabilizing feedbacks subjected to state and control constraints.

The approach proposed in this paper can be applied to deal with a wide variety of control objects that are much more general than those described by differential or discrete-time equations. For example, one can use this approach for software development projects. In fact, we already have tested the methodology provided by the proof of Theorem 9 and have developed a simple computer game. To use this methodology for computer algorithm synthesis one needs to choose some natural topology and then follow the proof of Theorem 9. In a computer game or a program where one needs to minimize some cost function this topology naturally is introduced by this cost function. In general the choice of an appropriate topology raises a number of difficult questions that can serve as topics for further research.

Acknowledgments. The author is grateful to Anton Nikitin for his editorial remarks and for reading the early version of this manuscript. The author is much obliged to anonymous referees for valuable suggestions. Some ideas presented in this paper were generated during (or shortly after) the author's visit to the University of Gent in June 1996. The author would like to thank Professor D. Aeyels for his kind support of this research and his hospitality.

REFERENCES

[1] D. Aeyels, *Stabilization of a class of nonlinear systems by a smooth feedback control*, Systems Control Lett., 5 (1985), pp. 289–294.

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

[3] J. M. Coron, *Necessary condition for feedback stabilization*, Systems Control Lett., 14 (1990), pp. 227–232.

[4] J. M. Coron, *Global asymptotic stabilization for controllable systems without drift*, Math. Control Signals Systems, 5 (1992), pp. 295–312.

[5] S. V. Emelyanov, S.K. Korovin, and I.G. Mamedov, *Variable-Structure Control Systems: Discrete and Digital*, Mir Publishers, Moscow, CRC Press, Boca Raton, FL, 1995.

[6] M. W. Hirsch, *Differential Topology,* Springer-Verlag, Berlin, New York, 1976.

[7] U. Itkis, *Control Systems of Variable Structure,* John Wiley, New York, c1976.

[8] N. G. Lloyd, *Degree Theory,* Cambridge University Press, London, UK, 1978.

[9] S. Nikitin, *Piecewise Smooth Stabilizing Extension*, in Proc. 2nd European Control Conference, ECC '93, Groningen, the Netherlands, pp. 50–52.

[10] S. Nikitin, *Global Controllability and Stabilization of Nonlinear Systems*, World Scientific, River Edge, NJ, 1994.

[11] S. Nikitin, *Topological necessary conditions of smooth stabilization in the large*, Systems Control Lett., 21 (1993), pp. 35–41.

[12] J. B. Pomet, *Explicit design of time-varying stabilizing control laws for a class of controllable systems without drift*, Systems Control Lett., 18 (1992), pp. 147–158.

[13] O. J. Sordalen and O. Egeland, *Exponential stabilization of nonholonomic chained systems*, IEEE Trans. Automat. Control, 35 (1995), pp. 35–49

[14] C. Samson, *Control of chained systems application to path following and time-varying point-stabilization of mobile robots*, IEEE Trans. Automat. Control, 40 (1995), pp. 64–77.

[15] A. F. Filippov, *Differential Equations with Discontinuous Right-Hand Sides*, Kluwer Academic Publishers, Norwell, MA, 1988.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 3, pp. 934–956

BIFURCATION STABILIZATION WITH LOCAL OUTPUT FEEDBACK[*]

GUOXIANG GU[†], XIANG CHEN[‡], ANDREW G. SPARKS[§], AND SIVA S. BANDA[§]

Abstract. Local output feedback stabilization with smooth nonlinear controllers is studied for parameterized nonlinear systems for which the linearized system possesses either a simple zero eigenvalue or a pair of imaginary eigenvalues and the bifurcated solution is unstable at the critical value of the parameter. It is assumed that the unstable mode corresponding to the critical eigenvalue of the linearized system is not linearly controllable. Results are established for bifurcation stabilization using output feedback where the critical mode can be either linearly observable or linearly unobservable. The stabilizability conditions are characterized in explicit forms that can be used to synthesize stabilizing controllers. The results obtained in this paper are applied to rotating stall control for axial flow compressors as an application example.

Key words. nonlinear systems, bifurcations, projection method, bifurcation stabilization, linear controllability/observability

AMS subject classifications. 93C10, 93C15, 93C60

PII. S0363012997320924

1. Introduction.

Stabilization of nonlinear control systems with smooth state feedback control has been studied by a number of people [4, 2, 3, 9, 19]. An interesting situation for nonlinear stabilization is the case when the linearized system has uncontrollable modes on the imaginary axis with the rest of the modes stable. This is the so-called *critical case* for which linearization theory is inadequate. It becomes more intricate if the underlying nonlinear system involves a real-valued parameter. At critical values of the parameter, the linearized system has unstable modes corresponding to eigenvalues on the imaginary axis and additional equilibrium solutions will be born. The bifurcated solutions may or may not be stable. The instability of the bifurcated solution may cause a "hysteresis loop" in bifurcation diagrams for both subcritical pitchfork bifurcations and Hopf bifurcations [11] and induce undesirable physical phenomena. This is manifested by rotating stall in axial flow compressors [1, 17, 18], which has received great attention for the past several years [5, 6, 7, 15, 16, 20]. Thus bifurcation stabilization is an important topic in nonlinear control.

Abed and Fu studied bifurcation stabilization using smooth local state feedback control [2, 3]. For a Hopf bifurcation, stabilization conditions were obtained for both the case where the critical modes of the linearized system are controllable and uncontrollable. For a stationary bifurcation, stabilization conditions were derived for the case where the critical mode of the linearized system is controllable. The uncontrollable case was investigated in [13], where normal forms of the nonlinear system are used. In this paper we study bifurcation stabilization via local output feedback

[*] Received by the editors May 2, 1997; accepted for publication (in revised form) April 28, 1998; published electronically March 22, 1999. This research was supported in part by grants from AFOSR, ARO, and LEQSF.

http://www.siam.org/journals/sicon/37-3/32092.html

[†] Department of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA 70803-5901 (ggu@ee.lsu.edu).

[‡] Department of Electrical Engineering, Lakehead University, 955 Oliver Road, Thunder Bay, Ontario, Canada P7B 5E1 (xchen@sleet.lakeheadu.ca).

[§] Control Analysis Section, WL/FIGC, 2210 Eighth Street, Suite 21, Wright-Patterson Air Force Base, OH 45433-7531 (sparks@batman.flight.wpafb.af.mil, banda@wl.wpafb.af.mil).

control laws that are smooth. Throughout the paper it is assumed that the critical mode of the linearized system is uncontrollable. Moreover only output measurement is available for feedback. It should be clear that measurement of all state variables is unrealistic in practice, especially when the underlying nonlinear system has a high order. Often some of state variables are more expensive or more difficult to measure than others. Hence the bifurcation stabilization problem studied in this paper has more engineering significance.

Two stabilization issues will be investigated. The first one is concerned with bifurcation stabilization where the critical mode of the linearized system is unobservable through linearized output measurement. Stabilizability conditions are established for both stationary bifurcations and Hopf bifurcations. Roughly speaking, it is shown that nonlinear controllers do not offer any advantage over the linear ones for bifurcation stabilization. Stabilizing controllers, if they exist, can be taken as linear ones. The second one is concerned with bifurcation stabilization where the critical mode of the linearized system is observable through linearized output measurement that includes state feedback as a special case. Stabilizability conditions are also obtained in this case. It is shown that linear controllers are adequate for stabilization of transcritical bifurcations and quadratic controllers are adequate for stabilization of pitchfork and Hopf bifurcations, respectively. It should be pointed out that the stabilization conditions obtained in this paper are characterized in an explicit form that provides synthesis procedures for design of stabilizing controllers, if they exist. Rotating stall control for axial flow compressors will be used as an application example to demonstrate the use of the stabilization results established in this paper.

The notation in this paper is standard. The collections of real and complex numbers are denoted by $\mathbf{R}$ and $\mathbf{C}$, respectively. If $c \in \mathbf{C}$, its complex conjugate is denoted by $\bar{c}$. The collection of real and complex vectors with size n are denoted by $\mathbf{R}^n$ and $\mathbf{C}^n$, respectively. A matrix M of size $p \times m$ can be viewed as a linear map from $\mathbf{C}^m$ to $\mathbf{C}^p$, and its transpose is denoted by M^T. For $m = p = n$, M is said to be stable if all its eigenvalues are in the open left half-plane. Notions such as linear controllability and linear observability can be found in [12].

2. Local bifurcation stability and projection method.

This section considers the stability issue for a bifurcated system using the projection method developed in [11]. The system under consideration is the following nth-order parametrized nonlinear system:

$$(1) \qquad \dot{x} = f(\gamma, x), \qquad f(\gamma, x_e) = 0 \;\; \forall\, \gamma \in (-\delta,\ \delta),$$

where $x \in \mathbf{R}^n$, γ is a real-valued parameter, and $\delta > 0$ is a sufficiently small real number. Without loss of generality, we can assume $x_e = 0$, i.e., $f(\gamma, 0) = 0$ in a small neighborhood of $\gamma = 0$, which is called the zero solution. The linearized system at the zero solution is given by

$$(2) \qquad \dot{x}_0 = L(\gamma)x_0, \qquad L(\gamma) = \left.\frac{df(\gamma, x)}{dx}\right|_{x=x_e=0}.$$

If $L(0)$ has one or more eigenvalues on the imaginary axis, then additional nonzero equilibrium solutions or bifurcated solutions will be born at $\gamma = 0$. It is assumed that $f(\cdot,\cdot)$ is sufficiently smooth such that the bifurcated solution $x_e \neq 0$, satisfying $f(\gamma, x_e) = 0$, is a smooth function of γ.

DEFINITION 2.1. *The nonlinear system in* (1) *is said to have local bifurcation stability if the bifurcated solution is locally asymptotically stable for sufficiently small* γ.

Two different types of bifurcations will be considered in this paper, and the determination of their local stability will be discussed in the following two subsections.

2.1. Local stability for stationary bifurcations. For stationary bifurcations, it is assumed that $L(\gamma)$ possesses a simple eigenvalue $\lambda(\gamma)$, depending smoothly on γ, satisfying

$$\lambda(0) = 0, \qquad \lambda'(0) = \frac{d\lambda}{d\gamma}(0) \neq 0, \tag{3}$$

while all other eigenvalues are stable in a neighborhood of $\gamma = 0$. It implies that the zero solution changes its stability as γ crosses 0. For instance, $\lambda'(0) < 0$ implies that the zero solution is locally stable for $\gamma > 0$ and becomes unstable for $\gamma < 0$. Furthermore additional equilibria $x_e \neq 0$ born at $\gamma = 0$ are smooth functions of γ by the smoothness of $f(\cdot,\cdot)$. Such bifurcated solutions are independent of time t and called stationary bifurcation. Thus $\gamma = 0$ is the critical value of the parameter and $\lambda(\gamma)$ is called the critical eigenvalue. The nonlinear system (1) at $\gamma = 0$ is referred to as the critical system. The bifurcated solution of the nonlinear system born at $\gamma = 0$ may or may not be locally stable. For simplicity, only double points [11] will be considered in this paper. A useful tool to determine local stability of the bifurcated solution and of the critical system is the projection method developed in [11] and advocated in [2, 11].

Let ℓ and r denote the left row and right column eigenvectors of $L(0)$, corresponding to the critical eigenvalue $\lambda(0) = 0$. Then $\ell r = 1$ by suitable normalization. Denote $\varepsilon = \ell x_e$, where $x_e \neq 0$ satisfying $f(\gamma, x_e) = 0$ is also an equilibrium solution of (1), or bifurcated solution in a small neighborhood of $\gamma = 0$. Then by [11] there exists a series expansion

$$\begin{bmatrix} x_e(\varepsilon) \\ \gamma(\varepsilon) \end{bmatrix} = \sum_{k=1}^{\infty} \begin{bmatrix} x_{ek} \\ \gamma_k \end{bmatrix} \varepsilon^k.$$

Since $f(\gamma, x)$ is sufficiently smooth, there exists a Taylor expansion near the origin of $\mathbf{R}^n$ of the form

$$\dot{x} = f(\gamma, x) = L(\gamma)x + Q(\gamma)[x, x] + C(\gamma)[x, x, x] + \cdots, \tag{4}$$

where $L(\gamma)x$, $Q(\gamma)[x, x]$, and $C(\gamma)[x, x, x]$ are vector-valued linear, quadratic, and cubic terms of $f(\gamma, x)$, respectively, having symmetric form in each of their entries, and can each be expanded into

$$L(\gamma)x = L_0 x + \gamma L_1 x + \gamma^2 L_2 x + \cdots, \qquad Q(\gamma)[x, x] = Q_0[x, x] + \gamma Q_1(x, x) + \cdots,$$

and $C(\gamma)[x, x, x] = C_0[x, x, x] + \gamma C_1[x, x, x] + \cdots$, where L_0, L_1, and L_2 are $n \times n$ constant matrices.

Let $\tilde{\lambda}(\gamma)$ be the critical eigenvalue of the linearized system matrix at the new (bifurcated) equilibrium close to the origin. Then $\tilde{\lambda}(0) = \lambda(0) = 0$ at $\gamma = 0$. There exists a series expansion [11]

$$\tilde{\lambda}(\varepsilon) = \sum_{i=1}^{\infty} \tilde{\lambda}_i \varepsilon^i = \tilde{\lambda}_1 \varepsilon + \tilde{\lambda}_2 \varepsilon^2 + \cdots.$$

The computation of the first two coefficients of $\tilde{\lambda}$ can proceed as follows [11, 2]:

- Step 1: Calculate $\lambda'(0) = \ell L_1 r$, where λ is a function of γ.
- Step 2: Set $x_{e1} = r$ and calculate $\gamma_1 = -\ell Q_0[r,r]/\lambda'(0)$.
- Step 3: Compute x_{e2} from equations $\ell x_{e2} = 0$ and $L_0 x_{e2} = -Q_0[r,r]$, and γ_2 from
$$\gamma_2 = -\frac{1}{\lambda'(0)}\left(\gamma_1 \ell L_1 x_{e2} + \gamma_1^2 \ell L_2 r + 2\ell Q_0[r,x_{e2}] + \gamma_1 \ell Q_1[r,r] + \ell C_0[r,r,r]\right).$$
- Step 4: Set $\tilde{\lambda}_1 = -\gamma_1\lambda'(0)$ and $\tilde{\lambda}_2 = -2\gamma_2\lambda'(0)$.

Local stability of a stationary bifurcation is given by the following lemma [3].

LEMMA 2.2. *Suppose that all eigenvalues of L_0 are stable except one critical eigenvalue. For the case $\gamma_1 \neq 0$, the branch of the bifurcated equilibrium solution is locally stable for γ sufficiently close to* 0 *if $\ell Q_0[r,r]\varepsilon < 0$ and unstable if $\ell Q_0[r,r]\varepsilon > 0$. For the case $\gamma_1 = 0$, the bifurcated solution is locally stable for γ sufficiently close to* 0 *if $\tilde{\lambda}_2 < 0$, and unstable if $\tilde{\lambda}_2 > 0$, where*
$$\tilde{\lambda}_2 = 2\ell\left(2Q_0[r,x_{e2}] + C_0[r,r,r]\right), \qquad x_{e2} = -(\ell^T\ell + L_0^T L_0)^{-1} L_0^T Q_0[r,r].$$

It should be clear that the local bifurcation for the case $\gamma_1 \neq 0$ is transcritical [21]. Thus the branch of the bifurcated solution at $\varepsilon > 0$ has the opposite stability property as the one at $\varepsilon < 0$. We are interested in the stabilization of the transcritical bifurcation for the branch of $\varepsilon > 0$, with no loss of generality. On the other hand, the local bifurcation for the case $\gamma_1 = 0$ and $\gamma_2 \neq 0$ is pitchfork [21], where both branches of the bifurcated solution share the same stability property.

2.2. Local stability for Hopf bifurcations. For a Hopf bifurcation, it is assumed that $L(\gamma)$ possesses a pair of complex eigenvalues $\lambda(\gamma), \bar{\lambda}(\gamma)$, dependent smoothly on γ, while all other eigenvalues are stable in a neighborhood of $\gamma = 0$. Denote $\lambda(\gamma) = \alpha(\gamma) + j\beta(\gamma)$ with $\alpha(\gamma), \beta(\gamma)$ real, and $j = \sqrt{-1}$ imaginary. It is assumed that

$$(5) \qquad \alpha(0) = 0, \qquad \beta(0) = \omega_c \neq 0, \qquad \alpha'(0) = \frac{d\alpha}{d\gamma}(0) \neq 0.$$

Thus $\lambda(\gamma)$ is a critical eigenvalue and so is its conjugate. It implies that the zero solution changes its stability as γ crosses 0. For instance $\alpha'(0) > 0$ implies that the zero solution is locally stable for $\gamma > 0$ and becomes unstable for $\gamma > 0$. Furthermore the Hopf bifurcation theorem asserts the existence of a one-parameter family $\{p_\varepsilon\}$, where $0 < \varepsilon \leq \varepsilon_0$, of nonconstant periodic solutions of (1) emerging from the zero solution at $\gamma = 0$. This is a nonstationary bifurcation. The positive real number ε is a measure of the amplitude of the periodic solution and ε_0 is sufficiently small. The periodic solutions $p_\varepsilon(t)$ have period near $2\pi/\omega_c$ and occur for parameter values γ given by a smooth function $\gamma(\varepsilon)$. Exactly one of the characteristic exponents of p_ε is near zero and is given by

$$(6) \qquad \tilde{\lambda}(\varepsilon) = \tilde{\lambda}_2\varepsilon^2 + \tilde{\lambda}_4\varepsilon^4 + \cdots = \sum_{i=1}^{\infty} \tilde{\lambda}_{2i}\varepsilon^{2i}.$$

Local stability of a Hopf bifurcation is hinged to the first nonzero coefficient of $\tilde{\lambda}(\varepsilon)$, denoted by $\tilde{\lambda}_{2N}$, $N \geq 1$. Generically $N = 1$.

Let the Taylor series of $f(\gamma, x)$ be of the form in (4) where $L_0 = L(0)$. An algorithm to compute $\tilde{\lambda}_2$ is quoted from [2]. See also [10].

- Step 1: Compute left row eigenvector ℓ and right column eigenvector r of L_0 corresponding to the critical eigenvalue of $\lambda(0) = j\omega_c$, normalized by setting $\ell r = 1$.
- Step 2: Solve column vectors μ and ν from the equations

$$-L_0\mu = \frac{1}{2}Q_0[r, \bar{r}], \qquad (2j\omega_c I - L_0)\nu = \frac{1}{2}Q_0[r, r], \qquad j = \sqrt{-1}.$$

- Step 3: The coefficient $\tilde{\lambda}_2$ is given by

$$\tilde{\lambda}_2 = 2\mathrm{Re}\left\{2\ell Q_0[r, \mu] + \ell Q_0[\bar{r}, \nu] + \frac{3}{4}\ell C_0[r, r, \bar{r}]\right\}.$$

Local stability of a Hopf bifurcation is given by the following lemma [3].

LEMMA 2.3. *Suppose all eigenvalues of L_0 are stable in a neighborhood of $\gamma = 0$ except the critical pair of complex eigenvalues. Then the Hopf bifurcation is stable if $\tilde{\lambda}_2 < 0$ and unstable if $\tilde{\lambda}_2 > 0$.*

3. Output feedback stabilization for stationary bifurcations.

We address the feedback stabilization problem for stationary bifurcations. The nonlinear control system under consideration has the form

$$\dot{x} = f(\gamma, x) + g(x, u), \qquad y = h(x), \qquad x \in \mathbf{R}^n, \tag{7}$$

where $f(\gamma, x)$ is the same as in (4), and $g(\cdot,\cdot)$ and $h(\cdot)$ are also smooth functions satisfying

$$g(x, 0) = 0 \ \forall x, \qquad h(0) = 0. \tag{8}$$

It is assumed that $u \in \mathbf{R}$, and $y \in \mathbf{R}^p$ with $p \geq 1$. Thus the nonlinear system (7) has only one control input, but may have more than one output measurement. Its Taylor series expansion is given by

$$\begin{aligned} \dot{x} = {} & L_0 x + \gamma L_1 x + \gamma^2 L_2 x + B_1 u + u\tilde{L}_1 x + Q_0[x, x] + B_2 u^2 + u\tilde{L}_{12}[x, x] \\ & + \gamma Q_1[x, x] + u^2\tilde{L}_2 x + u^2\tilde{Q}_1[x, x] + C_0[x, x, x] + B_3 u^3 + \cdots, \end{aligned} \tag{9}$$

where $\tilde{L}_1 x$ and $\tilde{Q}_1[x, x]$ are the linear and quadratic terms for the linear control component of $g(x, u)$, $\tilde{L}_2 x$ is the linear term for the quadratic control component and $\tilde{L}_{12}[x, x]$ is the quadratic term for the linear control component of $g(x, u)$, and B_1, B_2, and B_3 are the coefficient vectors of u, u^2, and u^3, respectively. It is assumed that L_0 has only one zero eigenvalue with the rest of the eigenvalues stable, and that the bifurcated solution born at $\gamma = 0$ is not locally stable. The assumption on stability of the nonzero eigenvalues of L_0 has no loss of generality. If some of the nonzero eigenvalues of L_0 are unstable, then any linear control method, such as pole placement [12], can be employed to stabilize those unstable modes corresponding to nonzero eigenvalues. It is the unstable mode corresponding to the critical eigenvalue $\lambda(0) = 0$ that renders linear control methods inadequate because of bifurcations.

We seek a smooth local output feedback control law of the form

$$u = K(y) = K_L y + K_Q[y, y] + K_C[y, y, y] + \cdots, \tag{10}$$

that stabilizes the bifurcated system (i.e., the closed-loop system admits bifurcation stability; see Definition 2.1), where $K_Q[\cdot,\cdot]$ and $K_C[\cdot,\cdot,\cdot]$ are the quadratic and cubic terms of $K(y)$, respectively, satisfying $K_Q[0,0] = K_C[0,0,0] = 0$. The output has a Taylor series expansion

$$y = h(x) = H_1x + H_2[x,x] + H_3[x,x,x] + \cdots, \tag{11}$$

where $H_1x, H_2[x,x]$, and $H_3[x,x,x]$ are the linear, quadratic, and cubic terms of $h(x)$, respectively, satisfying $H_2[0,0] = H_3[0,0,0] = 0$. Without loss of generality, it is assumed that linear, quadratic, and cubic terms of the feedback control law in (10) are of the form

$$\begin{aligned} K_Ly = K_1h(x), \qquad K_Q[y,y] &= K_2\tilde{H}_2[x,x] + K_2\tilde{H}_3[x,x,x] + \cdots, \\ K_C[y,y,y] &= K_3\tilde{\tilde{H}}_3[x,x,x] + \cdots, \end{aligned} \tag{12}$$

with K_1, K_2, and K_3 some constant matrices, and $\tilde{H}_2[0,0] = \tilde{H}_3[0,0,0] = \tilde{\tilde{H}}_3[0,0,0] = 0 \in \mathbf{R}^p$. Substituting the control law into (9), we get the closed-loop system equation in series form:

$$\dot{x} = L_0^*x + \gamma L_1^*x + Q_0^*[x,x] + \gamma^2 L_2^*x + \gamma Q_1^*[x,x] + C_0^*[x,x,x] + \cdots, \tag{13}$$

where the linear, quadratic, and cubic terms are given by

$$\begin{aligned} L_0^* = L_0 + B_1K_1H_1, \qquad L_1^* = L_1, \qquad L_2^* &= L_2, \qquad Q_1^*[x,x] = Q_1[x,x], \\ Q_0^*[x,x] &= Q_0[x,x] + B_1\left(K_1H_2[x,x] + K_2\tilde{H}_2[x,x]\right) + K_1H_1x\tilde{L}_1x + B_2(K_1H_1x)^2, \\ C_0^*[x,x,x] &= C_0[x,x,x] + \left(K_1H_2[x,x] + K_2\tilde{H}_2[x,x]\right)\tilde{L}_1x \\ &\quad + 2B_2K_1H_1x\left(K_1H_2[x,x] + K_2\tilde{H}_2[x,x]\right) \\ &\quad + (K_1H_1x)^2\tilde{L}_2x + B_1\left(K_1H_3[x,x,x] + K_2\tilde{H}_3[x,x,x] + K_3\tilde{\tilde{H}}_3[x,x,x]\right) \\ &\quad + K_1H_1x\tilde{L}_{12}[x,x] + K_1H_1x\tilde{Q}_1[x,x] + B_3(K_1H_1x)^3. \end{aligned}$$

Abed and Fu studied the same problem in [3] for the case of state feedback where the critical mode of the linearized system at $\gamma = 0$ is controllable. We will consider the case of output feedback where the critical mode of L_0 is linearly uncontrollable. It should be clear that, in practice, measurement of all state variables is unrealistic. Often only part of the state variables or a nonlinear function of part of the state variables is measurable. Under this circumstance, the critical mode of the linearized system may or may not be observable based on linearized output measurements. Hence the problem considered in this paper has more engineering significance than that of [3]. Because $y \in \mathbf{R}^p$ with $p \geq 1$, our results are also applicable to state feedback by taking $H_1 = I$ and $H_2 = H_3 = \cdots = 0$, which yields $y = x$. We will establish stabilizability conditions for bifurcated systems where the bifurcated solution is unstable near $\gamma = 0$ in the following two subsections.

3.1. Unobservable critical mode. Consider first when the critical mode of the linearized system corresponding to the zero eigenvalue at $\gamma = 0$ is not observable through the linearized output measurement H_1x. Note that by assumption the eigenvalue $\lambda(0) = 0$ is invariant under feedback control because of both uncontrollability and unobservability of the critical mode. Thus L_0^* also possesses the critical zero eigenvalue as L_0. Denote ℓ^* and r^* as the left row and right column eigenvectors for

L_0^* corresponding to the critical eigenvalue. Then it is easy to see that $\ell^* = \ell$ and $r^* = r$ again due to the uncontrollability and unobservability of the critical eigenvalue. This can be seen from the Popov, Belevitch, and Hautus (PBH) test [12]. Denote $\tilde{\lambda}^*$ as the critical eigenvalue of the linearized feedback system at the bifurcated solution. It is a function of $\varepsilon = \ell^* x_e = \ell x_e$ in the form of

$$\tilde{\lambda}^*(\varepsilon) = \tilde{\lambda}_1^*\varepsilon + \tilde{\lambda}_2^*\varepsilon^2 + \cdots . \tag{14}$$

Clearly the local feedback controller in (10) does not change the zero solution. Denote $L^*(\gamma)$ as the linearized system matrix for the closed-loop system at the zero solution. Its critical eigenvalue is denoted by $\lambda^*(\gamma)$. Since $L_1^* = L_1$,

$$\frac{d\lambda^*}{d\gamma}(0) = \lambda'(0) = \frac{d\lambda}{d\gamma}(0) \neq 0$$

is also invariant under feedback control. It follows that the bifurcated solution $x_e \equiv 0$ of the closed-loop system changes its stability and bifurcates at $\gamma = 0$ as well. The problem is whether or not the bifurcated solution can be stabilized with output feedback control. The next result is negative for the transcritical bifurcation.

THEOREM 3.1. *Consider the nonlinear control system* (9) *with output feedback control law* (10). *Suppose that the critical mode of* L_0 *is not linearly observable. Then for the case* $\tilde{\lambda}_1 > 0$, *there does not exist a feedback control law* $u = K(y)$ *that stabilizes the branch of the bifurcated solution at* $\varepsilon > 0$, *and* $\tilde{\lambda}_1^* = \tilde{\lambda}_1$ *is invariant under output feedback control in* (10).

Proof. Note that with the smooth feedback controller in (10), the closed-loop system is in the form of (13). The uncontrollability and unobservability of the critical mode for the linearized system at $\gamma = 0$ imply that both $\ell B_1 = 0$ and $H_1 r = 0$ by the PBH test [12]. Hence applying Lemma 2.2 to the nonlinear system in (13) gives the first coefficient of the critical eigenvalue for the branch of the bifurcated solution at $\varepsilon > 0$:

$$\begin{aligned}\tilde{\lambda}_1^* = \ell Q_0^*[r,r] &= \ell Q_0[r,r] + \ell B_1\left(K_1 H_2[r,r] + K_2\tilde{H}_2[r,r]\right)\\ &+ \ell K_1 H_1 r\tilde{L}_1 r + \ell B_2(K_1 H_1 r)^2 = \ell Q_0[r,r] = \tilde{\lambda}_1.\end{aligned}$$

It follows that the sign of $\tilde{\lambda}_1^*$ is the same as that of $\tilde{\lambda}_1$, which means we cannot alter the sign by feedback controller in (10). □

Although the stability property of a transcritical bifurcation with an unobservable critical mode cannot be altered by output feedback, the situation for pitchfork bifurcation is quite different. We have the following result.

THEOREM 3.2. *Consider the nonlinear control system* (9) *with output feedback control law* (10) *under the same hypothesis as in Theorem* 3.1. *For simplicity, assume in addition that* $\tilde{H}_2[r,r] = 0$. *Then for the case* $\tilde{\lambda}_1 = 0$, *there exists a smooth feedback control law* $u = K(y)$ *as in* (10) *that ensures* $\tilde{\lambda}_2^* < 0$, *i.e., stabilizes the bifurcated solutions, if and only if there exists a linear feedback control law* $u = K_1 y$ *that stabilizes the bifurcated solution. Moreover there exists a nonsingular matrix* $T \in \mathbf{R}^{n\times n}$ *such that*

$$TL_0T^{-1} = \begin{bmatrix} L_{00} & 0 \\ 0 & 0 \end{bmatrix}, \quad TB_1 = \begin{bmatrix} B_{11} \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} H_1 \\ \ell \end{bmatrix} T^{-1} = \begin{bmatrix} H_{11} & 0 \\ 0 & 1 \end{bmatrix}, \tag{15}$$

where $L_{00} \in \mathbf{R}^{(n-1)\times(n-1)}$, $B_{11} \in \mathbf{R}^{(n-1)\times 1}$, *and* $H_{11} \in \mathbf{R}^{p\times(n-1)}$. *Denote* I_{n-1} *as* *the identity matrix of size* $(n-1)\times(n-1)$, *and*

$$\rho = 4\ell Q_0[r, (\ell^T\ell + L_0^T L_0)^{-1}L_0^T B_1] - 2\ell\tilde{L}_1 r, \qquad \beta = \begin{bmatrix} I_{n-1} & 0 \end{bmatrix} TQ_0[r,r]. \tag{16}$$

Then the existence of the stabilizing feedback control law is equivalent to that

$$\tilde{\lambda}_2^* = \tilde{\lambda}_2 + K_1\left(H_{11}L_{00}^{-1}(B_{11}\tilde{\lambda}_2 + \rho\beta) - \rho H_2[r,r]\right) < 0 \tag{17}$$

and that $L_{00} + B_{11}K_1H_{11}$ *is stable for some* $K_1 \neq 0$.

Proof. Let x_{e2} and x_{e2}^* be the unique solutions for

$$\begin{aligned} &\ell x_{e2} = 0, \qquad L_0 x_{e2} = -Q_0[r,r], \qquad \ell x_{e2}^* = 0, \\ &(L_0 + B_1K_1H_1)x_{e2}^* = -Q_0[r,r] - B_1K_1H_2[r,r]. \end{aligned} \tag{18}$$

By the proof of Theorem 3.1, and the condition $\ell B_1 = 0$, $H_1 r = 0$,

$$\begin{aligned} 4\ell Q_0^*[r, x_{e2}^*] &= 2\ell\left(2Q_0[r, x_{e2}^*] + K_1H_1x_{e2}^*\tilde{L}_1 r\right), \\ 2\ell C_0^*[r,r,r] &= 2\ell\left(C_0[r,r,r] + K_1H_2[r,r]\tilde{L}_1 r + K_2\tilde{H}_2[r,r]\tilde{L}_1 r\right). \end{aligned}$$

The formula governing the vector x_{e2}^* as in (18) can be readily verified by noting that $Q_0^*[r,r] = Q_0[r,r] + B_1K_1H_2[r,r]$ due to $\tilde{H}_2[r,r] = 0$, and $H_1 r = 0$. Applying Lemma 2.2 again yields

$$\begin{aligned} \tilde{\lambda}_2^* &= 2\ell\Big(2Q_0[r, x_{e2}^*] + K_1H_1x_{e2}^*\tilde{L}_1 r\Big) + 2\ell\Big(C_0[r,r,r] + K_1H_2[r,r]\tilde{L}_1 r + K_2\tilde{H}_2[r,r]\tilde{L}_1 r\Big) \\ &= \tilde{\lambda}_2 + 4\ell\left(Q_0[r, x_{e2}^*] - Q_0[r, x_{e2}]\right) + 2\ell\left(K_1H_1x_{e2}^* + K_1H_2[r,r] + K_2\tilde{H}_2[r,r]\right)\tilde{L}_1 r \\ &= \tilde{\lambda}_2 - 4\ell Q_0[r, x_{e2} - x_{e2}^*] + 2(\ell\tilde{L}_1 r)K_1\left(H_1x_{e2}^* + H_2[r,r]\right). \end{aligned}$$

It follows that the bifurcated solution is stabilized if and only if there exists $K_1 \neq 0$ such that $\tilde{\lambda}_2^* < 0$ holds and the nonzero eigenvalues of $L_0 + B_1K_1H_1$ lie in the open left half-plane. It remains to be shown that stabilizability of the bifurcated solution is equivalent to the existence of $K_1 \neq 0$ such that (17) holds, plus stability of $L_{00} + B_{11}K_1H_{11}$. From (18), the equations

$$\ell(x_{e2} - x_{e2}^*) = 0, \qquad L_0(x_{e2} - x_{e2}^*) = B_1K_1\left(H_1x_{e2}^* + H_2[r,r]\right)$$

are obtained. Thus the solutions x_{e2} and x_{e2}^* solved from (18) satisfy

$$x_{e2} - x_{e2}^* = (\ell^T\ell + L_0^TL_0)^{-1}L_0^TB_1K_1\left(H_1x_{e2}^* + H_2[r,r]\right).$$

The expression for $\tilde{\lambda}_2^*$ can now be written as

$$\begin{aligned} \tilde{\lambda}_2^* &= \tilde{\lambda}_2 - 4\ell Q_0[r, \bar{B}_1K_1(H_1x_{e2}^* + H_2[r,r])] + 2(\ell\tilde{L}_1 r)K_1\left(H_1x_{e2}^* + H_2[r,r]\right) \\ &= \tilde{\lambda}_2 - \rho K_1\left(H_1x_{e2}^* + H_2[r,r]\right), \qquad \bar{B}_1 = (\ell^T\ell + L_0^TL_0)^{-1}L_0^TB_1, \end{aligned} \tag{19}$$

where ρ is the same as in (16). Since the critical mode of the linearized system is uncontrollable and unobservable, there exists a nonsingular matrix $T \in \mathbf{R}^{n\times n}$, such that (15) is true by Kalman decomposition. Denoting $L_0^* = L_0 + B_1K_1H_1$, the last two equations of (18) yield

$$\begin{aligned} TL_0^*x_{e2}^* &= \begin{bmatrix} L_{00} + B_{11}K_1H_{11} & 0 \\ 0 & 0 \end{bmatrix} Tx_{e2}^* = -T\left(Q_0[r,r] + B_1K_1H_2[r,r]\right) \\ &= -TQ_0[r,r] - \begin{bmatrix} B_{11} \\ 0 \end{bmatrix} K_1H_2[r,r], \qquad \ell x_{e2}^* = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} Tx_{e2}^* = 0, \end{aligned}$$

which implies that the last element of the column vector Tx_{e2}^* is zero. Denote

$$\alpha = \left[\begin{array}{cc} I_{n-1} & 0 \end{array}\right] Tx_{e2}^*, \qquad \beta = \left[\begin{array}{cc} I_{n-1} & 0 \end{array}\right] TQ_0[r,r],$$

where I_{n-1} is an identity matrix of size $(n-1)\times(n-1)$. Then x_{e2}^* is determined by α and T as

$$\text{(20)} \quad \alpha = -\left(L_{00} + B_{11}K_1H_{11}\right)^{-1}\left(\beta + B_{11}K_1H_2[r,r]\right), \qquad x_{e2}^* = T^{-1}\left[\begin{array}{c} \alpha \\ 0 \end{array}\right].$$

Applying (15) and (20) to the expression in (19) gives

$$\begin{aligned}
\text{(21)} \quad \tilde{\lambda}_2^* &= \tilde{\lambda}_2 - \rho K_1\left([H_1T^{-1}][Tx_{e2}^*] + H_2[r,r]\right) = \tilde{\lambda}_2 - \rho K_1\left(H_{11}\alpha + H_2[r,r]\right) \\
&= \tilde{\lambda}_2 + \rho K_1\Big(H_{11}\left(L_{00} + B_{11}K_1H_{11}\right)^{-1}\left(\beta + B_{11}K_1H_2[r,r]\right) - H_2[r,r]\Big) \\
&= \tilde{\lambda}_2 + \rho K_1H_{11}\left(L_{00} + B_{11}K_1H_{11}\right)^{-1}\beta \\
&\quad -\rho\Big(1 - K_1H_{11}\left(L_{00} + B_{11}K_1H_{11}\right)^{-1}B_{11}\Big)K_1H_2[r,r].
\end{aligned}$$

Denote $\bar{\lambda}_2 = \tilde{\lambda}_2 + \rho K_1H_{11}\left(L_{00} + B_{11}K_1H_{11}\right)^{-1}\beta$. Using the identity $\det(I+AB) = \det(I+BA)$ whenever the products AB and BA are compatible square matrices and $\det(A^{-1}) = 1/\det(A)$, the following sequence of equalities is true:

$$\begin{aligned}
\frac{\bar{\lambda}_2}{\tilde{\lambda}_2} &= \det\left(I + \frac{\rho}{\tilde{\lambda}_2}\left(L_{00} + B_{11}K_1H_{11}\right)^{-1}\beta K_1H_{11}\right) \\
&= \frac{1}{\det(L_{00} + B_{11}K_1H_{11})}\det\left(L_{00} + B_{11}K_1H_{11} + \frac{\rho}{\tilde{\lambda}_2}\beta K_1H_{11}\right) \\
&= \frac{\det(L_{00})}{\det(L_{00} + B_{11}K_1H_{11})}\det\left(I + \left(B_{11} + \frac{\rho}{\tilde{\lambda}_2}\beta\right)K_1H_{11}L_{00}^{-1}\right) \\
&= \frac{\det(L_{00})}{\det(L_{00} + B_{11}K_1H_{11})}\left(1 + K_1H_{11}L_{00}^{-1}\left(B_{11} + \frac{\rho}{\tilde{\lambda}_2}\beta\right)\right).
\end{aligned}$$

Similarly there holds

$$\begin{aligned}
1 - K_1H_{11}\left(L_{00} + B_{11}K_1H_{11}\right)^{-1}B_{11} &= \det\left(I - (L_{00} + B_{11}K_1H_{11})^{-1}B_{11}K_1H_{11}\right) \\
&= \frac{\det(L_{00})}{\det(L_{00} + B_{11}K_1H_{11})}.
\end{aligned}$$

Thus the expression of $\tilde{\lambda}_2^*$ in (21) can now be written as

$$\tilde{\lambda}_2^* = \frac{\det(L_{00})}{\det(L_{00} + B_{11}K_1H_{11})}\left(\tilde{\lambda}_2 + K_1\left(H_{11}L_{00}^{-1}(B_{11}\tilde{\lambda}_2 + \rho\beta) - \rho H_2[r,r]\right)\right).$$

Hence stabilization of a pitchfork bifurcation is equivalent to the existence of $K_1 \neq 0$ such that $L_{00} + B_{11}K_1H_{11}$ is stable, and

$$\tilde{\lambda}_2 + K_1\left(H_{11}L_{00}^{-1}(B_{11}\tilde{\lambda}_2 + \rho\beta) - \rho H_2[r,r]\right) < 0,$$

because stability of L_{00} and $L_{00} + B_{11}K_1H_{11}$ ensures that $\det(L_{00})$ and $\det(L_{00} + B_{11}K_1H_{11})$ have the same sign. This concludes the proof. □

Theorem 3.2 is surprising as it indicates that even though the critical mode of the linearized system is neither controllable nor observable, there exists an output feedback control law that stabilizes a pitchfork bifurcation under some mild conditions. This is in contrast to linear control theory. Moreover the nonlinear feedback control law does not offer any advantage over the linear ones as far as stabilization of stationary bifurcation is concerned. This is due to the fact that linear feedback control, though it has no influence on the stability of the linear term, changes the quadratic terms of the state-space equation that in turn determine the stability of the pitchfork bifurcation. It is worth pointing out that condition (17) in Theorem 3.2 is characterized in explicit form. It greatly simplifies the synthesis part for bifurcation stabilization. Indeed, all $K_1 \neq 0$ satisfying (17) can be easily parameterized and then substituted into $L_{00}^* = L_{00} + B_{11}K_1H_{11}$ to determine the right K_1 that ensures stability of L_{00}^*. To be specific for $p = 1$, K_1 satisfying (17) is a semi-infinite interval of the real line. Thus the stabilizing value of K_1 can be easily determined through root locus of

$$1 + K_1H_{11}(sI - L_{00})^{-1}B_{11} = 0, \tag{22}$$

where K_1 changes in the semi-infinite interval determined by inequality (17). If $p > 1$, then all stabilizing K_1's lie on one side of a hyper-plane in $\mathbf{R}^p$ that does not pass through the origin, as $K_1 = 0$ is not stabilizing. In this case one needs to search for the right K_1 on the given side of the hyperplane in $\mathbf{R}^p$ to ensure the stability of $L_{00}^* = L_{00} + B_{11}K_1H_{11}$, for which the parameterized root locus method can be used for (22). The assumption $\tilde{H}_2[r, r] = 0$ in Theorem 3.2 is not very restrictive if the critical mode of the linearized system is not observable through linearized output measurement. For instance, it holds for the case where the output measurement consists of linear combination of state variables. In the case $\tilde{H}_2[r, r] \neq 0$, the quadratic gain K_2 plays a role as well for stabilization of pitchfork bifurcation that will be illuminated further in the next subsection.

3.2. Observable critical mode.

This subsection is concerned with the case where the critical mode is linearly observable based on output measurement. Consider first the transcritical bifurcation. Without loss of generality, the branch of $\varepsilon > 0$ is assumed to be unstable for $\gamma > 0$. This is equivalent to $\tilde{\lambda}_1 > 0$. Our goal is to seek a controller of the form (10) that stabilizes the bifurcated solution for $\varepsilon > 0$ without changing the stability property of the zero solution. This problem is very different from past work [3] but has its significance in engineering application (see section 5). It is noted that by assumption the eigenvalue $\lambda(0) = 0$ is invariant under feedback control. Thus L_0^* also possesses the critical zero eigenvalue as L_0 does. Denote ℓ^* and r^* as the left row and right column eigenvectors for L_0^* corresponding to the critical eigenvalue. Then $\ell^* = \ell$ due to the uncontrollability of the critical eigenvalue. Denote $\tilde{\lambda}^*$ as the critical eigenvalue of L_0^* under feedback. It has the same form of the series expansion as in (14). However, $r^* \neq r$ in general due to $H_1r \neq 0$ by the observability of the critical mode. The next result concerns stabilization of transcritical bifurcations.

THEOREM 3.3. *Consider the nonlinear control system* (9) *with output feedback control law* (10)*. Suppose that the critical mode of the linearized system corresponding to the zero eigenvalue at $\gamma = 0$ is observable. Then for the case $\tilde{\lambda}_1 > 0$, i.e., $\ell Q_0[r, r] > 0$, there exists a nonlinear feedback control law $u = K(y)$ that stabilizes the given branch of the bifurcated solution at $\varepsilon > 0$ if and only if there exists a linear feedback control law that stabilizes the given branch of the bifurcated solution at $\varepsilon > 0$. Moreover* *there exists a nonsingular matrix $T \in \mathbf{R}^{n\times n}$ such that*

$$(23)\ TL_0T^{-1} = \begin{bmatrix} L_{00} & 0 \\ 0 & 0 \end{bmatrix}, \quad TB_1 = \begin{bmatrix} B_{11} \\ 0 \end{bmatrix}, \quad \begin{bmatrix} H_1 \\ \ell \end{bmatrix} T^{-1} = \begin{bmatrix} H_{11} & H_{12} \\ 0 & 1 \end{bmatrix},$$

where L_{00}, B_{11}, and H_{11} have the same sizes as in Theorem 3.2, respectively. Let ℓ_i be the ith element of ℓ, and $r^TQ_{0k}r$ be the kth element of $Q_0[r,r]$ with $Q_{0k} = Q_{0k}^T$ for $k = 1, 2, \ldots, n$. Partition

$$(T^{-1})^T\left(\sum_{k=1}^{n} \ell_k Q_{0k}\right)T^{-1} = \begin{bmatrix} \tilde{Q}_{00} & \tilde{Q}_{01} \\ \tilde{Q}_{10} & \tilde{Q}_{11} \end{bmatrix}, \qquad \tilde{Q}_{10} = \tilde{Q}_{01}^T, \qquad \tilde{Q}_{00} \in \mathbf{R}^{(n-1)\times(n-1)}.$$

Then $\tilde{Q}_{11} = \tilde{\lambda}_1 > 0$. The existence of a stabilizing feedback control law, subject to the same stability property for the zero solution, is equivalent to the existence of $K_1 \neq 0$ such that

(i) $\left(1 + K_1H_{11}L_{00}^{-1}B_{11}\right)\left(\tilde{\lambda}_1(1 + K_1H_{11}L_{00}^{-1}B_{11}) + aK_1H_{12}\right) + b(K_1H_{12})^2 < 0,$

(ii) $\lambda'(0)\left(\lambda'(0) + K_1(H_{11}\lambda'(0) - H_{12}d)L_{00}^{-1}B_{11}\right) > 0,$ *and*

(iii) $L_{00}^* = L_{00} + B_{11}K_1H_{11}$ *is stable,*

where $a = \tilde{d}_0 - 2\tilde{Q}_{10}L_{00}^{-1}B_{11}$, $b = \ell B_2 + ((L_{00}^{-1}B_{11})^T\tilde{Q}_{00} - \tilde{d})L_{00}^{-1}B_{11}$, $\begin{bmatrix} \tilde{d} & \tilde{d}_0 \end{bmatrix} = \ell\tilde{L}_1T^{-1}$with $\tilde{d}_0$ scalar, $d^T = \begin{bmatrix} I_{n-1} & 0 \end{bmatrix}(\ell L_1T^{-1})^T$, and $\lambda(\gamma)$ is the critical eigenvalue in (3).

Proof. From the proofs of Theorem 3.1 and of Lemma 2.2, the first nonzero coefficient of the critical eigenvalue for the feedback system has the expression

$$(24)\qquad \begin{aligned} \tilde{\lambda}_1^* &= \ell Q_0[r^*, r^*] + \ell B_1\left(K_1H_2[r^*, r^*] + K_2\tilde{H}_2[r^*, r^*]\right) \\ &\quad + \ell K_1H_1r^*\tilde{L}_1r^* + \ell B_2(K_1H_1r^*)^2 \\ &= \ell Q_0[r^*, r^*] + (\ell\tilde{L}_1r^*)(K_1H_1r^*) + \ell B_2(K_1H_1r^*)^2, \end{aligned}$$

due to $\ell B_1 = 0$ by the uncontrollability of the critical mode and the condition of the theorem. Because only the linear gain K_1 is present in the expression of $\tilde{\lambda}_1^*$, the existence of a stabilizing control law is equivalent to the existence of a linear stabilizing control law. It remains to be shown that conditions (i)–(iii) are equivalent to the existence of the stabilizing feedback control law with the same stability property for the zero solution retained. By Kalman decomposition a nonsingular matrix T exists such that (23) holds where the lower triangular Schur form of TL_0T^{-1} is used. The hypothesis on L_0 implies that $\det(L_{00}) \neq 0$, and thus $0 = \ell L_0 = (\ell T^{-1})(TL_0T^{-1})$ yields the form of ℓT^{-1} as in (23). Since

$$\ell r^* = (\ell T^{-1})(Tr^*) = 1, \quad L_0^*r^* = (L_0 + B_1K_1H_1)r^* = T(L_0 + B_1K_1H_1)T^{-1}(Tr^*) = 0,$$

the right eigenvector of L_0^* corresponding to the zero eigenvalue is found to be

$$(25)\quad r^* = T^{-1}\begin{bmatrix} \eta \\ 1 \end{bmatrix},$$

$$\eta = -(I_{n-1} + L_{00}^{-1}B_{11}K_1H_{11})^{-1}L_{00}^{-1}B_{11}K_1H_{12} = -\frac{L_{00}^{-1}B_{11}K_1H_{12}}{1 + K_1H_{11}L_{00}^{-1}B_{11}}.$$

By noticing that

$$K_1H_{11}\eta = -\frac{(K_1H_{11}L_{00}^{-1}B_{11})K_1H_{12}}{1+K_1H_{11}L_{00}^{-1}B_{11}} = -K_1H_{12} + \frac{K_1H_{12}}{1+K_1H_{11}L_{00}^{-1}B_{11}},$$

the expression for $K_1H_1r^*$ can be simplified as

$$\begin{aligned} K_1H_1r^* &= K_1(H_1T^{-1})(Tr^*) = K_1\begin{bmatrix} H_{11} & H_{12}\end{bmatrix}\begin{bmatrix}\eta \\ 1\end{bmatrix} \\ &= K_1H_{11}\eta + K_1H_{12} = \frac{K_1H_{12}}{1+K_1H_{11}L_{00}^{-1}B_{11}}. \end{aligned} \tag{26}$$

It is claimed that the condition (ii) is equivalent to the invariance of the stability property for the zero solution, provided that $L_{00}^* = L_{00} + B_{11}K_1H_{11}$ is stable. Indeed, denote λ^* as the critical eigenvalue of the linearized system under feedback control law $u = K(y)$. Then

$$\begin{aligned} \frac{d\lambda^*}{d\gamma}(0) &= \ell L_1r^* = (\ell L_1T^{-1})(Tr^*) = \begin{bmatrix} d & \lambda'(0)\end{bmatrix}\begin{bmatrix}\eta \\ 1\end{bmatrix} \\ &= d\eta + \lambda'(0), \qquad d = \ell L_1T^{-1}\begin{bmatrix} I_{n-1} \\ 0\end{bmatrix}, \end{aligned}$$

by the fact that if $K_1 = 0$, then $\eta = 0$, and thus $\lambda^*(\gamma)$ reduces to $\lambda(\gamma)$. Substituting the expression of η as in (25) yields that

$$\ell L_1r^* = -\frac{dL_{00}^{-1}B_{11}K_1H_{12}}{1+K_1H_{11}L_{00}^{-1}B_{11}} + \lambda'(0) = \frac{\lambda'(0) + K_1(H_{11}\lambda'(0) - H_{12}d)L_{00}^{-1}B_{11}}{1+K_1H_{11}L_{00}^{-1}B_{11}}.$$

It follows that (ii) is equivalent to ℓL_1r^* that has the same sign as $\lambda'(0)$ due to the stability assumption for L_{00}^* and L_{00}, that is in turn equivalent to that the zero solution for the feedback system changes its stability also at $\gamma = 0$, and shares the same stability property as that of an uncontrolled system. Thus the stability property of the zero solution is retained.

Similarly,

$$\ell\tilde{L}_1r^* = \frac{\tilde{d}_0 + K_1(H_{11}\tilde{d}_0 - H_{12}\tilde{d})L_{00}^{-1}B_{11}}{1+K_1H_{11}L_{00}^{-1}B_{11}}, \qquad \begin{bmatrix}\tilde{d} & \tilde{d}_0\end{bmatrix} = \ell\tilde{L}_1T^{-1},$$

with $\tilde{d}_0$ scalar. Combining with the expression in (26), there holds

$$\ell\tilde{L}_1r^*K_1H_1r^* = \left(\frac{\tilde{d}_0 + K_1(H_{11}\tilde{d}_0 - H_{12}\tilde{d})L_{00}^{-1}B_{11}}{(1+K_1H_{11}L_{00}^{-1}B_{11})^2}\right)K_1H_{12}. \tag{27}$$

For the term $\ell Q_0[r^*, r^*]$, there holds

$$\ell Q_0[r^*, r^*] = (Tr^*)^T(T^{-1})^TQ_\ell T^{-1}(Tr^*) = \begin{bmatrix}\eta^T & 1\end{bmatrix}\begin{bmatrix}\tilde{Q}_{00} & \tilde{Q}_{01} \\ \tilde{Q}_{10} & \tilde{Q}_{11}\end{bmatrix}\begin{bmatrix}\eta \\ 1\end{bmatrix}$$

for some real symmetric matrix Q_ℓ. To show that $\tilde{Q}_{11} = \tilde{\lambda}_1$, setting $K_1 = 0$, thus $a = 0$, leads to

$$\tilde{\lambda}_1 = \ell Q_0[r, r] = (Tr)^T(T^{-1})^TQ_\ell T^{-1}(Tr) = Q_{11},$$

because the last element of Tr is one, by $(\ell T^{-1})Tr = 1$. Therefore,

$$\ell Q_0[r^*, r^*] = \tilde{\lambda}_1 - \left(\frac{2\tilde{Q}_{10} L_{00}^{-1} B_{11}}{1 + K_1 H_{11} L_{00}^{-1} B_{11}} \right) K_1 H_{12} + \left(\frac{(L_{00}^{-1} B_{11})^T \tilde{Q}_{00} (L_{00}^{-1} B_{11})}{\left(1 + K_1 H_{11} L_{00}^{-1} B_{11}\right)^2} \right) (K_1 H_{12})^2. \tag{28}$$

Combining (26)–(28) with rearrangement yields the expression

$$\tilde{\lambda}_1^* \left(1 + K_1 H_{11} L_{00}^{-1} B_{11}\right)^2 = \tilde{\lambda}_1 \left(1 + K_1 H_{11} L_{00}^{-1} B_{11}\right)^2 + a \left(1 + K_1 H_{11} L_{00}^{-1} B_{11}\right) K_1 H_{12} + b (K_1 H_{12})^2,$$

which can be simplified to (i). The necessity of conditions (i)–(iii) is now clear. In brief, (iii) implies that $L_{00}^* = L_{00} + B_{11} K_1 H_{11}$ is stable. Hence

$$\left(1 + K_1 H_{11} L_{00}^{-1} B_{11}\right)^2 > 0.$$

Condition (i) then implies that $\tilde{\lambda}_1^* < 0$ ensures stability of the given branch of the bifurcated solution, while condition (ii) guarantees the invariance of the stability property of the zero solution under feedback. □

It is noticed that if $b = 0$, then (i) of Theorem 3.6 can be further simplified to

$$\tilde{\lambda}_1 (1 + K_1 H_{11} L_{00}^{-1} B_{11}) + a K_1 H_{12} < 0,$$

by the fact that condition (iii) and stability of L_{00} imply that

$$1 + K_1 H_{11} L_{00}^{-1} B_{11} = \frac{\det(L_{00} + B_{11} K_1 H_{11})}{\det(L_{00})} > 0.$$

Theorem 3.3 indicates that stabilization of a transcritical bifurcation is possible by using just a linear control law. More importantly the conditions (i)–(iii) also provide explicit formulas for synthesis of a stabilizing linear gain K_1. Indeed, for the case of $p = 1$, K_1 is scalar. The set of K_1 satisfying each of (i)–(iii) can be easily computed that are either finite intervals or semi-infinite intervals. In particular, the set of K_1 satisfying (iii) can be obtained through root locus. For the case $p > 1$, conditions (i)–(iii) offer more freedom for the synthesis of stabilizing linear gain K_1. On the other hand, the synthesis becomes more complex because there is more than one element for K_1.

REMARK 3.4. *Suppose that the linearized system has been transformed into the form of* (23). *Then the last entry of the state vector is the critical state variable. Measurement of this critical state variable corresponds to* $H_{12} \neq 0$. *Theorem* 3.6 *indicates that the feedback of the critical state variable is crucial. This is due to the fact that if* $H_{12} = 0$, *the sign of* $\tilde{\lambda}_1^*$ *remains the same as that of* $\tilde{\lambda}_1$, *and thus the given branch of the bifurcated solution at* $\varepsilon > 0$ *is not stabilizable that is exactly the result of Theorem* 3.1. *However, this fact does not imply that the measurement of noncritical state variables, which corresponds to* $H_{11} \neq 0$, *is unnecessary. By condition* (i) *of Theorem* 3.3, *the measurement of noncritical state variables becomes necessary if* $a = b = 0$. *If* $a \neq 0$ *and/or* $b < 0$, *however, the measurement of noncritical state variables are redundant and thus unnecessary for the purpose of bifurcation stabilization.*

Theorem 3.3 has implications for state feedback control:

$$u = K(x) = K_1 x + K_Q[x, x] + K_C[x, x, x] + \cdots,$$

with K_1 the linear state feedback gain and $K_Q[\cdot, \cdot]$ and $K_C[\cdot, \cdot, \cdot]$ the quadratic and cubic terms, respectively. The next result is a direct consequence of Theorem 3.3.

Corollary 3.5. *Suppose that all the hypotheses in Theorem* 3.3 *hold. Then there exists a nonlinear state feedback control law $u = K(x)$ that stabilizes the given branch of the bifurcated solution if and only if there exists a linear state feedback control law that stabilizes the given branch of the bifurcated solution. Moreover, with the same notation as in Theorem* 3.3, *the existence of stabilizing state feedback control law (subject to the same stability property for the zero solution) is equivalent to the existence of a* $K_1 = \left[\begin{array}{cc} K_{11} & K_{12} \end{array}\right] T \neq 0$ *such that*

$$\text{(i) } \left(1 + K_{11} L_{00}^{-1} B_{11}\right) \left(\tilde{\lambda}_1 (1 + K_{11} L_{00}^{-1} B_{11}) + a K_{12}\right) + b K_{12}^2 < 0,$$

$$\text{(ii) } \lambda'(0) \left(\lambda'(0) + (K_{11}\lambda'(0) - K_{12} d) L_{00}^{-1} B_{11}\right) > 0, \quad \textit{and}$$

$$\text{(iii) } L_{00}^* = L_{00} + B_{11} K_{11} \quad \textit{is stable.}$$

Proof. The theorem can be easily proven by setting $K_1 H_{11} \to K_{11}$, $K_1 H_{12} \to K_{12}$, and noting that

$$K_1 = \left[\begin{array}{cc} K_{11} & K_{12} \end{array}\right] T$$

with T the similarity transform, in the proof of Theorem 3.3. □

For a pitchfork bifurcation, i.e., $\tilde{\lambda}_1 = 0$, the situation is again different. We adopt an approach as in [3] by setting the linear term of the controller to zero. In fact, by the proof of Theorems 3.1 and 3.3, the nonzero gain K_1 will result in $\tilde{\lambda}_1^* \neq 0$, thereby changing the pitchfork bifurcation into a transcritical bifurcation for which only one branch of the bifurcated solution can be stable. Hence this is not a desirable situation other than some exceptional situations. It is noted that with $K_1 = 0$, the eigenvectors of L_0^* corresponding to the critical eigenvalue at $\gamma = 0$ satisfy $\ell^* = \ell$ and $r^* = r$. Hence both row and column eigenvectors of the critical eigenvalue are invariant under feedback control. Since $L_1^* = L_1$, there holds $\ell^* L_1^* r^* = \ell L_1 r = \lambda'(0)$. Thus the zero solution of the feedback system changes its stability at $\gamma = 0$ as well. The stabilizability of the bifurcated solution is given by the following result.

Theorem 3.6. *Consider the nonlinear control system* (9) *with output feedback control law* (10) *under the same hypothesis as in Theorem* 3.3. *Then for the case $\tilde{\lambda}_1 = 0$, there exists a feedback control law $u = K(y)$ subject to $K_1 = 0$, that ensures $\tilde{\lambda}_1^* = 0$ and $\tilde{\lambda}_2^* < 0$, i.e., changes the pitchfork bifurcation from subcritical into supercritical, if and only if*

$$\rho = 4\ell Q_0[r, (\ell^T \ell + L_0^T L_0)^{-1} L_0^T B_1] - 2\ell \tilde{L}_1 r \neq 0 \quad \textit{and} \quad \tilde{H}_2[r, r] \neq 0.$$

Note that the expression of ρ is the same as in (16).

Proof. By the proof of Theorem 3.2 and $\ell B_1 = 0$, there holds

$$\begin{aligned}
\tilde{\lambda}_2^* = 2\ell &\Big(2Q_0[r, x_{e2}^*] + C_0[r, r, r] + K_2 \tilde{H}_2[r, r] \tilde{L}_1 r \\
&+ \; K_1 H_1 r (\tilde{Q}_1[r, r] + K_1 H_1 r \tilde{L}_2 r + B_3 (K_1 H_1 r)^2)\Big) \\
&+ 2\ell K_1 \left((H_2[r, r] + H_1 x_{e2}^*) \tilde{L}_1 r + H_1 r \tilde{L}_1 x_{e2}^*\right) \\
&+ 4\ell B_2 K_1 H_1 r \left(K_1 (H_1 x_{e2}^* + H_2[r, r]) + K_2 \tilde{H}_2[r, r]\right).
\end{aligned}$$

Setting $K_1 = 0$ yields $\tilde{\lambda}_1^* = 0$. Since $Q_0^*[r,r] = Q_0[r,r] + B_1K_2\tilde{H}_2[r,r]$, x_{e2}^* can be solved from

$$L_0x_{e2}^* = -Q_0[r,r] - B_1K_2\tilde{H}_2[r,r], \qquad \ell x_{e2}^* = 0.$$

Combined with $\ell x_{e2} = 0$ and $L_0x_{e2} = -Q_0[r,r]$ yields

$$x_{e2}^* - x_{e2} = -(\ell^T\ell + L_0^TL_0)^{-1}L_0^TB_1K_2\tilde{H}_2[r,r].$$

Substituting the above into the expression of $\tilde{\lambda}_2^*$ gives

$$\begin{aligned}\tilde{\lambda}_2^* &= \tilde{\lambda}_2 + 4\ell Q_0[r, x_{e2}^* - x_{e2}] + 2\ell\tilde{L}_1rK_2\tilde{H}_2[r,r]\\ &= \tilde{\lambda}_2 - 4\ell Q_0[r,(\ell^T\ell + L_0^TL_0)^{-1}L_0^TB_1]K_2\tilde{H}_2[r,r] + 2\ell\tilde{L}_1rK_2\tilde{H}_2[r,r]\\ &= \tilde{\lambda}_2 - \left(4\ell Q_0[r,(\ell^T\ell + L_0^TL_0)^{-1}L_0^TB_1] - 2\ell\tilde{L}_1r\right)K_2\tilde{H}_2[r,r]\\ &= \tilde{\lambda}_2 - \rho K_2\tilde{H}_2[r,r].\end{aligned}$$

Since $\rho \neq 0$ and $\tilde{H}_2[r,r] \neq 0$, there exists $K_2 \neq 0$ such that $\tilde{\lambda}_2^* < 0$. Conversely, stability of the bifurcated solution implies that

$$\tilde{\lambda}_2^* = \tilde{\lambda}_2 - \rho K_2\tilde{H}_2[r,r] < 0,$$

which then implies that $\rho \neq 0$ and $\tilde{H}_2[r,r] \neq 0$ by the hypothesis that $\tilde{\lambda}_2 > 0$. □

It is noted that terms of order higher than two do not affect stability. Thus the stabilizing controllers can be taken as quadratic. Moreover if the output measurements consist of linear combinations of state variables, then the observability condition implies that $\tilde{H}_2[r,r] \neq 0$. In this case, stabilizability for pitchfork bifurcation is equivalent to $\rho \neq 0$. For bifurcation stabilization using state feedback in the case of a pitchfork bifurcation, the stabilizability condition is again equivalent to $\rho \neq 0$, because state feedback is a special case of output feedback satisfying observability.

4. Output feedback stabilization for Hopf bifurcation.

It is assumed that the linearized system matrix $L(\gamma)$ as in (2) has a pair of complex (critical) eigenvalues $\lambda(\gamma) = \alpha(\gamma) \pm j\beta(\gamma)$ such that $\alpha(0) = 0$ and $\alpha'(0) \neq 0$, while all other eigenvalues are stable. As explained in the previous section, this assumption has no loss of generality. The problem to be investigated in this section is the stabilization of Hopf bifurcations with output feedback control if the Hopf bifurcation born at $\gamma = 0$ for the nonlinear system in (1) is unstable. We consider first the case when the pair of critical modes corresponding to the pair of complex eigenvalues $\lambda(\gamma)$ are neither controllable nor observable. According to the PBH test [12], both left and right eigenvectors corresponding to the pair of critical eigenvalues satisfy

$$\ell B_1 = \bar{\ell}B_1 = H_1r = H_1\bar{r} = 0.$$

It follows that when the feedback controller (10) is employed, $L_0^* = L_0 + B_1K_1H_1$ retains the pair of critical eigenvalues $\pm j\omega_c$ at $\gamma = 0$. Denote ℓ^* and r^* as the left and right eigenvectors of L_0^* corresponding to the pair of critical eigenvalues, respectively. Then there holds $\ell^* = \ell$ and $r^* = r$, and thus a Hopf bifurcation is again born at $\gamma = 0$ for which the zero solution changes its stability as γ crosses zero [11]. The next result gives the condition on stabilizability of a Hopf bifurcation.

THEOREM 4.1. *Consider the nonlinear control system* (9) *with output feedback control law* (10). *Suppose that* $\tilde{\lambda}_2 > 0$ *with* $\tilde{\lambda}(\varepsilon)$ *as in* (6) *and the critical modes of* L_0

are neither controllable nor observable in the sense that $\ell B_1 = 0$ *and* $H_1 r = 0$*. For simplicity, assume in addition that* $\tilde{H}_2[r,r] = H_2[r,r] = 0$*. Let* $Q_0^*[x,x]$ *and* C_0^* *be as in the proof of Theorem* 3.1*. Let* μ *and* ν *be as in section* 2.2*, and define* μ^* *and* ν^* *by*

$$-L_0^*\mu^* = \frac{1}{2}Q_0^*[r,\bar{r}], \qquad (2j\omega_c I - L_0^*)\,\nu^* = \frac{1}{2}Q_0^*[r,r],$$

where $L_0^* = L_0 + B_1 K_1 H_1$*. Then there exists a feedback control law* $u = K(y)$ *that stabilizes the Hopf bifurcation if and only if there exists a linear feedback control law* $u = K_1 y$ *that stabilizes the Hopf bifurcation. Moreover the existence of a stabilizing feedback control law is equivalent to*

$$(29)\quad \tilde{\lambda}_2^* = \tilde{\lambda}_2 + \mathrm{Re}\left\{\frac{K_1H_1L_0^{-1}Q_0[r,\bar{r}]\theta}{1+K_1H_1L_0^{-1}B_1} + \frac{K_1H_1(L_0-2j\omega_c I)^{-1}Q_0[r,r]\phi}{1+K_1H_1(L_0-2j\omega_c I)^{-1}B_1}\right\} < 0,$$

and the noncritical eigenvalues of L_0^* *lie on open left half-plane for some* $K_1 \neq 0$*, where*

$$(30)\quad \theta = -\left(\ell\tilde{L}_1 r - 2\ell Q_0[r, L_0^{-1}B_1]\right), \quad \phi = \frac{1}{2}\left(\ell\tilde{L}_1\bar{r} - 2\ell Q_0[\bar{r}, (-2j\omega_c I + L_0)^{-1}B_1]\right).$$

Proof. Denote $\tilde{\lambda}^*(\varepsilon)$ as the function in (6) for the controlled system, and $\tilde{\lambda}_2^*$ as the first coefficient of $\tilde{\lambda}^*(\varepsilon)$. From the proof of Theorem 3.1,

$$Q_0^*[r,r] = Q_0[r,r], \qquad \ell Q_0^*[r,r] = \ell Q_0[r,r], \qquad \ell C_0^*[r,r,\bar{r}] = \ell C_0[r,r,\bar{r}],$$

by using $\ell B_1 = \bar{\ell}B_1 = 0$, $H_1 r = H_1\bar{r} = 0$, and $\tilde{H}_2[r,r] = H_2[r,r] = 0$. Thus there hold

$$\begin{aligned} -2(L_0 + B_1K_1H_1)\mu^* &= Q_0[r,\bar{r}],\ -2L_0\mu = Q_0[r,\bar{r}], \\ -2(-2j\omega_c I + L_0 + B_1K_1H_1)\nu^* &= Q_0[r,r],\ -2(-2j\omega_c I + L_0)\nu^* = Q_0[r,r]. \end{aligned}$$

The above equalities imply that

$$\mu^* - \mu = -L_0^{-1}B_1K_1H_1\mu^*, \qquad \nu^* - \nu = -(-2j\omega_c I + L_0)^{-1}B_1K_1\nu^*.$$

Although $Q_0^*[r,r] = Q_0[r,r]$, $Q_0^*[r,x] \neq Q_0[r,x]$ for $x \neq r$ due to the feedback term. In fact,

$$2\ell Q_0^*[r,\mu^*] = 2\ell Q_0[r,\mu^*] + \ell K_1H_1\mu^*\tilde{L}_1 r,\ 2\ell Q_0^*[\bar{r},\nu^*] = 2\ell Q_0[\bar{r},\nu^*] + \ell K_1H_1\nu^*\tilde{L}_1\bar{r}.$$

The coefficient $\tilde{\lambda}_2^*$ can now be computed as

$$\begin{aligned} \tilde{\lambda}_2^* &= \mathrm{Re}\left\{2\ell(2Q_0[r,\mu^*] + Q_0[\bar{r},\nu^*]) + 3C_0[r,r,\bar{r}]/4\right\} + K_1H_1\mathrm{Re}\left\{2\mu^*\ell\tilde{L}_1 r + \nu^*\ell\tilde{L}_1\bar{r}\right\} \\ &= \tilde{\lambda}_2 + \mathrm{Re}\left\{2\ell(2Q_0[r,\mu^*-\mu] + Q_0[\bar{r},\nu^*-\nu])\right\} + K_1H_1\mathrm{Re}\left\{2\mu^*\ell\tilde{L}_1 r + \nu^*\ell\tilde{L}_1\bar{r}\right\} \\ &= \tilde{\lambda}_2 + K_1H_1\mathrm{Re}\Big\{2\mu^*\left(\ell\tilde{L}_1 r - 2\ell Q_0[r, L_0^{-1}B_1]\right) \\ &\qquad\qquad + \nu^*\left(\ell\tilde{L}_1\bar{r} - 2\ell Q_0[\bar{r}, (-2j\omega_c I + L_0)^{-1}B_1]\right)\Big\}. \end{aligned}$$

By $\mu^* = -0.5(L_0+B_1K_1H_1)^{-1}Q_0[r,\bar{r}]$, and $\nu^* = -0.5(L_0+B_1K_1H_1-2j\omega_c I)^{-1}Q_0[r,r]$,

$$\tilde{\lambda}_2^* = \tilde{\lambda}_2 + K_1H_1\mathrm{Re}\left\{(L_0 + B_1K_1H_1)^{-1}Q_0[r,\bar{r}]\theta + (L_0 + B_1K_1H_1 - 2j\omega_c I)^{-1}Q_0[r,r]\phi\right\},$$

where θ and ϕ are as in (30). For any square nonsingular matrices M and $M+NF$,

$$F(M+NF)^{-1}G = F(I+M^{-1}NF)^{-1}M^{-1}G = \frac{FM^{-1}G}{1+FM^{-1}N}, \tag{31}$$

whenever $FM^{-1}N$ is a scalar. The expression for $\tilde{\lambda}_2^*$ as in (29) can now be verified. Hence stabilization of the Hopf bifurcation with the nonlinear controllers in (10) is equivalent to the existence of $K_1 \neq 0$ such that $\tilde{\lambda}_2^* < 0$, and the rest of the eigenvalues of L_0^* remain in the open left half plane. □

The result in Theorem 4.1 is similar to the stabilizability of a pitchfork bifurcation in Theorem 3.2 where nonlinear controllers do not offer any advantage over linear ones in terms of bifurcation stabilization. Hence if the pair of critical modes of the linearized system is uncontrollable and unobservable, linear controllers are adequate for bifurcation stabilization. Moreover a necessary condition for stabilizability of a Hopf bifurcation as in Theorem 4.1 is that

$$\|H_1 L_0^{-1} Q_0[r,\bar{r}]\theta\| + \|H_1(L_0 - 2j\omega_c I)^{-1} Q_0[r,r]\phi\| \neq 0,$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbf{R}^p$. If the above holds, all K_1 such that $\tilde{\lambda}_2^* < 0$ can be easily parameterized, especially for scalar K_1 that are a collection of a limited number of finite intervals, or semi-infinite intervals, while stability of the noncritical eigenvalues of L_0 can be tested using root locus of (22) as argued in the previous section.

In the rest of the section, we study the case when the stabilizability condition in Theorem 4.1 is not satisfied. Clearly additional sensors must be deployed such that $H_1 r \neq 0$ is valid in order for a Hopf bifurcation to be stabilizable. Although linear controllers can be investigated, it is much easier to consider the class of nonlinear controllers in (10) with $K_1 = 0$ as discussed in [2]. The next result generalizes the result in [2] to output feedback stabilization.

THEOREM 4.2. *Consider the nonlinear control system* (9) *with output feedback control law in* (10) *subject to* $K_1 = 0$. *Suppose that* $\tilde{\lambda}_2 > 0$ *with* $\tilde{\lambda}(\varepsilon)$ *as* (6), μ, ν *as in section* 2.2, *and the critical mode of* L_0 *is observable through linearized output measurement* $y = H_1 x$. *Then there exists a feedback control law* $u = K(y)$ *that stabilizes the Hopf bifurcation if and only if*

$$\begin{aligned} \operatorname{Re}\Big\{ & \tilde{H}_2[r,\bar{r}]\ell\left(2Q_0[r, L_0^{-1}B_1] - \tilde{L}_1 r\right) \\ & + \tilde{H}_2[r,r]\ell\left(Q_0[\bar{r}, (2j\omega_c I - L_0)^{-1}B_1] - 0.5\tilde{L}_1\bar{r}\right)\Big\} \neq 0. \end{aligned} \tag{32}$$

If the above condition holds, stabilizing controllers can be taken as quadratic.

Proof. With $K_1 = 0$, both critical eigenvalues and left/right eigenvectors are invariant under feedback. Moreover

$$\begin{aligned} L_0^* = L_0, \qquad Q_0^*[x,z] &= Q_0[x,z] + B_1 K_2 \tilde{H}_2[x,z], \qquad \ell Q_0^*[x,z] = \ell Q_0[x,z] \\ 3\ell C_0^*[r,r,\bar{r}] &= 3\ell C_0[r,r,\bar{r}] + \ell K_2\left(\tilde{H}_2[r,r]\tilde{L}_1\bar{r} + 2\tilde{H}_2[r,\bar{r}]\tilde{L}_1 r\right) \end{aligned}$$

by $\ell B_1 = 0$. It follows that

$$\mu^* = -0.5L_0^{-1}\left(Q_0[r,\bar{r}] + B_1 K_2 \tilde{H}_2[r,\bar{r}]\right) = \mu + \Delta\mu, \qquad \Delta\mu = -0.5L_0^{-1}B_1 K_2 \tilde{H}_2[r,\bar{r}],$$

$$\nu^* = \nu + \Delta\nu, \qquad \Delta\nu = -0.5(2j\omega_c I - L_0)^{-1} B_1 K_2 \tilde{H}_2[r,\bar{r}].$$

By the property of the quadratic term,

$$Q_0[r,\mu+\Delta\mu] = Q_0[r,\mu] + Q_0[r,\Delta\mu] = Q_0[r,\mu] - 0.5K_2\tilde{H}_2[r,\bar{r}]Q_0[r,L_0^{-1}B_1],$$
$$Q_0[\bar{r},\nu+\Delta\nu] = Q_0[\bar{r},\nu] + Q_0[\bar{r},\Delta\nu] = Q_0[\bar{r},\nu] - 0.5K_2\tilde{H}_2[r,\bar{r}]Q_0[\bar{r},(2j\omega_c I - L_0)^{-1}B_1].$$

After lengthy calculations, we finally obtain

$$\begin{aligned}\tilde{\lambda}_2^* = \tilde{\lambda}_2 - K_2\mathrm{Re}\Big\{&\tilde{H}_2[r,\bar{r}]\ell\left(2Q_0[r,L_0^{-1}B_1] - \tilde{L}_1 r\right)\\ &+ \tilde{H}_2[r,r]\ell\left(Q_0[\bar{r},(2j\omega_c I - L_0)^{-1}B_1] - 0.5\tilde{L}_1\bar{r}\right)\Big\}.\end{aligned}$$

Because $\tilde{\lambda}_2 > 0$, the Hopf bifurcation of the uncontrolled system is unstable. Hence stabilization requires $\tilde{\lambda}_2^* < 0$ to hold, which implies the condition in (32). Conversely, if (32) holds, then there exists K_2 such that $\tilde{\lambda}_2^* < 0$, which ensures the stability of the Hopf bifurcation for the feedback system. Since only the quadratic term of the nonlinear controller is involved in determination of $\tilde{\lambda}_2^*$, the stabilizing controller can be taken as quadratic. □

It is worth pointing out that condition (32) in Theorem 4.2 reduces to that of [2] if $\tilde{H}_2[r,r] = \tilde{H}_2[r,\bar{r}]$ is a nonzero real number, which was derived for state feedback control laws.

5. Applications to rotating stall control. Rotating stall is a flow instability in axial flow compressors of gas turbines, induced by bifurcations. It cannot only lead to large penalties in performance, but also cause catastrophe. Hence there is a growing interest in suppression of rotating stall using feedback control in order to extend the stable operating range, and to improve turbine-based aeroengines for axial flow compressors. A number of control laws are proposed in [8, 14, 16, 20] using rotating stall control, and are shown to be effective based on the low-order Moore–Greitzer model [18]. In this section established results on stationary bifurcations in section 3 will be applied to the Moore–Greitzer model and are shown to yield identical results as in [8, 14, 16, 20], thereby validating our results on bifurcation stabilization. Clearly our results are more general and apply to broader bifurcation instability problems other than rotating stall.

The post-stall model developed by Moore and Greitzer is of the form

$$\dot{\Psi} = \frac{1}{\beta^2}\left(\Phi - (\gamma+u)\sqrt{\Psi} + 1\right), \tag{33}$$

$$\dot{\Phi} = -\Psi + \psi_c(\Phi) + 6c_3\Phi R, \qquad \psi_c(\Phi) = c_0 + c_1\Phi + c_3\Phi^3, \tag{34}$$

$$\dot{R} = \sigma R(1 - \Phi^2 - R), \tag{35}$$

where Φ is the average flow rate, Ψ the pressure rise, R the amplitude square of the disturbance flow ($R = A^2$), and u the actuating signal implemented with throttle, which are all nondimensionalized. An obvious equilibrium (Ψ_e, Φ_e, R_e) for $u = 0$ satisfies

$$R_e = 0, \qquad \Psi_e = \psi_c(\Phi_e), \qquad \Psi_e = \frac{1}{\gamma}(1+\Phi_e)^2. \tag{36}$$

It can be easily shown that there exists $\gamma_c > 0$ such that the above equilibrium is stable for $\gamma > \gamma_c$ but unstable for $\gamma < \gamma_c$ [17]. Denote

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \Psi - \Psi_e \\ \Phi - \Phi_e \\ R \end{bmatrix}, \qquad g(x) = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} \frac{\sqrt{\Psi}}{\beta^2},$$

with $\Psi = x_1 + \Psi_e$. Then

$$\dot{x} = f(\delta\gamma, x) + g(x)u, \qquad \delta\gamma = \gamma - \gamma_c. \tag{37}$$

Thus the equilibrium in (36) is the local zero solution for $u = 0$, and both Ψ_e and Φ_e are functions of γ. Moreover the linearized system at the origin possesses exactly one zero eigenvalue at $\gamma = \gamma_c$, which implies that γ_c is the critical value. The equilibrium in (36) at the critical value of γ is determined as (see [16, 14, 20]):

$$R_c = 0, \qquad \Phi_c = 1, \qquad \Psi_c = \Psi_c(\Phi_c) = c_0 + c_1 + c_3,$$
$$\gamma_c = \frac{2}{\sqrt{\Psi_c}}, \qquad c_0 = 8/3, \qquad c_1 = 1.5, \qquad c_3 = -0.5.$$

Clearly the nonlinear system (37) can be expanded in the same form as (9) with

$$L_0 = \begin{bmatrix} -\frac{\gamma_c\beta^{-2}}{2\sqrt{\Psi_c}} & \beta^{-2} & 0 \\ -1 & 0 & 6c_3 \\ 0 & 0 & 0 \end{bmatrix}, \qquad B_1 = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} \frac{\sqrt{\Psi_c}}{\beta^2}, \tag{38}$$

$$Q_0[x,x] = \begin{bmatrix} \frac{\gamma_c}{8\beta^2\Psi_c^{\frac{3}{2}}} x_1^2 \\ 3c_3\Phi_c x_2^2 + 6c_3x_2x_3 \\ -\sigma x_3^2 - 2\sigma\Phi_c x_2 x_3 \end{bmatrix},$$

$$\tilde{L}_1 = \begin{bmatrix} -\frac{1}{2\beta^2\sqrt{\Psi_c}} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; C_0[x,x,x] = \begin{bmatrix} -\frac{\gamma_c}{16\beta^2\Psi_c^{\frac{5}{2}}} x_1^3 \\ c_3 x_2^3 \\ -\sigma x_2^2 x_3 \end{bmatrix}, \; r = \begin{bmatrix} 6c_3 \\ \frac{3c_3\gamma_c}{\sqrt{\Psi_c}} \\ 1 \end{bmatrix}, \tag{39}$$

and $\ell = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$. Thus the critical mode of the linearized system is uncontrollable. By the assumption in [16, 14, 20], $\sigma > 0$ and $\beta > 0$. For the uncontrolled system,

$$\tilde{\lambda}_1 = \ell Q_0[r,r] = -\sigma\left(1 + \frac{6c_3\gamma_c}{\sqrt{\Psi_c}}\right) = -\sigma\left(1 + \frac{12c_3}{\Psi_c}\right) > 0, \quad \lambda'(0) = -2\sigma\sqrt{\Psi_c} < 0.$$

Therefore it is a transcritical bifurcation and the bifurcated solution for $R_e > 0$ is unstable. Corollary 3.5 is now applied to compute a linear state feedback gain K_1 that stabilizes the bifurcated solution at $R_e > 0$. Recall that R is the amplitude squared of the disturbance flow. Thus $R_e < 0$ has no physical meaning. Let T be the required similarity transformation as in (23). Then

$$K_1 = \begin{bmatrix} k_\Psi & k_\Phi & k_R \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} \end{bmatrix} T, \quad K_{11} = \begin{bmatrix} k_1 & k_2 \end{bmatrix}, \quad K_{12} = k_3. \tag{40}$$

Straightforward computation gives

$$T = \begin{bmatrix} 1 & 0 & -6c_3 \\ 0 & 1 & -\frac{3c_3\gamma_c}{\sqrt{\Psi_c}} \\ 0 & 0 & 1 \end{bmatrix}, \qquad L_{00} = \begin{bmatrix} -\frac{\gamma_c\beta^{-2}}{2\sqrt{\Psi_c}} & \beta^{-2} \\ -1 & c_2 + 3c_3\Phi_c^2 \end{bmatrix}, \qquad B_{11} = \begin{bmatrix} -\frac{\sqrt{\Psi_c}}{\beta^2} \\ 0 \end{bmatrix},$$

which yields $a = -2\sqrt{\Psi_c}\sigma$, $b = 0$, and $d = \tilde{d} = 0$. Thus (i)–(iii) of Corollary 3.5 are equivalent to

$$\text{(i) } -\sigma(1 - k_2\sqrt{\Psi_c})\left[\left(1 + \frac{12c_3}{\Psi_c}\right)(1 - k_2\sqrt{\Psi_c}) + 2\sqrt{\Psi_c}k_3\right] < 0, \tag{41}$$

$$\text{(ii) } \lambda'(0)\left(1 + K_{11}L_{00}^{-1}B_{11}\right) = \lambda'(0)(1 - k_2\sqrt{\Psi_c}) < 0 \tag{42}$$

$$\text{(iii) } k_1 > -\Psi_c^{-3/2}, \quad k_2 < \Psi_c^{-1/2}. \tag{43}$$

Note that (iii) implies that (ii) is true, and (i) is reduced to

$$(1 - 6/\Psi_c)\,(1-k_2\sqrt{\Psi_c})+2\sqrt{\Psi_c}k_3 > 0, \quad k_1 = k_\Psi,\;\; k_2 = k_\Phi,\;\; k_3 = -3k_\Psi-\frac{3}{\Psi_c}k_\Phi+k_R,$$

where $c_3 = -1/2$ is used. After simplification, the stabilizing state feedback gain satisfies

$$\text{(44)} \qquad \frac{6-\Psi_c}{\Psi_c^{3/2}} < -6k_\Psi - k_\Phi + 2k_R, \qquad k_\Phi < \Psi_c^{-1/2}, \qquad k_\Psi > -\Psi_c^{-3/2},$$

which are exactly the same as in [14]. Taking $k_R = k_\Phi = 0$ with $c_0 = 8/3$, $c_1 = 1.5$, and $c_3 = -0.5$ yields

$$-\frac{3}{11}\sqrt{\frac{3}{11}} = -\Psi_c^{-3/2} < k_\Psi < \frac{1}{6\sqrt{\Psi_c}} - \Psi_c^{-3/2} = -\frac{7}{66}\sqrt{\frac{3}{11}},$$

which is the same condition obtained in [8]. Taking $k_\Psi = k_\Phi = 0$ yields

$$k_R > 3\Psi_c^{-3/2} - \frac{1}{2\sqrt{\Psi_c}} = 0.1175,$$

which implies that $k_R = 0.5$ as in [16, 20] is a stabilizing gain too.

It should be pointed out that the results of [16, 20] can also be obtained from Theorem 3.6 directly. A more interesting case is that condition (44) can be obtained using Theorem 3.2 as well. Indeed, set $R = A^2$ with A the amplitude of disturbance flow. Then (35) can be written as

$$\text{(45)} \qquad \dot{A} = 0.5\sigma A(1 - \Phi^2 - A^2).$$

Together with (33) and (34), the Moore–Greitzer model has new coordinates (Ψ, Φ, A) and rotating stall corresponds to a subcritical pitchfork bifurcation that is unstable due to $\tilde{\lambda}_1 = 0$ and $\tilde{\lambda}_2 = -\sigma(\Psi_c - 6)/\Psi_c > 0$ [16, 20]. Let $x_1 = \Psi - \Psi_e$, $x_2 = \Phi - \Phi_e$, and $x_3 = A$, and let the output measurement be given by

$$y = H_1 x + H_2[x,x] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ x_3^2 \end{bmatrix}.$$

Then it is easy to see that the critical mode of the linearized system at $\gamma = \gamma_c$ is neither controllable nor observable in light of

$$L_0 = \begin{bmatrix} -\frac{\gamma_c\beta^{-2}}{2\sqrt{\Psi_c}} & \beta^{-2} & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad B_1 = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} \frac{\sqrt{\Psi_c}}{\beta^2},$$

$$H_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \ell^T = r = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

which are of the same form as in (15) with $T = I$. By Theorem 3.2, linear controllers suffice for stabilization which have the form

$$\text{(46)} \qquad u = K_1 y = K_1(H_1 x + H_2[x,x]) = k_\Psi x_1 + k_\Phi x_2 + k_{A^2} x_3^2.$$

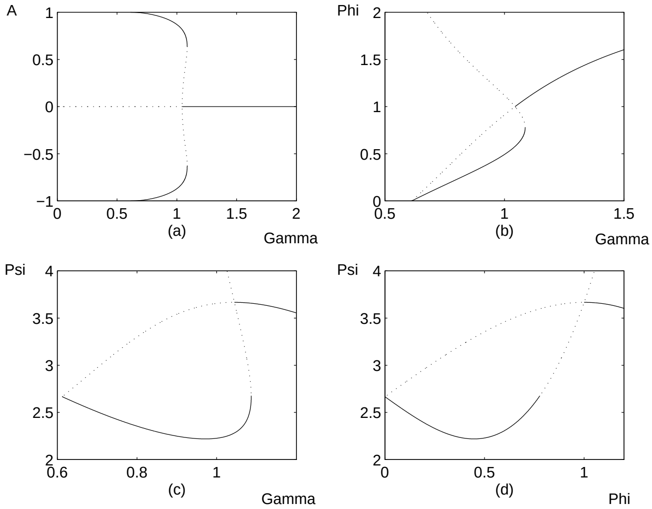


FIG. 1. *Bifurcation diagrams without feedback control.*

Direct computation gives

$$\rho = 2\sqrt{\Psi_c}, \qquad \beta = \begin{bmatrix} 0 \\ -2 \end{bmatrix}, \qquad H_{11}L_{00}^{-1}\left(B_{11}\tilde{\lambda}_2 + \rho\beta\right) = \begin{bmatrix} 6\sigma\sqrt{\Psi_c} \\ \sigma\sqrt{\Psi_c} \end{bmatrix}.$$

Hence the stabilizing conditions in Theorem 3.2 are equivalent to

$$-\frac{\Psi_c - 6}{\Psi_c^{3/2}} + 6k_\Psi + k_\Phi - 2k_{A^2} < 0, \qquad k_\Phi < \Psi_c^{-1/2}, \qquad k_\Psi > -\Psi_c^{-3/2},$$

which are exactly the same as in (44) as $k_R = k_{A^2}$ by $R = A^2$.

Numerical simulations. The compressor model with the following parameters [8, 16, 18, 20]:

$$\lambda = 1.75, \qquad H = 0.18, \qquad W = 0.25, \qquad B = 2, \qquad a = 1/3.5,$$
$$c_0 = 8/3, \qquad c_1 = 1.5, \qquad c_3 = -0.5, \qquad l_c = 8, \qquad l_F = \infty,$$

is used to illustrate the application of our results on bifurcation stabilization. Figure 1 shows four simulation plots for the uncontrolled compression system where (a) is the bifurcation diagram of A_e vs. γ with a solid line for stability and a dotted line for instability. It shows a subcritical pitchfork bifurcation associated with the hysteresis loop in rotating stall. The bifurcation diagrams in Figure 1(b)–(d) are obtained from Figure 1(a) using the relations satisfied for steady equilibrium solutions. The hysteresis loop in Figure 1(a) clearly has adverse effects when using the throttle as control actuator. This is seen from the fact that the operating points (Ψ_e, Φ_e, A_e) are not single-valued functions of γ.

We now apply control law (46) by choosing $K_\Phi = 0$ and $K_{A^2} = 0$. Then a stabilizing K_Ψ must lie in the interval of $(-\frac{3}{11}\sqrt{\frac{3}{11}}, -\frac{7}{66}\sqrt{\frac{3}{11}})$. Using a linear feedback control law with $K_\Psi = -0.0885$, the bifurcation plots are shown in Figure 2. This control law changes the pitchfork bifurcation in Figure 1(a) from subcritical to supercritical as shown in Figure 2(a). The bifurcation diagrams in Figure 2(b)–(d) show that the adverse effects of the hysteresis loop are eliminated, and (Ψ_e, Φ_e, A_e) are all single-valued functions of γ. The details can be found in [8].

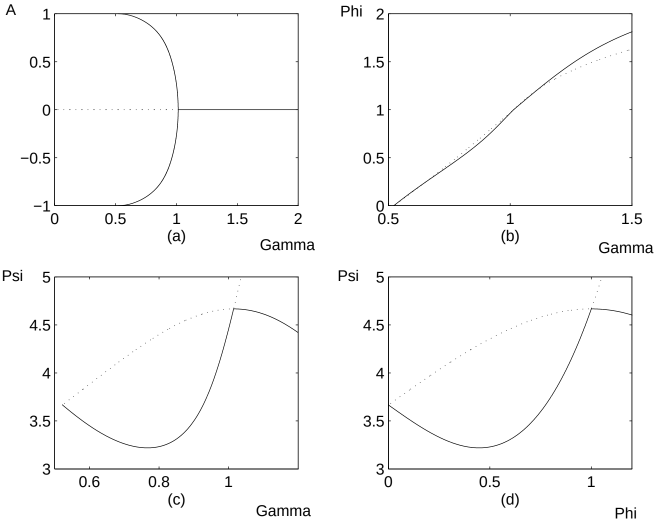


FIG. 2. *Bifurcation diagrams with feedback control.*

6. Conclusion. This paper investigated bifurcation stabilization using smooth local output feedback controllers for parameterized nonlinear systems where the critical mode of the linearized system is uncontrollable. Stabilizability conditions were established for both the case where the critical mode is linearly unobservable and the case where it is observable through output measurement. The latter case includes state feedback as a special case. It was shown that nonlinear controllers do not offer any advantage over the linear ones for bifurcation stabilization if the critical mode of the linearized system is unobservable. For the case that the critical mode of the linearized system is observable, it was shown that linear controllers are adequate for stabilization of transcritical bifurcation, and quadratic controllers are adequate for stabilization of pitchfork and Hopf bifurcations, respectively. Stabilization conditions for nonlinear bifurcations with single critical parameters were characterized in explicit form which can be used to synthesize stabilizing controllers, if they exist. The applicability of the bifurcation stabilization results was demonstrated for rotating stall control of axial flow compressors.

REFERENCES

[1] E.H. Abed, P.K. Houpt, and W.M. Hosny, *Bifurcation analysis of surge and rotating stall in axial flow compressors*, J. Turbomachinery, 115 (1993), pp. 817–824.

[2] E.H. Abed and J.H. Fu, *Local feedback stabilization and bifurcation control*, I. *Hopf bifurcation*, Systems Control Lett., 7 (1986), pp. 11–17.

[3] E.H. Abed and J.H. Fu, *Local feedback stabilization and bifurcation control*, II. *Stationary bifurcation*, Systems Control Lett., 8 (1986), pp. 467–473.

[4] D. Aeyels, *Stabilization of a class of nonlinear systems by a smooth feedback control*, Systems Control Lett., 5 (1985), pp. 467–473.

[5] O.O. Badmus, S. Chowdhury, E.M. Eveker, C.N. Nett, and C.J. Rivera, *A simplified approach for control of rotating stall—Part* 1/2, in 29th Joint Propulsion Conference and Exhibit, 1993, AIAA paper 93-2229/2334.

[6] J. Baillieul, S. Dahlgren, and B. Lehman, *Nonlinear control design for systems with bifurcations with applications to stabilization and control of compressors*, in Proceedings of the

34th IEEE Conf. Dec. and Contr., 1995, pp. 3062–3067.

[7] R.L. Behnken, R. D'Andrea, and R.M. Murray, *Control of rotating stall in a low-speed axial flow compressor using pulsed air injection: Modeling, simulations, and experimental validation*, Proceedings of the 34th IEEE Conf. Dec. and Contr., 1995, pp. 3056–3061.

[8] X. Chen, G. Gu, P. Martin, and K. Zhou, *Rotating Stall Control via Bifurcation Stabilization*, Automatica J. IFAC, 34 (1998), pp. 437–443.

[9] R.W. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometry Control Theory, R. W. Brokett, R.S. Millman, and H.J. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 181–191.

[10] L.N. Howard, *Nonlinear bifurcations*, in Nonlinear Oscillations in Biology, F.C. Hoppensteadt, ed., AMS, Providence, RI, 1979, pp. 1–68.

[11] G. Iooss and D.D. Joseph, *Elementary Stability and Bifurcation Theory*, Springer-Verlag, New York, 1980.

[12] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[13] Wei Kang, *Bifurcation control via state feedback for single input nonlinear systems: Part* I *and Part* II, in Proceedings of the 36th IEEE Conf. Dec. and Contr., 1997, pp. 1162–1167.

[14] A. Krener, *The Feedbacks Which Soften the Primary Bifurcation of MG*3, preprint, 1995.

[15] M. Krstic, J.M. Protz, J.D. Paduano, and P.V. Kokotovic, *Backstepping designs for jet engine stall and surge control*, Proceedings of the 34th IEEE Conf. Dec. and Contr., 1995, pp. 3049–3055.

[16] D.-C. Liaw and E.H. Abed, *Active control of compressor stall inception: A bifurcation-theoretical approach*, Automatica J. IFAC, 32 (1996), pp. 109–116.

[17] F.E. McCaughan, *Bifurcation analysis of axial flow compressor stability*, SIAM J. Appl. Math., 50 (1990), pp. 1232–1253.

[18] F.K. Moore and E.M. Greitzer, *A theory of post-stall transients in axial compressors: Part* I*—development of the equations*, ASME J. of Engr. for Gas Turbines and Power, 108 (1986), pp. 68–76.

[19] H.J. Sussmann, *Subanalytic sets and feedback control*, J. Differential Equations, 31 (1979), pp. 31–52.

[20] H.O. Wang, R.A. Adomaitis, and E.H. Abed, *Nonlinear analysis and control of rotating stall in axial flow compressors*, in American Control Conference, Baltimore, MD, 1994, pp. 2317–2321.

[21] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, New York, 1990.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 3, pp. 957–995

BACKSTEPPING CONTROLLER DESIGN FOR NONLINEAR STOCHASTIC SYSTEMS UNDER A RISK-SENSITIVE COST CRITERION*

ZIGANG PAN† AND TAMER BAŞAR‡

Abstract. This paper develops a methodology for recursive construction of optimal and near-optimal controllers for strict-feedback stochastic nonlinear systems under a risk-sensitive cost function criterion. The design procedure follows the integrator backstepping methodology, and the controllers obtained guarantee any desired achievable level of long-term average cost for a given risk-sensitivity parameter θ. Furthermore, they lead to closed-loop system trajectories that are bounded in probability, and in some cases asymptotically stable in the large. These results also generalize to nonlinear systems with strongly stabilizable zero dynamics. A numerical example included in the paper illustrates the analytical results.

Key words. stochastic differential equation, stochastic stability, risk-sensitive control, integrator backstepping, zero dynamics

AMS subject classifications. 93E15, 93E20, 90D25, 93C10, 90A46

PII. S0363012996307059

1. Introduction.

The topic of designing globally stabilizing controllers for nonlinear systems has been an intense area of research in recent years. A class of nonlinear systems that have attracted particular interest consists of those that can be transformed into linear time-invariant systems under a state diffeomorphism and state feedback—to so-called feedback linearizable nonlinear systems [10]. For this class of nonlinear systems, a general and flexible control design strategy was introduced in the early 1990s [15], which is now known as *integrator backstepping*. This methodology provides a general recursive constructive tool to design controllers for nonlinear systems that are given in the *strict-feedback* form and for systems that are feedback equivalent to such systems. Since the early 1990s, several additional results have been obtained for strict-feedback nonlinear systems using the integrator backstepping method; some selective references in this area are [18], [26], and [12]. This methodology has also been used to design controllers for a class of nonholonomic nonlinear systems [14]. The recent book [19] contains an up-to-date coverage of this topic, with an extensive list of references.

The integrator backstepping methodology provides a considerable amount of flexibility in the design process. This is reflected in terms of the choice of the stabilizing control laws and the additive Lyapunov functions that can be prescribed at each step of the recursive construction. One of the interesting issues that arise in the application of backstepping methodology is the identification of appropriate choices of these

*Received by the editors July 22, 1996; accepted for publication (in revised form) March 25, 1998; published electronically March 23, 1999.

http://www.siam.org/journals/sicon/37-3/30705.html

†Center for Control Engineering and Computation, Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106. Present address: Department of Electrical Engineering, Polytechnic University, Six Metrotech Center, Brooklyn, NY 11201 (pan@isaacs.poly.edu). The research of this author was supported in part by National Science Foundation CAREER grant NSF-ECS-9702702 and by AFOSR grant F49620-95-0-0409.

‡Coordinated Science Laboratory, University of Illinois, 1308 West Main Street, Urbana, IL 61801 (tbasar@decision.csl.uiuc.edu). The research of this author was supported in part by the U.S. Department of Energy grant DOE-DEFG-02-94ER13939 and in part by National Science Foundation grant NSF-ECS-93-12807.

design flexibilities at each step of the recursion, so that an improved control design is achieved at the end of the recursive construction. One approach to this problem is the inverse optimal control design [6], [7]. In [7], for example, it has been shown that given a controlled Lyapunov function for a nonlinear system, one can construct a controller that is optimal for one out of a certain class of desirable cost functions.

Another approach for reducing the choices of integrator backstepping design is to design the controller to guarantee a prespecified level of performance with respect to a given cost criterion. In a recent paper [23], we have introduced such an approach and used the backstepping design tool for guaranteed disturbance attenuation in the framework of the worst-case design methodology. This has involved obtaining controllers that achieve a zero value for a particular parametrized nonlinear differential game.

It has been known for some time that differential game problems are closely related to risk-sensitive stochastic control problems with exponentiated cost [11], [13], [25], [28], [29]. In particular, it has been shown that a general nonlinear/nonquadratic risk-sensitive stochastic control problem for continuous diffusions has an equivalent representation as a stochastic differential game [1]. If the noise intensity vanishes, then a particular (large deviation) limit of the stochastic differential game (and also that of the risk-sensitive stochastic control problem) yields a deterministic differential game. All these connections have motivated, in recent years, accelerated research on the topic of risk-sensitive stochastic control.

Because of the connections alluded to above between deterministic worst-case designs and risk-sensitive stochastic control designs for nonlinear systems, the question arises as to whether the counterpart of the results of [23] can be obtained under a risk-sensitive control formulation, using an integrator backstepping methodology. Such a study would of course also require the development of a "backstepping" tool for stochastic systems. This is precisely the problem addressed in this paper.

To this end, we introduce a class of strict-feedback stochastic nonlinear systems, along with an exponential cost function to be optimized, which is parametrized by a risk-sensitivity parameter θ. For this system, we seek to design a controller that would guarantee any desired achievable level of long-term average cost. It turns out that the construction of a controller for the risk-sensitive cost function is quite different from that of the nonlinear H^∞ problem presented in [23], the reason being that the equivalence between the deterministic dynamic game and the exponential cost problem holds only as the intensity of the system noise diminishes to zero. We present here an explicit recursive construction for the nonlinear control laws that guarantee any prespecified level of long-term average cost and lead to boundedness in probability of the closed-loop system trajectory. Three special cases are discussed in detail. One case is that in which the vector fields for the disturbance vanish at the origin of the system; in this case, the control design can actually guarantee a zero long-term average cost, and the closed-loop system becomes asymptotically stable in the large. Another special case we discuss in the paper pertains to the limit when the risk-sensitivity parameter converges to zero—the case of the risk-neutral performance index. The third special case discussed is that in which the system noise level diminishes to zero, for which we show that the control design procedure converges (in some sense) to the H^∞ control design procedure of [23], uniformly on compact subsets of the entire state space. These results admit immediate extensions to a class of nonlinear systems with strongly stabilizable zero dynamics (section 6). The theoretical findings are further illustrated in section 7 on a second-order nonlinear system.

The paper is organized as follows. In section 2, we present a precise formulation for the problem of guaranteed cost control design and provide some general theorems regarding the uncontrolled version of the exponential cost problem. We discuss the recursive construction procedure for a guaranteed cost controller design in sections 3 and 4; the former discusses a single typical backstepping process for nonlinear systems, which serves as a building block for the more general derivation given in the latter. Extensive discussions of various limits and specializations of the results are included in section 5. In section 6, we extend the result to a class of more general nonlinear systems involving nontrivial zero dynamics. A numerical example is presented in section 7. The paper presents some conclusions and remarks on future research directions in section 8. An appendix includes proofs of two results that are utilized in the main body of the paper.

2. Problem formulation.

We consider the following noise-prone strict-feedback nonlinear system, given by the Itô stochastic differential equation (SDE):[1]

$$
\begin{aligned}
&\text{(2.1a)} && dx_1 = (x_2 + f_1(x_1))\,dt + h_1'(x_1)\,dw,\\
& && \vdots \qquad\quad \vdots\\
&\text{(2.1b)} && dx_{n-1} = (x_n + f_{n-1}(x_1,\dots,x_{n-1}))\,dt + h_{n-1}'(x_1,\dots,x_{n-1})\,dw,\\
&\text{(2.1c)} && dx_n = (f_n(x_1,\dots,x_n) + b(x_1,\dots,x_n)u)\,dt + h_n'(x_1,\dots,x_n)\,dw,\\
&\text{(2.1d)} && y = x_1\,.
\end{aligned}
$$

Here, $x := (x_1,\dots,x_n)'$ is the n-dimensional state vector, with a fixed initial value $x(0)$; u is the scalar control input; w is the q-dimensional vector-valued standard Wiener process; and y is the scalar output. In compact form, the SDE (2.1) can be expressed as

$$dx = (f(x) + G(x)u)\,dt + H(x)\,dw. \tag{2.2}$$

Note that, in addition to the strict-feedback form introduced in [15] in a deterministic setting, the above system incorporates additive stochastic disturbance inputs, where the nonlinear function multiplying the disturbance terms are also in triangular form. The underlying probability space is taken to be the triplet $(\Omega, \mathcal{F}, \boldsymbol{P})$, where Ω is the sample space, $\mathcal{F}$ is a filtration, and $\boldsymbol{P}$ is the probability measure.

For the nonlinear system (2.1), we make the following basic assumption as a starting point of our study.

Assumption A1. The nonlinear functions f_i and h_i are $\mathcal{C}^\infty$ in all their arguments (or simply are smooth), $i = 1,\dots,n$. The nonlinear functions b and $1/b$ are $\mathcal{C}^2$ in all their arguments. The functions f_i, $i = 1,\dots,n$, vanish at $x = 0$: $f_i(x_1,\dots,x_i)|_{x=0} = 0$, $i = 1,\dots,n$. □

The first part of Assumption A1 is a standard smoothness assumption for this class of nonlinear systems; the condition imposed on f_i at $x = 0$ is to ensure that the origin is an equilibrium point of the deterministic (unperturbed) part of the system.

The control input u depends on the current value of the state and hence is generated by

$$u(t) = \mu(x(t)), \tag{2.3}$$

[1] In a deterministic form, systems having the structure (2.1) arise in several applications (see [19]), one of which is robotic manipulators [27]. Here, we have taken the disturbance input to be a Wiener process instead of a deterministic unknown function. We should also note that stochastic systems in the state-space form (2.2) with general dynamics can be brought to the strict-feedback form (2.1) via a diffeomorphic transformation as shown in [22]. This therefore justifies the study of stochastic systems in the form of (2.1).

where the mapping $\mu : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz. We denote the class of all such controllers by $\mathcal{M}$.

The objective of the controller design is to maintain a finite (and in fact arbitrarily small, positive) long-term average value for the following *risk-sensitive* cost function:

$$J_\theta(\mu) = \limsup_{t_f \to \infty} \frac{2}{\theta t_f} \ln \left\{ \boldsymbol{E} \left\{ \exp \left[\frac{\theta}{2} \left(\int_0^{t_f} y^2(t)\, dt \right) \right] \right\} \right\}, \tag{2.4}$$

where θ is the risk-sensitivity parameter. When θ is positive, the cost functional weights heavily the large deviations of y (from a base zero) through the exponential operator, which leads to a risk-averse control design. The larger θ is, the more conservative is the controller. On the other hand, if θ is negative, the optimization will lead to a risk-seeking controller design. If we take $\theta \to 0$, then the cost function converges to a standard integral cost, with the underlying stochastic control problem also known as a *risk-neutral* problem. In this paper, we will study only the case in which θ is positive. The risk-seeking case (i. e., with $\theta < 0$) can be studied analogously but will not be covered here.

An important point to note here is that in the performance function (2.4) there is no weighting on the control input, and hence control can be chosen arbitrarily large without incurring any cost. This, coupled with the strict-feedback structure, allows an arbitrarily small positive long-term average value to be guaranteed for the closed-loop system, for any given risk-sensitivity parameter θ—a result that will actually be established in this paper. The smaller the long-term average cost is, the larger the control effort will generally be. Similarly, the larger the value of θ is, the higher will be the induced weighting on large excursions of the output y, which implies a better systems response if the long-term average cost is maintained at a constant value, but at the expense of a larger control effort. Motivated by these considerations, we embed the original problem of minimizing (2.4) in a larger class of "cost-level satisfaction" problems, where the objective is to find a controller that leads to satisfaction of a given bound on the risk-sensitive cost. This is captured in the following definition.

DEFINITION 2.1. *Given a risk-sensitivity parameter θ, a controller μ is said to achieve a guaranteed risk-sensitive cost $R_c \geq 0$, if the following inequality holds:*

$$J_\theta(\mu) = \limsup_{t_f \to \infty} \frac{2}{\theta t_f} \ln \left\{ \boldsymbol{E} \left\{ \exp \left[\frac{\theta}{2} \left(\int_0^{t_f} (y^2(t) + l(x(t)))\, dt \right) \right] \right\} \right\} \leq R_c, \tag{2.5}$$

where the function $l(x(t))$ is nonnegative and is chosen by the designer.

Our objective, succinctly stated, is given $\theta > 0$ and any $R_c > 0$ ($R_c = 0$ under the additional assumption $H(0) = 0$), to construct a controller that achieves a risk-sensitive cost of R_c. To achieve this, our approach to the problem is recursive construction of the value function using the *integrator backstepping* procedure. To this end, we first present a result that deals with the issues of existence, uniqueness, stability, and performance of a control free risk-sensitive problem. Such a risk-sensitive problem will correspond to the closed-loop system of (2.1) under a particular smooth control design.

We note at this point that the nonlinear continuous-time risk-sensitive stochastic control problem has been treated before by many authors, including the work in [5], [3], [1], [4], and [25]. The most recent paper in this area is [20], in which the general risk-sensitive control problem has been treated in both finite- and infinite-horizon cases under some global Lipschitz conditions. Yet the results presented in all these references are not completely suitable for the class of problems under consideration in this paper, since we are considering here the infinite-horizon case with arbitrary (globally non-Lipschitzian) nonlinearities. In contradistinction with the early references, though, we are interested here in designing a suboptimal control strategy rather than an optimal one, which avoids the measure transformation that is commonly used for this type of a problem.

Let us now consider the SDE

$$dx = f(x)\,dt + H(x)\,dw \tag{2.6}$$

along with a risk-sensitive cost function:

$$J_\theta = \limsup_{t_f\to\infty} \frac{2}{\theta t_f} \ln\left\{\boldsymbol{E}\left\{\exp\left[\frac{\theta}{2}\int_0^{t_f} q(x(t))\,dt\right]\right\}\right\}. \tag{2.7}$$

Let us introduce the value function, $W(t_f - t; x)$, associated with the finite-horizon version of this problem (on interval $[t, t_f]$):

$$W(t_f - t; x) = \frac{2}{\theta}\ln\left\{\boldsymbol{E}\left\{\exp\left[\frac{\theta}{2}\int_t^{t_f} q(x(s))\,ds\right]\right\}\right\} \tag{2.8}$$

for a system with initial time-state pair (t, x). The Hamilton–Jacobi–Bellman (HJB) equation satisfied by W is [1]

$$\frac{\partial W}{\partial t} + \frac{\partial W}{\partial x} f(x) + \frac{\theta}{4}\frac{\partial W}{\partial x} H(x)H'(x)\left(\frac{\partial W}{\partial x}\right)' + \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 W}{\partial x^2} H(x)H'(x)\right) + q(x) = 0. \tag{2.9}$$

Generally (and in particular if $H(x)H'(x) > \delta I\ \forall x \in \mathbb{R}^n$, for some $\delta > 0$), the value function $W(t_f - t; x)$ becomes unbounded as the terminal time t_f goes to infinity. The rate at which it goes to infinity corresponds to the long-term average cost for the infinite-horizon problem. Also in view of the results of [20], it is therefore reasonable to assume the following structure for W:

$$W(t_f - t; x) = V(x) + (t_f - t)R_c,$$

where R_c is a desired long-term average cost.

In terms of this constant, $V(x)$ satisfies the following HJB equation:

$$\frac{\partial V}{\partial x} f(x) + \frac{\theta}{4}\frac{\partial V}{\partial x} H(x)H'(x)\left(\frac{\partial V}{\partial x}\right)' + \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 V}{\partial x^2} H(x)H'(x)\right) + q(x) = R_c. \tag{2.10}$$

Since we are interested in finding a controller that guarantees the desired long-term average cost R_c, we can relax the equality in (2.10) to an inequality, and thus search for a $V(x)$ that satisfies the following HJB inequality:

$$\frac{\partial V}{\partial x} f(x) + \frac{\theta}{4}\frac{\partial V}{\partial x} H(x)H'(x)\left(\frac{\partial V}{\partial x}\right)' + \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 V}{\partial x^2} H(x)H'(x)\right) + q(x) = \Delta(x) \le R_c, \tag{2.11}$$

where R_c is the desired long-term average risk-sensitive cost that uniformly bounds the function $\Delta(x)$. We will prove that the existence of a $\mathcal{C}^2$ solution to the above

HJB inequality yields the desired existence, uniqueness, stability, and performance requirements for the original risk-sensitive control problem defined by (2.6) and (2.7).

First, we recall the stochastic process notions of *bounded in probability* and *asymptotic stability in the large*, as introduced in Chapter 1, Section 4 and Chapter 5, Section 4 of the classical book [17]. These notions will be useful in our development.

DEFINITION 2.2. *A stochastic process* $\{x(t), t \geq 0\}$ *is said to be* bounded in probability *if*

$$\lim_{c\to\infty} \sup_{0\leq t<\infty} \boldsymbol{P}\{|x(t)| > c\} = 0. \tag{2.12}$$

DEFINITION 2.3. *Consider the SDE* (2.6), *with* $H(0) = 0$. *The solution* $x(t) \equiv 0$ *of such an SDE is said to be* asymptotically stable in the large *if for any* $\epsilon > 0$,

$$\lim_{|x_0|\to 0} \boldsymbol{P}\{\sup_{t\geq 0} |x(t)| \geq \epsilon\} = 0 \tag{2.13}$$

and for any initial condition x_0,

$$\boldsymbol{P}\{\lim_{t\to\infty} x(t) = 0\} = 1. \tag{2.14}$$

DEFINITION 2.4. *A scalar function* $V : \mathbb{R}^n \to \mathbb{R}$ *is* positive definite *if* $V(0) = 0$ *and* $V(x) > 0\ \forall x \in \mathbb{R}^n \setminus \{0\}$. V *is* radially unbounded *if* $V(x) \to \infty$ *as* $|x| \to \infty$.

We are now in a position to state (and prove) the following theorem.

THEOREM 2.5. *Consider the SDE* (2.6) *with the risk-sensitive cost function* (2.7). *Let* $\theta > 0$ *and* $R_c \geq 0$ *be fixed. Assume that the vector field* $f(x)$ *and the matrix-valued function* $H(x)$ *are* $\mathcal{C}^1$ *in their arguments, and the function* $q(x)$ *is nonnegative definite and is* $\mathcal{C}^2$ *in its argument. Further assume that there exists a positive definite and radially unbounded* $\mathcal{C}^2$ *value function* $V(x)$ *satisfying the HJB inequality* (2.11). *Then the following statements are true.*

1. *There exists an almost surely (a.s.)* $\boldsymbol{P}$ *unique solution to the SDE* (2.6) *on* $[0, \infty)$.
2. *The cost* (2.7) *is upper bounded by* R_c.
3. *If, in addition,* $V(x)$ *and* $q(x)$ *satisfy the linear bound relationship*[2]

$$V(x) \leq c_1 q(x) + c_2 \qquad \forall x \in \mathbb{R}^n \tag{2.15}$$

for some positive constants c_1, c_2, *then the solution* $x(t)$ *to the SDE* (2.6) *is bounded in probability.*

4. *If* $R_c = 0$, $H(0) = 0$, *and the function* $q(x)$ *is positive definite, then the SDE* (2.6) *is asymptotically stable in the large.*

Proof. Consider the function $\tilde{W}(x) := V(x) + R_c$. By the HJB inequality (2.11), we have

$$\begin{aligned}\frac{\partial \tilde{W}}{\partial x} f(x) + \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 \tilde{W}}{\partial x^2} H(x)H(x)'\right) &= -\frac{\theta}{4}\frac{\partial V}{\partial x} H(x)H(x)' \left(\frac{\partial V}{\partial x}\right)' - q(x) + \Delta(x)\\ &\leq \Delta(x) \leq R_c \leq \tilde{W}(x)\end{aligned}$$

for any $x \in \mathbb{R}^n$. Since the function $V(x)$ is radially unbounded, $\tilde{W}$ is also radially unbounded. Then, by Theorem 4.1 of Chapter 3 of [17], we conclude that there exists

[2] In section 4, we will provide an explicit construction for a function V that satisfies this bound.

a solution to the SDE (2.6) on the infinite interval $[0,\infty)$, and that the solution is a.s. $\boldsymbol{P}$ unique. This completes the proof of the first statement.

For the second statement, we apply Itô's rule to the stochastic process $V(x(t))$ generated by composing $V(\cdot)$ with the stochastic process $x(t),\ t\geq 0$:

$$\begin{aligned} dV(x(t)) &= \left(\frac{\partial V}{\partial x}f(x)+\frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 V}{\partial x^2}H(x)H'(x)\right)\right)dt+\frac{\partial V}{\partial x}H(x)\,dw \\ &= \frac{\partial V}{\partial x}H(x)\,dw-\left(\frac{\theta}{4}\frac{\partial V}{\partial x}H(x)H'(x)\left(\frac{\partial V}{\partial x}\right)'+q(x)-\Delta(x)\right)dt, \end{aligned}$$

where in the second line we have made use of (2.11).

Integrating both sides of this equation, we arrive at

$$\begin{aligned} \int_0^{t_f} q(x)\,dt &= \int_0^{t_f}\left(\frac{\partial V}{\partial x}H(x)\,dw-\left(\frac{\theta}{4}\frac{\partial V}{\partial x}H(x)H'(x)\left(\frac{\partial V}{\partial x}\right)'-\Delta(x)\right)dt\right) \\ &\quad +V(x(0))-V(x(t_f)) \\ &\leq V(x(0))+\int_0^{t_f}\left(\frac{\partial V}{\partial x}H(x)\,dw-\frac{\theta}{4}\frac{\partial V}{\partial x}H(x)H'(x)\left(\frac{\partial V}{\partial x}\right)'dt\right)+R_c t_f, \end{aligned}$$

where in the second line we have used the bound $\Delta(x)\leq R_c$ and replaced $V(x(t_f))$ by its lower bound 0. Using this expression in the cost function J_θ, we have

$$\begin{aligned} J_\theta \leq \limsup_{t_f\to\infty}\frac{2}{\theta t_f}\ln&\left\{\boldsymbol{E}\left\{\exp\left[\int_0^{t_f}\left(\frac{\theta}{2}\frac{\partial V}{\partial x}H(x)\,dw-\frac{\theta^2}{8}\left|H'(x)\left(\frac{\partial V}{\partial x}\right)'\right|^2 dt\right)\right]\right\}\right\} \\ &+R_c. \end{aligned}$$

The stochastic process

$$\zeta(t):=\exp\left[\int_0^t\left(\frac{\theta}{2}\frac{\partial V}{\partial x}H(x)\,dw-\frac{\theta^2}{8}\frac{\partial V}{\partial x}H(x)H'(x)\left(\frac{\partial V}{\partial x}\right)'dt\right)\right] \tag{2.16}$$

is almost surely positive and is a supermartingale (see the appendix). Hence,

$$\boldsymbol{E}(\zeta(t_f))\leq 1 \qquad \forall t_f>0.$$

Using this bound in the inequality for J_θ yields the desired bound

$$J_\theta\leq R_c.$$

This establishes the second statement of the theorem.

The third statement follows from Theorem 2 of Chapter 3, Section 13 of [8], which says that, for any $\delta>0$,

$$\sup_{0\leq t<\infty}\boldsymbol{P}\{V(x(t))>\delta\}\leq\frac{1}{\delta}\left(V(x(0))+\frac{c_2+R_c}{c_1}\right)$$

because of the HJB inequality (2.11).

Since V is a positive definite and radially unbounded function, there exists a continuous strictly increasing scalar function $\alpha : [0, \infty) \to [0, \infty)$ with the property $\lim_{s\to\infty} \alpha(s) = \infty$ such that (see Lemma 3.5 in [16])

$$V(x) \geq \alpha(|x|).$$

Consequently,

$$\begin{aligned}\sup_{0\leq t<\infty} \boldsymbol{P}\{|x(t)| > \delta\} &= \sup_{0\leq t<\infty} \boldsymbol{P}\{\alpha(|x|) > \alpha(\delta)\} \leq \sup_{0\leq t<\infty} \boldsymbol{P}\{V(x(t)) > \alpha(\delta)\} \\ &\leq \frac{1}{\alpha(\delta)} \left(V(x(0)) + \frac{c_2 + R_c}{c_1} \right).\end{aligned}$$

The third statement follows by taking the limit $\delta \to \infty$.

When $R_c = 0$, the last statement is a direct consequence of Theorem 4.4 of Chapter 5 of [17].

This completes the proof of the theorem. □

In the development to follow, we will make extensive use of an equivalent form of the HJB equation (2.11), which is made precise by the following proposition.

PROPOSITION 2.6. *A $\mathcal{C}^2$ function $V(x)$ satisfies the HJB equation* (2.11) *if and only if the following algebraic relationship holds:*

$$\begin{aligned}\frac{\partial V}{\partial x} f(x)\, dt + \frac{\partial V}{\partial x} H(x)\, dw + \frac{1}{2} \mathrm{Tr}\left(\frac{\partial^2 V}{\partial x^2} H(x) H'(x) \right) dt \\ = \sigma(x)\, dw - \left(\frac{\theta}{4} \sigma(x)\sigma'(x) + q(x) - \Delta(x) \right) dt,\end{aligned} \tag{2.17}$$

where $\sigma(x) = \frac{\partial V}{\partial x} H(x)$ for all $x \in \mathbb{R}^n$.

Proof. Sufficiency. By the hypothesis of the proposition, we have

$$\begin{aligned}\frac{\partial V}{\partial x} f(x)\, dt + \frac{\partial V}{\partial x} H(x)\, dw + \frac{1}{2} \mathrm{Tr}\left(\frac{\partial^2 V}{\partial x^2} H(x) H'(x) \right) dt \\ = \sigma(x)\, dw - \left(\frac{\theta}{4} \sigma(x)\sigma'(x) + q(x) - \Delta(x) \right) dt.\end{aligned} \tag{2.18}$$

Taking expectations on both sides of the equation, we get

$$\frac{\partial V}{\partial x} f(x)\, dt + \frac{1}{2} \mathrm{Tr}\left(\frac{\partial^2 V}{\partial x^2} H(x) H'(x) \right) dt = -\left(\frac{\theta}{4} \sigma(x)\sigma'(x) + q(x) - \Delta(x) \right) dt.$$

Therefore, by dividing both sides by dt, we have

$$\frac{\partial V}{\partial x} f(x) + \frac{\theta}{4} \sigma(x)\sigma'(x) + q(x) + \frac{1}{2} \mathrm{Tr}\left(\frac{\partial^2 V}{\partial x^2} H(x) H'(x) \right) = \Delta(x).$$

Since this holds for any initial condition $x \in \mathbb{R}^n$, the function $V(x)$ satisfies the HJB equation (2.11). Substituting this into (2.18), we have

$$\frac{\partial V}{\partial x} H(x)\, dw = \sigma(x)\, dw.$$

By Lemma A.2 in the appendix, we have

$$\frac{\partial V}{\partial x} H(x) = \sigma(x).$$

Then, it holds for all $x \in \mathbb{R}^n$. Substitution of this equality into (2.18) leads to the HJB equation (2.11).

This completes the proof of the sufficiency part of the proposition.

Necessity. Since the HJB equation (2.11) is satisfied by the $\mathcal{C}^2$ function $V(x)$,

$$\begin{aligned}\left(\frac{\partial V}{\partial x} f(x) + \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 V}{\partial x^2} H(x)H'(x)\right)\right) dt + \frac{\partial V}{\partial x} H(x)\, dw \\ = \frac{\partial V}{\partial x} H(x)\, dw - \left(\frac{\theta}{4}\sigma(x)\sigma'(x) + q(x) - \Delta(x)\right) dt.\end{aligned}$$

This completes the proof of the necessity part of the proposition. □

REMARK 2.1. *We note that in Proposition* 2.6, *the left-hand side of* (2.6) *equals the Itô differential of the stochastic process $V(x(t))$ with $x(t)$ generated by the SDE*

$$dx = f(x)\, dt + H(x)\, dw, \qquad x(0) = x, \tag{2.19}$$

whenever it exists. As a consequence of the property that the function $V(x)$ satisfies the HJB equation (2.11), *the Itô differential of $V(x(t))$ exists under additional growth conditions delineated in Theorem* 2.5. *These growth conditions will automatically be satisfied by the recursive construction procedure to be described later. For notational simplicity, we henceforth identify $dV(x(t))$ with*

$$\left(\frac{\partial V}{\partial x} f(x) + \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 V}{\partial x^2} H(x)H'(x)\right)\right) dt + \frac{\partial V}{\partial x} H(x)\, dw$$

for a candidate $\mathcal{C}^2$ value function $V(x)$ without verifying the existence of the Itô differential. The validity of this correspondence, which is equivalent to the existence of the Itô differential of $V(x(t))$, will be justified by a later application of Theorem 2.5.

We now turn, in the next two sections, to the development of a recursive construction procedure for a risk-sensitive controller with a desired guaranteed cost. In the next section, we first study a typical step in the recursive design, which provides a technical tool in the construction of a cost bounding controller for the full-order system.

3. Recursive design.

In this section, we present a systematic recursive design methodology for constructing a guaranteed cost controller for the risk-sensitive stochastic control problem.

Consider the following two-level problem:

$$d\eta = (f_1(\eta) + g_1(\eta)\xi)\, dt + H_1(\eta)\, dw, \tag{3.1a}$$

$$d\xi = (f_2(\eta, \xi) + v)\, dt + H_2(\eta, \xi)\, dw, \tag{3.1b}$$

where η is of dimension $n_1 > 0$, and ξ is a scalar. The exogenous disturbance process w is as defined in section 2. The scalar v is the control variable for this two-level system. All nonlinear functions involved are assumed to be smooth.

Suppose that the control design for the η system has been completed with ξ as the control input, and the first component of η as the output y. Further assume that this design has resulted in a smooth positive definite and radially unbounded value function $V_1(\eta)$ and a smooth control law $\alpha_1(\eta)$ (for ξ) when the running cost in (2.4) has an additional additive nonnegative term $q_1(\eta)$. Let a guaranteed time-average cost level be $R_1 \geq 0$, which can be picked as *zero* if $H_1(0) = 0$, and positive (but arbitrarily close to *zero*) otherwise. Further suppose that the following four induction hypotheses hold.

1. The equilibrium condition $\alpha_1(0) = 0$.
2. If $\xi = \alpha_1(\eta)$, the value function V_1 for this system satisfies the HJB inequality

$$\frac{\partial V_1}{\partial \eta}(f_1(\eta) + g_1(\eta)\alpha_1(\eta)) + \frac{\theta}{4}\frac{\partial V_1}{\partial \eta} H_1(\eta)H_1'(\eta)\left(\frac{\partial V_1}{\partial \eta}\right)' + y^2 + q_1(\eta) + \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 V_1}{\partial \eta^2} H_1(\eta)H_1'(\eta)\right) =: \Delta_1(\eta) \leq R_1 \tag{3.2}$$

3. The weighting function $q_1(\eta)$ is smooth and positive definite, and there exist constants $c_1 > 0$, $c_2 > 0$, $c_3 > 0$, and $c_4 > 0$, such that

$$q_1(\eta) \geq c_1 \eta' \eta \qquad \forall |\eta| \leq c_2 \tag{3.3}$$

and

$$V_1(\eta) \leq c_3\, q_1(\eta) + c_4 \qquad \forall \eta \in \mathbb{R}^{n_1}. \tag{3.4}$$

4. When $H_1(0) = 0$, the value function V_1 and control law α_1 can be chosen in such a way to make the long-term average cost *zero* (i.e., $R_1 = 0$).

These hypotheses lead to the conclusion that the η system admits a guaranteed risk-sensitive cost R_1 with risk-sensitivity parameter θ, i.e.,

$$\limsup_{t_f \to \infty} \frac{2}{\theta t_f} \ln \left\{ \boldsymbol{E} \left\{ \exp\left[\frac{\theta}{2} \left(\int_0^{t_f} (y^2(t) + q_1(\eta(t)))\, dt \right) \right] \right\} \right\} \leq R_1$$

and is stochastically bounded, if we can set $\xi = \alpha_1(\eta)$. When $H_1(0) = 0$, the guaranteed cost R_1 is identically zero. In this case, the closed-loop system is asymptotically stable in the large.

Among the conditions above, we stress the importance of the quadratic growth condition on the function q_1 of the induction hypothesis 3. This growth condition is crucial for the function q_1 to bound the quadratic variation term that results from the ξ subsystem when R_2 is identically zero. This point will be elucidated shortly.

In the design of the control law α_1 and the value function V_1, we have taken ξ as a control input. In reality, ξ is the state of an SDE driven by the control input v. Hence, to complete the design process, we next consider the control design for the entire (η, ξ) system, based on the knowledge of the control design for the η subsystem just completed. Our objective for the design of the (η, ξ) system is to achieve the time-average cost bound

$$\limsup_{t_f \to \infty} \frac{2}{\theta t_f} \ln \left\{ \boldsymbol{E} \left\{ \exp\left[\frac{\theta}{2} \left(\int_0^{t_f} (y^2 + (1-\lambda) q_1(\eta) + q_2(\eta, \xi))\, dt \right) \right] \right\} \right\} \leq R_1 + R_2$$

and ensure stochastic boundedness of the closed-loop system, where the constant $\lambda \in (0,1)$ is a design constant. In the above, q_2 is another arbitrary nonnegative functional chosen by the control designer, and R_2 is a nonnegative constant, reflecting the design specification. Note further that $q_1(\eta)$ has been rescaled by a factor of $1-\lambda$. When $H_1(0) = 0$ and $H_2(0,0) = 0$, we take R_2 to be zero, which reflects our desire to guarantee an asymptotically stable (in the sense of Definition 2.3) closed-loop system for this special case. Otherwise, R_2 is taken to be positive.

To obtain a controller for the composite system, we consider the state transformation

$$z := \xi - \alpha_1(\eta), \tag{3.5}$$

where z represents the deviation of the random process ξ from its desired trajectory $\alpha_1(\eta)$.

Let us now introduce a convention that is adopted for the rest of this section. Any function symbol marked with an overbar will denote a function defined in terms of the transformed state variables $(\eta', z)'$, such as $\bar{a}$ denoting the equivalent form of function a (of $(\eta', \xi)'$) in terms of transformed state variables, $(\eta', z)'$.

In terms of this transformed state variable, we can rewrite the SDE (3.1) as follows:

$$\begin{aligned} d\eta &= (f_1 + g_1\alpha_1 + g_1 z)\, dt + H_1\, dw \\ dz &= \left(f_2 + v - \frac{\partial \alpha_1}{\partial \eta}(f_1 + g_1\alpha_1 + g_1 z) - \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 \alpha_1}{\partial \eta^2} H_1 H_1' \right) \right) dt \\ &\quad + \left(H_2 - \frac{\partial \alpha_1}{\partial \eta} H_1 \right) dw. \end{aligned}$$

We introduce the following terms:

$$\begin{aligned} \bar{a}_2(\eta, z) &:= f_2 - \frac{\partial \alpha_1}{\partial \eta}(f_1 + g_1\alpha_1 + g_1 z) - \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 \alpha_1}{\partial \eta^2} H_1 H_1' \right), \\ \bar{d}_2(\eta, z) &:= H_2 - \frac{\partial \alpha_1}{\partial \eta} H_1. \end{aligned}$$

We note here that the function $\bar{a}_2$ includes the quadratic variation of α_1 and does not vanish at $\eta = 0$, in general. Since $\bar{d}_2$ is a smooth (nonlinear) function of (η, z), we can rewrite it as follows, where $\bar{d}_{22}$ is also smooth:

$$\bar{d}_2(\eta, z) =: \bar{d}_{21}(\eta) + z\bar{d}_{22}(\eta, z).$$

Using the preceding notation, we have

$$dz = (\bar{a}_2 + v)\, dt + (\bar{d}_{21} + z\bar{d}_{22})\, dw.$$

Consider the following smooth function $V_2(\eta, \xi)$ as a candidate value function for the (η, ξ) system:

$$\bar{V}_2(\eta, z) := V_1(\eta) + \Xi(\eta) z^2, \tag{3.6}$$

where Ξ is some positive smooth function yet to be determined. In forming the value function V_2, we have added to the previous value function V_1 the term Ξz^2, instead of a simple z^2. This is motivated by the fact that the quadratic variation term of the difference $V_2 - V_1$ does not in general vanish at $z = 0$. To guarantee the desired long-term average performance, we have to introduce the Ξ function into the design, as will be clear shortly.

Using Itô's rule, the random process $\bar{V}_2(\eta(t), z(t))$ satisfies the following SDE: [3]

$$\begin{aligned} d\bar{V}_2(\eta(t), z(t)) &= dV_1(\eta(t)) + \begin{bmatrix} z^2 \left(\dfrac{\partial \Xi}{\partial \eta} \right)' \\ 2z\Xi \end{bmatrix}' \begin{bmatrix} (f_1 + g_1\alpha_1 + g_1 z)\, dt + H_1\, dw \\ (\bar{a}_2 + v)\, dt + (\bar{d}_{21} + z\bar{d}_{22})\, dw \end{bmatrix} \\ &\quad + \frac{1}{2}\mathrm{Tr}\left(\begin{bmatrix} z^2 \dfrac{\partial^2 \Xi}{\partial \eta^2} & 2z \left(\dfrac{\partial \Xi}{\partial \eta} \right)' \\ 2z \dfrac{\partial \Xi}{\partial \eta} & 2\Xi \end{bmatrix} \begin{bmatrix} H_1 \\ \bar{d}_{21} + z\bar{d}_{22} \end{bmatrix} \begin{bmatrix} H_1 \\ \bar{d}_{21} + z\bar{d}_{22} \end{bmatrix}' \right) dt \end{aligned}$$

[3] All these differentials exist, as the explicit construction of V_1, $\bar{V}_2$, ... will reveal.

Since the differential for $V_1(\eta(t))$ is

$$dV_1(\eta(t)) = -\left(\frac{\theta}{4}\frac{\partial V_1}{\partial \eta} H_1 H_1' \left(\frac{\partial V_1}{\partial \eta}\right)' + y^2 + q_1 - \frac{\partial V_1}{\partial \eta} g_1 z - \Delta_1\right) dt \\ + \frac{\partial V_1}{\partial \eta} H_1 \, dw,$$

we further have

$$\text{(3.7)}\quad d\bar{V}_2(\eta(t), z(t)) = -y^2\, dt - q_1\, dt + \bar{\sigma}_2\, dw - \frac{\theta}{4}\frac{\partial V_1}{\partial \eta} H_1 H_1' \frac{\partial V_1}{\partial \eta}'\, dt + \frac{\partial V_1}{\partial \eta} g_1 z\, dt \\ + \Delta_1\, dt + \begin{bmatrix} z^2 \left(\dfrac{\partial \Xi}{\partial \eta}\right)' \\ 2z\Xi \end{bmatrix}' \begin{bmatrix} f_1 + g_1\alpha_1 + g_1 z \\ \bar{a}_2 + v \end{bmatrix} dt \\ + \frac{1}{2}\mathrm{Tr}\left(\begin{bmatrix} z^2 \dfrac{\partial^2 \Xi}{\partial \eta^2} & 2z\left(\dfrac{\partial \Xi}{\partial \eta}\right)' \\ 2z\dfrac{\partial \Xi}{\partial \eta} & 2\Xi \end{bmatrix} \begin{bmatrix} H_1 H_1' & H_1 \bar{d}_2' \\ \bar{d}_2 H_1' & (\bar{d}_{21} + z\bar{d}_{22})(\bar{d}_{21} + z\bar{d}_{22})' \end{bmatrix} \right) dt,$$

where

$$\bar{\sigma}_2(\eta, z) := \begin{bmatrix} \dfrac{\partial \bar{V}_2}{\partial \eta} & \dfrac{\partial \bar{V}_2}{\partial z} \end{bmatrix} \begin{bmatrix} H_1 \\ \bar{d}_2 \end{bmatrix} \equiv \frac{\partial V_1}{\partial \eta} H_1 + \begin{bmatrix} z^2 \left(\dfrac{\partial \Xi}{\partial \eta}\right)' \\ 2z\Xi \end{bmatrix}' \begin{bmatrix} H_1 \\ \bar{d}_2 \end{bmatrix}.$$

We seek a control law α_2 such that $\bar{V}_2$ satisfies the following HJB inequality:

$$\text{(3.8)}\quad \begin{bmatrix} \dfrac{\partial V_2}{\partial \eta} \\ \dfrac{\partial V_2}{\partial \xi} \end{bmatrix}' \begin{bmatrix} f_1(\eta) + g_1(\eta)\xi \\ f_2(\eta, \xi) + \alpha_2(\eta, \xi) \end{bmatrix} + y^2 + (1-\lambda) q_1(\eta) + q_2(\eta, \xi) \\ + \frac{\theta}{4} \begin{bmatrix} \dfrac{\partial V_2}{\partial \eta} \\ \dfrac{\partial V_2}{\partial \xi} \end{bmatrix}' \begin{bmatrix} H_1(\eta) \\ H_2(\eta, \xi) \end{bmatrix} \begin{bmatrix} H_1(\eta) \\ H_2(\eta, \xi) \end{bmatrix}' \begin{bmatrix} \dfrac{\partial V_2}{\partial \eta} \\ \dfrac{\partial V_2}{\partial \xi} \end{bmatrix} \\ + \frac{1}{2}\mathrm{Tr}\left(\begin{bmatrix} \dfrac{\partial^2 V_2}{\partial \eta^2} & \dfrac{\partial^2 V_2}{\partial \eta \partial \xi} \\ \dfrac{\partial^2 V_2}{\partial \eta \partial \xi} & \dfrac{\partial^2 V_2}{\partial \xi^2} \end{bmatrix} \begin{bmatrix} H_1(\eta) \\ H_2(\eta, \xi) \end{bmatrix} \begin{bmatrix} H_1(\eta) \\ H_2(\eta, \xi) \end{bmatrix}' \right) \le R_1 + R_2$$

for some nonnegative function q_2, by making use of the equivalence given in Proposition 2.6. To this end, we add and subtract $\frac{\theta}{4}\bar{\sigma}_2(\eta, z)\bar{\sigma}_2'(\eta, z)\, dt$ and $\bar{\beta}(\eta, z) z^2\, dt$ to the right-hand side of (3.7). This yields

$$d\bar{V}_2(\eta(t), z(t)) = -y^2\, dt - q_1\, dt - \bar{\beta} z^2\, dt + \bar{\sigma}_2\, dw - \frac{\theta}{4}\bar{\sigma}_2\bar{\sigma}_2'\, dt + \Delta_1\, dt \\ + \bar{\beta} z^2\, dt + \frac{\theta}{4}\bar{\sigma}_2\bar{\sigma}_2'\, dt + \frac{\partial V_1}{\partial \eta} g_1 z\, dt - \frac{\theta}{4}\frac{\partial V_1}{\partial \eta} H_1 H_1' \frac{\partial V_1}{\partial \eta}'\, dt \\ + \begin{bmatrix} z^2 \left(\dfrac{\partial \Xi}{\partial \eta}\right)' \\ 2z\Xi \end{bmatrix}' \begin{bmatrix} f_1 + g_1\alpha_1 + g_1 z \\ \bar{a}_2 + v \end{bmatrix} dt$$

$$+\frac{1}{2}\mathrm{Tr}\left(\left[\begin{array}{cc} z^2\dfrac{\partial^2\Xi}{\partial\eta^2} & 2z\left(\dfrac{\partial\Xi}{\partial\eta}\right)' \\ 2z\dfrac{\partial\Xi}{\partial\eta} & 2\Xi \end{array}\right]\left[\begin{array}{cc} H_1H_1' & H_1\bar{d}_2' \\ \bar{d}_2H_1' & (\bar{d}_{21}+z\bar{d}_{22})(\bar{d}_{21}+z\bar{d}_{22})' \end{array}\right]\right) dt,$$

where the function $\bar{\beta}(\eta, z)$ is to be determined shortly.

By multiplying out various terms above, we arrive at

$$\begin{aligned} d\bar{V}_2(\eta(t), z(t)) &= -y^2\,dt - q_1\,dt + \bar{\sigma}_2\,dw - \frac{\theta}{4}\bar{\sigma}_2\bar{\sigma}_2'\,dt + \Delta_1\,dt - \bar{\beta}z^2\,dt \\ &+\bar{\beta}z^2\,dt + \frac{\theta}{4}\left[\begin{array}{c} z^2\left(\dfrac{\partial\Xi}{\partial\eta}\right)' \\ 2z\Xi \end{array}\right]'\left[\begin{array}{c} H_1 \\ \bar{d}_2 \end{array}\right]H_1'\left(\frac{\partial V_1}{\partial\eta}\right)'\,dt \\ &+\frac{\theta}{4}\frac{\partial V_1}{\partial\eta}H_1\left[\begin{array}{c} H_1 \\ \bar{d}_2 \end{array}\right]'\left[\begin{array}{c} z^2\left(\dfrac{\partial\Xi}{\partial\eta}\right)' \\ 2z\Xi \end{array}\right]dt + \frac{\partial V_1}{\partial\eta}g_1 z\,dt \\ &+\frac{\theta}{4}\left[\begin{array}{c} z^2\left(\dfrac{\partial\Xi}{\partial\eta}\right)' \\ 2z\Xi \end{array}\right]'\left[\begin{array}{c} H_1' \\ \bar{d}_2 \end{array}\right]\left[\begin{array}{c} H_1' \\ \bar{d}_2 \end{array}\right]'\left[\begin{array}{c} z^2\left(\dfrac{\partial\Xi}{\partial\eta}\right)' \\ 2z\Xi \end{array}\right]dt \\ &+\left[\begin{array}{c} z^2\left(\dfrac{\partial\Xi}{\partial\eta}\right)' \\ 2z\Xi \end{array}\right]'\left[\begin{array}{c} f_1+g_1\alpha_1+g_1z \\ \bar{a}_2+v \end{array}\right]dt + \frac{1}{2}\mathrm{Tr}\left(z^2\frac{\partial^2\Xi}{\partial\eta^2}H_1'H_1\right) \\ &+\frac{1}{2}\mathrm{Tr}\left(4z\left(\frac{\partial\Xi}{\partial\eta}\right)'\bar{d}_2H_1 + 2\Xi(z\bar{d}_{21}\bar{d}_{22}' + z\bar{d}_{22}\bar{d}_{21}' + z^2\bar{d}_{22}\bar{d}_{22}')\right)dt \\ &+\Xi\bar{d}_{21}\bar{d}_{21}'\,dt. \end{aligned}$$

In the above SDE, we can select the control to cancel out all terms except those on the first and last lines. The control law that accomplishes this is the following:

$$\begin{aligned} \text{(3.9)}\quad \bar{\alpha}_{2b}(\eta, z) = -\frac{1}{2\Xi}\Bigg(&\bar{\beta}z + \frac{\theta}{4}\left[\begin{array}{c} z\left(\dfrac{\partial\Xi}{\partial\eta}\right)' \\ 2\Xi \end{array}\right]'\left[\begin{array}{c} H_1 \\ \bar{d}_2 \end{array}\right]H_1'\left(\frac{\partial V_1}{\partial\eta}\right)' \\ &+\frac{\theta}{4}\frac{\partial V_1}{\partial\eta}H_1\left[\begin{array}{c} H_1 \\ \bar{d}_2 \end{array}\right]'\left[\begin{array}{c} z\left(\dfrac{\partial\Xi}{\partial\eta}\right)' \\ 2\Xi \end{array}\right] + \frac{\partial V_1}{\partial\eta}g_1 \\ &+\frac{\theta}{4}\left[\begin{array}{c} z\left(\dfrac{\partial\Xi}{\partial\eta}\right)' \\ 2\Xi \end{array}\right]'\left[\begin{array}{c} H_1 \\ \bar{d}_2 \end{array}\right]\left[\begin{array}{c} H_1 \\ \bar{d}_2 \end{array}\right]'\left[\begin{array}{c} z^2\left(\dfrac{\partial\Xi}{\partial\eta}\right)' \\ 2z\Xi \end{array}\right] \\ &+\left[\begin{array}{c} z\left(\dfrac{\partial\Xi}{\partial\eta}\right)' \\ 2\Xi \end{array}\right]'\left[\begin{array}{c} f_1+g_1\alpha_1+g_1z \\ \bar{a}_2 \end{array}\right] \\ &+\frac{1}{2}\mathrm{Tr}\left(z\frac{\partial^2\Xi}{\partial\eta^2}H_1H_1' + 4\left(\frac{\partial\Xi}{\partial\eta}\right)'\bar{d}_2H_1' + 2\Xi(\bar{d}_{21}\bar{d}_{22}' + \bar{d}_{22}\bar{d}_{21}' + z\bar{d}_{22}\bar{d}_{22}')\right)\Bigg). \end{aligned}$$

Even though this controller leads to a guaranteed risk-sensitive cost design, in general it fails to preserve the equilibrium at $(\eta, \xi) = (0,0)$, except for the case when $H_1(0) = 0$ and $H_2(0,0) = 0$. To preserve the equilibrium at the origin, we remove the bias term from (3.9) to arrive at the modified controller

$$v = \bar{\alpha}_2(\eta, z) := \bar{\alpha}_{2b}(\eta, z) - \frac{\Xi(0)}{\Xi(\eta)} \bar{\alpha}_{2b}(0,0), \tag{3.10}$$

which, by its construction, preserves the equilibrium at $(\eta, \xi) = (0,0)$.

Using this control law, the differential of $\bar{V}_2$ becomes

$$\begin{aligned} d\bar{V}_2(\eta(t), z(t)) = -y^2\,dt - q_1\,dt + \bar{\sigma}_2\,dw - \frac{\theta}{4}\bar{\sigma}_2\bar{\sigma}_2'\,dt + \Delta_1\,dt - \bar{\beta} z^2\,dt \\ + \Xi \bar{d}_{21}\bar{d}_{21}'\,dt + 2z\Xi(0)\bar{\alpha}_{2b}(0,0)\,dt, \end{aligned}$$

where the last term is due to the removal of the constant bias in the controller.

To establish the desired HJB inequality (3.8), we will first pick the function Ξ and then β in the following steps. The function $\bar{d}_{21}(\eta)$ is further decomposed into constant and η-dependent parts:

$$\begin{aligned} \bar{d}_{21c} &:= \bar{d}_{21}(0), \\ \bar{d}_{21v}(\eta) &:= \bar{d}_{21}(\eta) - \bar{d}_{21c}. \end{aligned}$$

We choose Ξ such that for any positive R_2 and $\forall \eta \in \mathbb{R}^{n_1}$,

$$\Xi(\eta)\bar{d}_{21}(\eta)\bar{d}_{21}'(\eta) \le \frac{R_2}{2} + \lambda q_1(\eta), \tag{3.11}$$

and furthermore,

$$\Xi(\eta)\bar{d}_{21v}(\eta)\bar{d}_{21v}'(\eta) \le \lambda q_1(\eta). \tag{3.12}$$

One possible choice of the function Ξ can be made in the following two steps. For the first step, we will select a function Ξ_1 such that the inequality (3.12) is satisfied with Ξ_1 instead of Ξ. Since $\bar{d}_{21v}(\eta)$ is a smooth function that vanishes at $\eta = 0$, it can be written as

$$\bar{d}_{21v}(\eta) = \eta' \bar{D}_{21v}(\eta),$$

where $\bar{D}_{21v}$ is a smooth matrix-valued function of appropriate dimension. Let $\lambda_M = \lambda_{\max}(\bar{D}_{21v}(0)\bar{D}_{21v}'(0))$, where $\lambda_{\max}(M)$ denotes the maximum eigenvalue of a symmetric matrix M. Note that the following function is smooth:

$$\chi_1(\eta) = \frac{\lambda q_1(\eta)}{1 + \bar{d}_{21v}(\eta)\bar{d}_{21v}'(\eta)}.$$

Then, by the smoothness of $\bar{D}_{21v}$, there exists a constant $c_5 \in (0, c_2]$ such that

$$\frac{\lambda c_1 \eta'\eta}{(0.5 + \lambda_M)\eta'\eta} - \chi_1(\eta) \ge \frac{\lambda c_1}{1 + \lambda_M} \quad \forall\ 0 < |\eta| \le c_5$$

and

$$\bar{d}_{21v}(\eta)\bar{d}_{21v}'(\eta) = \eta'\bar{D}_{21v}(\eta)\bar{D}_{21v}'(\eta)\eta \le (0.5 + \lambda_M)\eta'\eta \quad \forall\ |\eta| \le c_5.$$

We define the following function:

$$\chi_2(\eta) = \begin{cases} \exp\left(\dfrac{\eta'\eta}{\eta'\eta - c_5^2}\right), & |\eta| < c_5, \\ 0, & |\eta| \geq c_5, \end{cases}$$

which is clearly smooth. Then, Ξ_1 can be chosen as

$$\Xi_1(\eta), = \frac{\lambda c_1}{1+\lambda_M}\chi_2(\eta) + \chi_1(\eta). \tag{3.13}$$

Ξ_1 is smooth and

$$\begin{aligned} \Xi_1(\eta)\bar{d}_{21v}(\eta)\bar{d}'_{21v}(\eta) &\leq \left(\frac{\lambda c_1}{1+\lambda_M} + \chi_1(\eta)\right)\bar{d}_{21v}(\eta)\bar{d}'_{21v}(\eta) \\ &\leq \lambda c_1 \eta'\eta \frac{\bar{d}_{21v}(\eta)\bar{d}'_{21v}(\eta)}{(0.5+\lambda_M)\eta'\eta} \leq \lambda q_1(\eta) \qquad \forall 0 < |\eta| < c_5 \end{aligned}$$

and

$$\Xi_1(\eta)\bar{d}_{21v}(\eta)\bar{d}'_{21v}(\eta) \leq \chi_1(\eta)\bar{d}_{21v}(\eta)\bar{d}'_{21v}(\eta) \leq \lambda q_1(\eta) \qquad \forall |\eta| \geq c_5.$$

Furthermore, $\Xi_1(\eta) > 0$, $\forall \eta \in \mathbb{R}^{n_1}$. Therefore, we have

$$\Xi_1(\eta)\bar{d}_{21v}(\eta)\bar{d}'_{21v}(\eta) \leq \lambda q_1(\eta) \qquad \forall \eta \in \mathbb{R}^{n_1}.$$

For the second step, we will select the function Ξ based on Ξ_1 derived above. Let $C > 0$ be a constant such that

$$C\left(\bar{d}_{21c}\bar{d}'_{21c} + \sqrt{\bar{d}_{21c}\bar{d}'_{21c}}\right) \leq \frac{R_2}{2}.$$

Given the above choice of the function Ξ_1, we can then select the function Ξ as

$$\Xi(\eta) = \frac{C\Xi_1(\eta)}{\sqrt{\Xi_1^2(\eta) + C^2\left(1+\sqrt{\bar{d}_{21c}\bar{d}'_{21c}}\right)^2}}. \tag{3.14}$$

It is clear that Ξ is smooth, and it satisfies the inequalities (3.11) and (3.12) since

$$\begin{aligned} &\Xi(\eta) < \frac{\Xi_1(\eta)}{1+\sqrt{\bar{d}_{21c}\bar{d}'_{21c}}}, \qquad \Xi(\eta) < C, \qquad \Xi(\eta) > 0 \qquad \forall \eta \in \mathbb{R}^{n_1}, \\ &\bar{d}_{21}(\eta)\bar{d}'_{21}(\eta) = \bar{d}_{21c}\bar{d}'_{21c} + 2\bar{d}_{21c}\bar{d}'_{21v}(\eta) + \bar{d}_{21v}(\eta)\bar{d}'_{21v}(\eta) \\ &\qquad \leq \bar{d}_{21c}\bar{d}'_{21c} + \frac{\bar{d}_{21c}\bar{d}'_{21c}}{\sqrt{\bar{d}_{21c}\bar{d}'_{21c}}} + \bar{d}_{21v}(\eta)\bar{d}'_{21v}(\eta)\sqrt{\bar{d}_{21c}\bar{d}'_{21c}} + \bar{d}_{21v}(\eta)\bar{d}'_{21v}(\eta), \\ &\Xi(\eta)\bar{d}_{21}(\eta)\bar{d}'_{21}(\eta) \leq \Xi(\eta)\left(\bar{d}_{21c}\bar{d}'_{21c} + \sqrt{\bar{d}_{21c}\bar{d}'_{21c}}\right) \\ &\qquad +\Xi(\eta)\bar{d}_{21v}(\eta)\bar{d}'_{21v}(\eta), \left(1+\sqrt{\bar{d}_{21c}\bar{d}'_{21c}}\right) < \frac{R_2}{2} + \lambda q_1(\eta) \qquad \forall \eta \in \mathbb{R}^{n_1}. \end{aligned}$$

After fixing the choice of the function Ξ, we next choose the positive function β to be

$$\bar{\beta}(\eta, z) = \frac{2\Xi^2(0)\bar{\alpha}_{2b}^2(0,0)}{R_2} + \bar{\beta}_a(\eta, z), \tag{3.15}$$

$$\bar{\beta}_a(\eta, z) \geq c_6\Xi(\eta) + c_7, \tag{3.16}$$

where $c_6 > 0$ and $c_7 > 0$ are positive constants. Then, we have

$$2z\Xi(0)\bar{\alpha}_{2b}(0,0) \leq \frac{R_2}{2} + \bar{\beta}(\eta,z)z^2 - \bar{\beta}_a(\eta,z)z^2. \tag{3.17}$$

As a result of these steps,

$$\begin{aligned} d\bar{V}_2(\eta(t),z(t)) &= -y^2\,dt - q_1\,dt + \bar{\sigma}_2\,dw - \frac{\theta}{4}\bar{\sigma}_2\bar{\sigma}_2'\,dt + \Delta_1\,dt - \bar{\beta}z^2\,dt \\ &\quad + \Xi\bar{d}_{21}\bar{d}_{21}'\,dt + 2\Xi(0)\bar{\alpha}_{2b}(0,0)z\,dt \\ &=: -y^2\,dt - (1-\lambda)q_1\,dt - q_2\,dt + \bar{\sigma}_2\,dw - \frac{\theta}{4}\bar{\sigma}_2\bar{\sigma}_2'\,dt + \Delta_1\,dt + R_2\,dt, \end{aligned}$$

where q_2 is defined by

$$\begin{aligned} \bar{q}_2(\eta,z) &:= \lambda q_1(\eta) + \bar{\beta}(\eta,z)z^2 + R_2 - \Xi(\eta)\bar{d}_{21}(\eta)\bar{d}_{21}'(\eta) - 2\Xi(0)\bar{\alpha}_{2b}(0,0)z \\ &\geq \bar{\beta}_a(\eta,z)z^2. \end{aligned}$$

By Proposition 2.6, we conclude that the function V_2 satisfies the HJB inequality (3.8).

REMARK 3.1. *Consider the special case $H_1(0) = 0$ and $H_2(0,0) = 0$. In this case, $\bar{d}_{21c} = 0$. Therefore, any choice of the function Ξ that satisfies the inequality* (3.12) *satisfies the inequality* (3.11) *with $R_2 = 0$. Fix such a choice of Ξ. In this special case, we observe that $\bar{\alpha}_{2b}(0,0) = 0$. This implies that the choice for the function $\bar{\beta}$*

$$\bar{\beta}(\eta,z) = \bar{\beta}_a(\eta,z) \tag{3.18}$$

leads to satisfaction of the inequality (3.17) *with $R_2 = 0$. Hence, in this special case, the above choices of functions Ξ and $\bar{\beta}$ lead to a controller design with $R_2 = 0.0$*

We next check the satisfaction of the control law α_2 and value function V_2 with respect to the four induction hypotheses made at the beginning of this section.

1. The control law (3.10) is smooth and meets the equilibrium condition $\alpha_2(0,0) = 0$ by construction.

2. The function V_2 is smooth and satisfies the HJB inequality (3.8).

3. The inequality

$$(1-\lambda)q_1(\eta) + q_2(\eta,\xi) \geq (1-\lambda)c_1\eta'\eta + c_7z^2 \geq c_1'\left|\left[\begin{array}{c}\eta\\ \xi\end{array}\right]\right|^2 \quad \forall\left|\left[\begin{array}{c}\eta\\ \xi\end{array}\right]\right| \leq c_2' \tag{3.19}$$

holds for some constants $c_1' > 0$ and $c_2' > 0$, and the left-hand side is a smooth function. We also have the following inequality:

$$\bar{V}_2(\eta,z) \leq c_3q_1(\eta) + c_4 + \bar{\beta}_a(\eta,z)z^2 \leq c_3'((1-\lambda)q_1(\eta) + q_2(\eta,z)) + c_4, \tag{3.20}$$

which holds $\forall(\eta,z) \in \mathbb{R}^{n_1+1}$, for some $c_3' > 0$, because of (3.16).

4. In the special case of $H_1(0) = 0$ and $H_2(0,0) = 0$, by the induction hypothesis, there exist a control law α_1 and a corresponding value function V_1 to guarantee a desired long-term average cost $R_1 = 0$ for the η dynamics. By Remark 3.1, we have found the design to yield $R_2 = 0$. Hence the pair α_2 and V_2 are smooth and guarantee a zero long-term average cost (i.e., $R_1 + R_2 = 0$) for the two-level problem (3.1). Furthermore, the guaranteed incremental cost function for the two-level problem (3.1) is the smooth function $(1-\lambda)q_1 + q_2$.

Hence, the control law α_2 and value function V_2 meet all four induction hypotheses.

This completes the guaranteed cost design for this two-level problem. We summarize these findings in the following lemma.

LEMMA 3.1. *Consider the SDE equation* (3.1), *and assume the following.*

(i) *Given an arbitrary* $R_1 > 0$, *there exists a smooth positive definite and radially unbounded value function* $V_1(\eta)$ *and a smooth fictitious control law* $\alpha_1(\eta)$ *(for* ξ*), such that* V_1 *satisfies the HJB inequality* (3.2), *as well as the bounds* (3.3) *and* (3.4) *with a smooth incremental cost function* $q_1(\eta)$, *and* $\alpha_1(0) = 0$.

(ii) *When* $H_1(0) = 0$, *the guaranteed long-term average cost level* R_1 *can be chosen to be* 0.

Then, the following statements hold.

1. *There exist a smooth value function* $V_2(\eta, \xi)$ *and a smooth control law* $\alpha_2(\eta, \xi)$ *as given by* (3.6) *and* (3.10), *respectively, where the former satisfies the HJB inequality* (3.8) *with the additional positive constant* R_2.
2. *The weighting function* $(1-\lambda)q_1(\eta) + q_2(\eta, \xi)$ *satisfies the bounds* (3.19) *and* (3.20), *and is smooth.*
3. *The control law* $v = \alpha_2(\eta, \xi)$ *has the property* $\alpha_2(0,0) = 0$, *and under it the closed-loop system trajectory is bounded in probability.*
4. *If* $H_1(0) = 0$ *and* $H_2(0,0) = 0$, *the constant* $R_1 + R_2$ *can be chosen as zero, which further implies that the closed-loop system is asymptotically stable in the large.*

4. Controller design for the full-order system. By using the backstepping Lemma 3.1 that we have just derived, we can design a feedback controller that leads to a guaranteed risk-sensitive cost R_c, for an arbitrary positive R_c and risk-sensitivity parameter θ. The design follows n steps of integrator backstepping. As suggested by Lemma 3.1, at each step of the backstepping one has to accommodate some positive quadratic variation intrinsic to that particular step. Accordingly, we first split the desired risk-sensitive cost R_c into n pieces, each of which is positive. When $h_1(0) = \cdots = h_n(0, \ldots, 0) = 0$, the control design can actually guarantee a zero long-term average, i.e., $R_c = 0$. Let

$$R_c = R_1 + \cdots + R_n,$$

where $R_i > 0$, $i = 1, \ldots, n$. The controller design follows n steps of integrator backstepping, as elucidated below:

Step 1: Here, we are dealing with a scalar system:

$$dx_1 = (f_1(x_1) + x_2)\,dt + h_1'(x_1)\,dw.$$

The controller to be designed is to guarantee a risk-sensitive cost R_1.

For this purpose, we select the value function V_1 as

$$V_1(x_1) = \frac{1}{1 + 2h_{1c}'h_{1c}/R_1} x_1^2 =: \Xi_1 x_1^2,$$

where h_{1c} is the constant part of the function h_1:

$$h_1(x_1) = h_{1c} + h_{1v}(x_1)x_1.$$

Then, using Itô's rule for $V_1(x_1(t))$, we obtain the following relationship:

$$dV_1(x_1(t)) = 2\Xi_1 x_1(f_1(x_1) + x_2)\,dt + 2\Xi_1 x_1 h_1'(x_1)\,dw + \mathrm{Tr}(\Xi_1 h_1'(x_1)h_1(x_1))\,dt.$$

Choose the virtual control input to cancel out all the terms that involve a multiplier of x_1:

$$\alpha_{1b}(x_1) = -\frac{1}{2\Xi_1}x_1 - \frac{1}{2\Xi_1}\beta x_1 - f_1(x_1) - \frac{\theta\Xi_1}{2}h_1'(x_1)h_1(x_1)x_1$$
$$-\frac{1}{2}h_{1v}'(x_1)h_{1v}(x_1)x_1 - h_{1c}'h_{1v}(x_1),$$

where β is a constant to be chosen later.

To preserve the equilibrium at the origin, we define the control law as follows:

$$x_2 = \alpha_{1b}(x_1) - \alpha_{1b}(0) =: \alpha_1(x_1), \tag{4.1}$$

This choice of the virtual control input leads to

$$dV_1(x_1(t)) = 2\Xi_1 x_1 h_1'\,dw - (x_1^2 + \beta x_1^2 + \theta\Xi_1^2 h_1' h_1 x_1^2 + \Xi_1 h_{1c}' h_{1c} + 2\Xi_1\alpha_{1b}(0)x_1)\,dt.$$

Choose

$$\beta = \frac{4\Xi_1^2\alpha_{1b}^2(0)}{R_1} + c_2,$$

where $c_2 > 0$. Then, V_1 satisfies the HJB inequality (2.11) for the x_1 subsystem under α_1 with guaranteed cost R_1:

$$\frac{\partial V_1}{\partial x_1}(f_1 + \alpha_1) + \frac{\theta}{4}\frac{\partial V_1}{\partial x_1}h_1'h_1\left(\frac{\partial V_1}{\partial x_1}\right)' + x_1^2 + \frac{1}{2}\beta x_1^2 + \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 V_1}{\partial x_1^2}h_1 h_1'\right) \le R_1.$$

It is easy to check to see that all four induction hypotheses are satisfied. This completes the first step of the integrator backstepping.

For Steps 2, 3, etc., we can repeatedly apply the backstepping lemma, Lemma 3.1. At a typical step i, we can design the risk-sensitive controller with the increase in the long-term average cost bounded by R_i. We then continue this process until the second to last step $n-1$, where we design the virtual control input

$$x_n = \alpha_{n-1}(x_1, \ldots, x_{n-1}).$$

At Step n, we let $v = bu$, and use the construction of Lemma 3.1 to lead to a smooth control law:

$$v = \alpha_n(x_1, \ldots, x_n).$$

Under the working assumption, A1, we finally obtain

$$u = \frac{1}{b(x_1, \ldots, x_n)}\,\alpha_n(x_1, \ldots, x_n). \tag{4.2}$$

This completes the construction of a risk-sensitive controller, guaranteeing any desired level of positive long-term average cost. This construction is now made precise in the following theorem.

THEOREM 4.1. *Consider the strict-feedback stochastic nonlinear system described by the SDE* (2.1), *satisfying Assumption* A1. *For any risk-sensitivity parameter* $\theta > 0$ *and any desired long-term average cost level* $R_c > 0$, *there exists a smooth nonlinear feedback control* (4.2) *that achieves the level* R_c. *Such a controller can be constructed* *using the integrator backstepping procedure and by a repeated application of Lemma* 3.1. *The resulting closed-loop system trajectory is bounded in probability.*

If the functions h_i, $i = 1, \dots, n$, vanish at $x = 0$, then the controller (4.2) *achieves a zero long-term average cost for the SDE* (2.1). *In this case, the closed-loop system is asymptotically stable in the large, in addition to being bounded in probability.*

Proof. The result is an immediate application of Theorem 2.5, in view of the backstepping construction outlined above. Since the existence (and construction) of function V satisfying the HJB inequality is implied by the construction presented before the statement of the theorem, the boundedness in probability follows immediately.

When h_i vanishes at $x = 0$, the construction implies that the constant R_c can be chosen to be zero, by Remark 3.1. Since the weighting function is chosen to be positive definite, by Theorem 2.5 the system is asymptotically stable in the large, in addition to being bounded in probability. □

5. Two special cases.

In this section, we discuss some implications of the results presented above as the design parameters approach specific limits. In particular, we are interested in two types of limiting processes. The first is known as the risk-neutral limit, with $\theta \downarrow 0$. The other process is known as the large deviation limit, where the noise intensity decreases to zero.

5.1. Risk-neutral case.

If $\theta \downarrow 0$, the limiting problem is the one with a risk-neutral cost function:

$$J_0(\mu) = \limsup_{t_f \to \infty} \frac{1}{t_f} \boldsymbol{E} \left\{ \int_0^{t_f} (y^2(t) + l(x(t)))\, dt \right\} \leq R_c, \tag{5.1}$$

where R_c is again the desired long-term average cost bound.

The control design procedure presented in the previous two sections equally applies here, leading to a controller

$$u = \mu(x)$$

and a value function $V(x)$ such that

$$\frac{\partial V}{\partial x} f^u(x) + \frac{1}{2} \mathrm{Tr} \left(\frac{\partial^2 V}{\partial x^2} H(x) H'(x) \right) + y^2 + l(x) = \Delta(x) \leq R_c, \tag{5.2}$$

where

$$f^u(x) = \begin{bmatrix} f_1(x_1) + x_2 \\ \vdots \\ f_{n-1}(x_1, \dots, x_{n-1}) + x_n \\ f_n(x_1, \dots, x_n) + b(x_1, \dots, x_n)\mu(x) \end{bmatrix} \quad H(x) = \begin{bmatrix} h_1'(x_1) \\ \vdots \\ h_n'(x_1, \dots, x_n) \end{bmatrix}.$$

This inequality immediately implies that the long-term average cost for the risk-neutral cost function (5.1) is bounded by R_c.

Therefore, we have the following corollary to Theorem 4.1.

Corollary 5.1. *Consider the strict-feedback stochastic nonlinear system* (2.1) *satisfying Assumption* A1, *along with the risk-neutral cost function* (5.1). *For any desired long-term average cost level $R_c > 0$, there exists a corresponding smooth nonlinear feedback control that guarantees it. Such a controller can be constructed using the integrator backstepping procedure presented in the previous two sections with the* *risk-sensitivity parameter set at $\theta = 0$. The resulting closed-loop system trajectories are bounded in probability.*

If the functions h_i, $i = 1, \ldots, n$, vanish at $x = 0$, then the controller μ can be chosen to guarantee $R_c = 0$. In this case, the closed-loop system is asymptotically stable in the large, in addition to being bounded in probability.

5.2. Large deviation limit. Another limiting scenario is that in which the noise intensity in the system dynamics asymptotically vanishes at the rate of ϵ, for some small positive parameter ϵ, while the cost function heavily penalizes any deviation of the output from zero at the rate of $1/\epsilon$. Letting $\theta = \frac{1}{\gamma^2\epsilon^2}$, for some positive design parameter γ (whose role and significance will become clear shortly), we now have the following problem formulation:

$$\text{(5.3a)} \qquad dx_1 = (x_2 + f_1(x_1))\,dt + \epsilon\tilde{h}_1'(x_1)\,dw,$$

$$\vdots \qquad\qquad \vdots$$

$$\text{(5.3b)} \qquad dx_{n-1} = (x_n + f_{n-1}(x_1, \ldots, x_{n-1}))\,dt + \epsilon\tilde{h}_{n-1}'(x_1, \ldots, x_{n-1})\,dw,$$

$$\text{(5.3c)} \qquad dx_n = (f_n(x_1, \ldots, x_n) + b(x_1, \ldots, x_n)u)\,dt + \epsilon\tilde{h}_n'(x_1, \ldots, x_n)\,dw,$$

$$\text{(5.3d)} \qquad y = x_1,$$

$$\text{(5.4)} \qquad J_\theta(\mu) = \limsup_{t_f \to \infty} \frac{2}{\theta t_f} \ln \left\{ \boldsymbol{E} \left\{ \exp\left[\frac{\theta}{2} \left(\int_0^{t_f} (y^2 + l(x(t)))\,dt \right) \right] \right\} \right\} \le R_c.$$

It is a well-known fact in the risk-sensitive optimal control literature (see [1] or [25]) that the large deviation limit, as $\epsilon \to 0$, of the optimal risk-sensitive controller for this problem leads exactly to the H^∞ central controller with disturbance attenuation parameter γ, provided that the solution to either one exists. For the problem at hand, the corresponding H^∞ control problem involves the state dynamics

$$\text{(5.5a)} \qquad \dot{x}_1 = x_2 + f_1(x_1) + \tilde{h}_1'(x_1)w,$$

$$\vdots \qquad\qquad \vdots$$

$$\text{(5.5b)} \qquad \dot{x}_{n-1} = x_n + f_{n-1}(x_1, \ldots, x_{n-1}) + \tilde{h}_{n-1}'(x_1, \ldots, x_{n-1})w,$$

$$\text{(5.5c)} \qquad \dot{x}_n = f_n(x_1, \ldots, x_n) + b(x_1, \ldots, x_n)u + \tilde{h}_n'(x_1, \ldots, x_n)w,$$

$$\text{(5.5d)} \qquad y = x_1$$

and the following worst-case (game) cost function:

$$\text{(5.6)} \qquad J_\gamma(\mu, \nu) = \int_0^\infty (y^2(t) + l(x(t)) - \gamma^2|w(t)|^2)\,dt,$$

where w is generated by some opponent through $w(t) = \nu(t, x(t))$.

This is precisely the problem that was addressed in [23]. The interesting question to ask here is whether the backstepping construction of risk-sensitive controllers for (5.3), (5.4), also converges to the backstepping construction for the H^∞ controller for (5.5), (5.6) presented in [23]. Since both the design procedures presented here and the ones in [23] are suboptimal schemes, the convergence cannot be guaranteed by the well-established large deviation limit for risk-sensitive optimal control problems.

It turns out that the answer to this question is still yes. We examine again the single-stage backstepping design presented in section 3. The difference between the recursive construction presented in [23] and that of this paper is as follows. First of all, the choice of Ξ is solely determined by the inequalities (3.11) and (3.12). For any fixed compact set K, Ξ can be chosen to be constant, say 1/2, on this compact set for sufficiently small $\epsilon > 0$, such that, $\forall \eta \in K$,

$$\mathrm{Tr}\left(\Xi(\eta)\bar{d}_{21}(\eta)\bar{d}_{21}'(\eta)\right) \leq \frac{1}{\epsilon^2}\left(\frac{R_2}{2} + \lambda q_1(\eta)\right)$$

and

$$\mathrm{Tr}\left(\Xi(\eta)\bar{d}_{21v}(\eta)\bar{d}_{21v}'(\eta)\right) \leq \frac{\lambda c_3 \eta'\eta}{\epsilon^2}.$$

This compact set expands to the entire state space as $\epsilon \downarrow 0$. On this compact set, the added value function in this case is exactly the same as that for the H^∞ problem. Consider the controller (3.9); all the terms on the first three lines are exactly the same as in the controller of the H^∞ problem. The function $\bar{a}_2$ is within $O(\epsilon^2)$ of its counterpart in the H^∞ problem. The terms on the last line are of $O(\epsilon^2)$. Therefore, on this compact set, the risk-sensitive design provides a $O(\epsilon^2)$ approximation to the H^∞ solution.

Hence, we can state the following remark to capture our observation.

Remark 5.1. *Consider the strict-feedback stochastic nonlinear system* (5.3) *satisfying Assumption* A1 *with the risk-sensitive cost function* (5.4). *For any desired long-term average cost level $R_c > 0$, there exists a corresponding nonlinear feedback control that guarantees it. Such a controller approximates the backstepping controller for the nonlinear H^∞ control problem* [23] *up to $O(\epsilon^2)$ on any given compact set $K \subset \mathbb{R}^n$.*

6. An extension.

We identify and present here an immediate (natural) extension of the results presented heretofore to a more general class of nonlinear systems, with nontrivial zero dynamics. Consider the class of nonlinear systems described by the following SDE:

$$\begin{aligned}
&(6.1\text{a}) \quad d\phi = f_0(\phi, x_1)\,dt + H_0(\phi, x_1)\,dw,\\
&(6.1\text{b}) \quad dx_1 = (b_1(\phi, x_1)x_2 + f_1(\phi, x_1))\,dt + h_1'(\phi, x_1)\,dw,\\
&\qquad\quad \vdots \qquad\quad \vdots\\
&(6.1\text{c})\, dx_{n-1} = (b_{n-1}(\phi, x_1, \ldots, x_{n-1})x_n + f_{n-1}(\phi, x_1, \ldots, x_{n-1}))\,dt\\
&\qquad\qquad\qquad + h_{n-1}'(\phi, x_1, \ldots, x_{n-1})\,dw,\\
&(6.1\text{d}) \quad dx_n = (f_n(\phi, x_1, \ldots, x_n) + b(\phi, x_1, \ldots, x_n)u)\,dt + h_n'(\phi, x_1, \ldots, x_n)\,dw,\\
&(6.1\text{e}) \quad y = p(\phi, x_1)
\end{aligned}$$

where ϕ is an n_0-dimensional state vector, which can be viewed as the state of the zero dynamics of the system (see Chapter 9 of [10]). We note that these ϕ dynamics are affected only by x_1 and not by the remaining components of the state x. Furthermore, the virtual control input variables $x_2, \ldots, x_n$ are premultiplied by nonlinear functions $b_1, \ldots, b_{n-1}$. For this class of nonlinear stochastic systems, we introduce the following two assumptions in addition to A1.

Assumption A2. The nonlinear functions f_0, H_0, b_i, and $1/b_i$ are $\mathcal{C}^\infty$ in all their arguments, $i = 1, \ldots, n-1$. The output function p is smooth. □

Assumption A3. Given a risk-sensitivity parameter $\theta_0 > 0$, there exists a smooth virtual control law $\alpha_0(\phi)$, with $\alpha_0(0) = 0$, and a positive definite and radially unbounded smooth value function $V_0(\phi)$, such that the following PDE is satisfied for some positive definite function $q_0(\phi)$:

$$\begin{aligned}(6.2)\ &\frac{\partial V_0}{\partial \phi} f_0(\phi, \alpha_0(\phi)) + \frac{\theta_0}{4}\frac{\partial V_0}{\partial \phi} H_0(\phi, \alpha_0(\phi)) H_0'(\phi, \alpha_0(\phi)) \left(\frac{\partial V_0}{\partial \phi}\right)' \\ &+ \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 V_0}{\partial \phi^2} H_0(\phi, \alpha_0(\phi)) H_0'(\phi, \alpha_0(\phi))\right) + p^2(\phi, \alpha_0(\phi)) + q_0(\phi) = \Delta_0(\phi) \leq R_0.\end{aligned}$$

Furthermore, $q_0(\phi)$ is smooth and satisfies the assumptions of induction for the backstepping procedure, and the following growth conditions:

$$q_0(\phi) \geq c_1 \phi'\phi \qquad \forall |\phi| \leq c_2 \tag{6.3}$$

and

$$V_0(\phi) \leq c_3\, q_0(\phi) + c_4 \qquad \forall \phi \in \mathbb{R}^{n_0} \tag{6.4}$$

for some constants $c_1 > 0$, $c_2 > 0$, $c_3 > 0$, and $c_4 > 0$. □

We should note that A2 is quite a standard assumption for this class of nonlinear system, ensuring that the state variables $x_1, \ldots, x_n$ are completely controllable; A3, on the other hand, is a strong stabilizability condition on the ϕ dynamics, ensuring that there is a virtual control law for x_1 that guarantees a long-term average value of R_0 for the risk-sensitive cost function

$$\limsup_{t_f \to \infty} \frac{2}{\theta_0 t_f} \ln\left\{\boldsymbol{E}\left\{\exp\left[\frac{\theta_0}{2}\left(\int_0^{t_f} y^2(t)\, dt\right)\right]\right\}\right\}$$

for a given value of θ_0. This is stated as an assumption, because for a given $\theta_0 > 0$ there may not exist a control law (for x_1) to return an arbitrarily small value for the risk-sensitive cost function above. This is partly due to the appearance of both ϕ and x_1 as arguments of p.

Under these assumptions, it is possible to construct recursively a nonlinear feedback controller that achieves any given long-term average cost level $R_c > R_0$ for a given risk-sensitivity parameter $\theta \leq \theta_0$. We note that, as stated earlier, we may no longer be able to achieve an arbitrarily small long-term average cost for arbitrarily large values of the risk-sensitivity parameter. This performance limitation is solely due to the presence of the zero dynamics of the system.

This now brings us to the following theorem.

THEOREM 6.1. *Consider the class of stochastic nonlinear systems described by SDE* (6.1), *satisfying Assumptions* A1, A2, *and* A3. *For any risk-sensitivity parameter $\theta \leq \theta_0$ and any desired long-term average cost level $R_c > R_0$, there exists a nonlinear feedback control law μ, designed using the integrator backstepping procedure that achieves the level R_c. The resulting closed-loop system trajectory is bounded in probability.*

If $R_0 = 0$ is a feasible choice, and the nonlinear functions H_0, h_i, $i = 1, \ldots, n$, all vanish at the origin of the system, then the nonlinear controller constructed with $R_0 = 0$ achieves a zero long-term average cost for the SDE (6.1). *In this case, the closed-loop system is asymptotically stable in the large, in addition to being bounded in probability.*

Proof. We need only establish the result when $n = 1$. For the case when $n > 1$, the result follows directly from the recursive design procedure described in section 3.

Consider the following SDE:

$$\begin{aligned} d\phi &= f_0(\phi, x_1)\,dt + H_0(\phi, x_1)\,dw \\ dx_1 &= (b_1(\phi, x_1)u + f_1(\phi, x_1))\,dt + h_1'(\phi, x_1)\,dw \\ y &= p(\phi, x_1)\,. \end{aligned} \tag{6.5}$$

We will introduce the variable

$$v = b_1(\phi, x_1)u$$

and design a control for v, which, in turn, yields a control law for u, under the working Assumption A2.

By Assumption A3, the function V_0 satisfies the following HJB inequality for any $\theta \le \theta_0$:

$$\begin{aligned} &\frac{\partial V_0}{\partial \phi} f_0(\phi, \alpha_0(\phi)) + \frac{\theta}{4}\frac{\partial V_0}{\partial \phi} H_0(\phi, \alpha_0(\phi))H_0(\phi, \alpha_0(\phi))' \left(\frac{\partial V_0}{\partial \phi}\right)' \\ &+\frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 V_0}{\partial \phi^2} H_0(\phi, \alpha_0(\phi))H_0(\phi, \alpha_0(\phi))'\right) + p^2(\phi, \alpha_0(\phi)) + q_0(\phi) = \Delta_0(\phi) \le R_0. \end{aligned} \tag{6.6}$$

Given this, we introduce the transformed state variable

$$z = x_1 - \alpha_0(\phi).$$

Then, by Itô's rule of differentiation, we have

$$dz = (\bar{a}_1 + v)\,dt + \bar{d}_1\,dw,$$

where

$$\begin{aligned} \bar{a}_1(\phi, z) &= \bar{f}_1(\phi, z) - \frac{\partial \alpha_0}{\partial \phi}\bar{f}_0(\phi, z) - \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 \alpha_0}{\partial \phi^2}\bar{H}_0(\phi, z)\bar{H}_0'(\phi, z)\right), \\ \bar{d}_1(\phi, z) &= \bar{h}_1'(\phi, z) - \frac{\partial \alpha_0}{\partial \phi}\bar{H}_0(\phi, z). \end{aligned}$$

Introduce a smooth function $\bar{V}_1$ as a candidate value function

$$\bar{V}_1(\phi, z) = V_0(\phi) + \Xi(\phi)z^2, \tag{6.7}$$

where Ξ is a positive smooth function yet to be determined. We again set up the proper conditions to prove that $\bar{V}_1$ satisfies the desired HJB inequality by using the equivalent form of Proposition 2.6.

Define the smooth functions $\bar{f}_{01}$, $\bar{H}_{01}$, $\bar{d}_{11}$, and $\bar{d}_{12}$ as follows:

$$\begin{aligned} &f_0(\phi, x_1) = f_0(\phi, \alpha_0(\phi)) + \bar{f}_{01}(\phi, z)z, \qquad H_0(\phi, x_1) = H_0(\phi, \alpha_0(\phi)) + \bar{H}_{01}(\phi, z)z, \\ &\bar{d}_{11}(\phi) = \bar{d}_1(\phi, 0), \qquad \bar{d}_1(\phi, z) = \bar{d}_1(\phi, 0) + \bar{d}_{12}(\phi, z)z. \end{aligned}$$

The Itô differential of $\bar{V}_1(\phi(t), z(t))$ is given by

$$\begin{aligned} d\bar{V}_1(\phi(t), z(t)) = {}&\left(\frac{\partial V_0}{\partial \phi} f_0(\phi, \alpha_0(\phi)) + \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 V_0}{\partial \phi^2} H_0(\phi, \alpha_0(\phi))H_0'(\phi, \alpha_0(\phi))\right)\right) dt \\ &+z(\bar{m}(\phi, z) + 2\Xi(\phi)v)\,dt + \bar{\sigma}(\phi, z)\,dw + \mathrm{Tr}(\Xi(\phi)\bar{d}_{11}(\phi)\bar{d}_{11}'(\phi))\,dt, \end{aligned}$$

where

$$\begin{aligned}\bar{m}(\phi,z) &= \frac{\partial V_0}{\partial \phi}\bar{f}_{01} + \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 V_0}{\partial \phi^2}(2H_0\bar{H}_{01}' + z\bar{H}_{01}\bar{H}_{01}')\right) + z\frac{\partial \Xi}{\partial \phi}\bar{f}_0 + 2\Xi\bar{a}_1 \\ &\quad + \frac{1}{2}\mathrm{Tr}\left(z\frac{\partial^2 \Xi}{\partial \phi^2}\bar{H}_0\bar{H}_0' + 4\frac{\partial \Xi}{\partial \phi}\bar{H}_0\bar{d}_1' + 4\Xi\bar{d}_{11}\bar{d}_{12}' + 2\Xi\bar{d}_{12}\bar{d}_{12}'z\right), \\ \bar{\sigma}(\phi,z) &= \frac{\partial V_0}{\partial \phi}H_0 + z\bar{\sigma}_{11}, \\ \bar{\sigma}_{11}(\phi,z) &= \frac{\partial V_0}{\partial \phi}\bar{H}_{01} + z\frac{\partial \Xi}{\partial \phi}\bar{H}_0 + 2\Xi\bar{d}_1.\end{aligned}$$

Using the HJB inequality (6.6), we arrive at the following stochastic differential inequality:

$$\begin{aligned}d\bar{V}_1(\phi(t),z(t)) &= \left(-p^2 - q_0 + \Delta_0 - \frac{\theta}{4}\bar{\sigma}\bar{\sigma}'\right)dt + \bar{\sigma}\,dw \\ &\quad + z\left(\bar{m} + \frac{\theta}{4}\left(2\frac{\partial V_0}{\partial \phi}H_0\bar{\sigma}_{11}' + z\bar{\sigma}_{11}\bar{\sigma}_{11}'\right) + 2\Xi v\right)dt + \mathrm{Tr}(\Xi\bar{d}_{11}\bar{d}_{11}')\,dt.\end{aligned}$$

We next define a smooth nonlinear function $\bar{p}_1$ by

$$p(\phi,x_1) = p(\phi,\alpha_0(\phi)) + \bar{p}_1(\phi,z)z\,,$$

and introduce the control law

$$v = \bar{\alpha}_1(\phi,z) = \bar{\alpha}_{1b}(\phi,z) - \frac{\Xi(0)}{\Xi(\phi)}\bar{\alpha}_{1b}(0,0), \tag{6.8}$$

where

$$\bar{\alpha}_{1b}(\phi,z) = \frac{1}{2\Xi}\left(-2p\bar{p}_1 - \bar{p}_1^2 z - \bar{\beta}z - \bar{m} - \frac{\theta}{4}\left(2\frac{\partial V_0}{\partial \phi}H_0\bar{\sigma}_{11}' + z\bar{\sigma}_{11}\bar{\sigma}_{11}'\right)\right)$$

and $\bar{\beta}(\phi,z)$ is a positive nonlinear function to be determined shortly.

Under this control law, the Itô differential of $\bar{V}_1(\phi(t),z(t))$ satisfies

$$\begin{aligned}d\bar{V}_1(\phi(t),z(t)) &= \left(-y^2 - q_0 - \bar{\beta}z^2 + \Delta_0 - \frac{\theta}{4}\bar{\sigma}\bar{\sigma}'\right)dt + \bar{\sigma}\,dw + 2z\Xi(0)\bar{\alpha}_{1b}(0,0)\,dt \\ &\quad + \mathrm{Tr}(\Xi\bar{d}_{11}\bar{d}_{11}')\,dt.\end{aligned}$$

Now, following a selection process similar to that of section 3 for the functions Ξ and $\bar{\beta}$, we can show that

$$d\bar{V}_1(\phi(t),z(t)) \le \left(-y^2 - q_1 + R_1 - \frac{\theta}{4}\bar{\sigma}\bar{\sigma}'\right)dt + \bar{\sigma}\,dw$$

for any $R_1 > R_0$, and the nonlinear weighting function q_1 is smooth and satisfies the induction assumptions of section 3. This then completes the construction for this first step of the backstepping procedure.

When $n > 1$, we can repeatedly apply Lemma 3.1 to construct the nonlinear cost-bounding controller for the overall system.

This completes the proof of the theorem. $\square$

7. An example. In this section, we illustrate the design procedure presented in the main body of the paper by a second-order example and compare the resulting control design with that of [23] and the risk neutral case as $\theta \downarrow 0$. The second-order system is taken as

$$\begin{bmatrix} dx_1 \\ dx_2 \end{bmatrix} = \begin{bmatrix} x_1^2 + x_2 \\ 0 \end{bmatrix} dt + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u\, dt + \begin{bmatrix} 0 & 1 \\ 1 & x_1 \end{bmatrix} dw, \tag{7.1a}$$

$$y = x_1, \tag{7.1b}$$

where u is the control input and w is a two-dimensional standard Wiener process. For this SDE, we will consider two types of cost functions, namely, the risk-sensitive cost function and the risk-neutral cost function. We will take the risk-sensitivity parameter $\theta = 1$ and a desired guaranteed long-term average cost of $R_c = 4$. This corresponds to the following choice of design criterion for the risk-sensitive case:

$$J_{rs}(\mu) = \limsup_{t_f \to \infty} \frac{2}{t_f} \ln \left\{ \boldsymbol{E} \left\{ \exp \left[\frac{1}{2} \left(\int_0^{t_f} (y^2(t) + l(x(t)))\, dt \right) \right] \right\} \right\} \leq 4 \tag{7.2}$$

and the following one for the risk-neutral case:

$$J_{rn}(\mu) = \limsup_{t_f \to \infty} \frac{1}{t_f} \boldsymbol{E} \left\{ \int_0^{t_f} (y^2(t) + l(x(t)))\, dt \right\} \leq 4. \tag{7.3}$$

The corresponding nonlinear H^∞-control problem of [23] (see also (5.5) and (5.6) in subsection 5.2), on the other hand, is described by the following second-order dynamics:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_1^2 + x_2 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u + \begin{bmatrix} 0 & 1 \\ 1 & x_1 \end{bmatrix} w, \tag{7.4a}$$

$$y = x_1, \tag{7.4b}$$

where w is to be chosen by some opponent to disrupt the control objective. The associated (game-theoretic) cost function is

$$J_{rb}(\mu, \nu) = \int_0^\infty (y^2(t) + l(x(t)) - w'(t)w(t))\, dt. \tag{7.5}$$

Here we have picked $\epsilon = 1$ and $\gamma = 1$—a choice that is consistent with the specific choice $\theta = 1$ for the risk-sensitivity parameter, because of the relationship $\theta = \frac{1}{\epsilon^2\gamma^2}$.

Carrying out the risk-sensitive backstepping design procedure, we arrive at the following value function and control law:

$$V_{rs}(x) = 0.5x_1^2 + 0.0391251 \left(x_2 + 2.25x_1 + x_1^2\right)^2, \tag{7.6a}$$

$$\begin{aligned} \mu_{rs}(x) = {} & -4.93429x_2 - 19.9442x_1 - 7.02849x_1^2 - 2.26409x_1x_2 - 2.66023x_1^3 \\ & -0.176063x_1^2x_2 - 0.176063x_1^4. \end{aligned} \tag{7.6b}$$

An important observation to make here is that the constant multiplying the second term in the value function V_{rs} above is much smaller than that of the first term. This constant is determined by the bounds (3.11) and (3.12) for the recursive design.

Now, for the risk-neutral case, the counterparts of (7.6a) and (7.6b) are, respectively,

$$V_{rn}(x) = 0.5x_1^2 + 0.0406138 \left(x_2 + 2x_1 + x_1^2\right)^2, \tag{7.7a}$$

$$\mu_{rn}(x) = -4.47237x_2 - 17.2558x_1 - 4.47237x_1^2 - 2x_1x_2 - 2x_1^3. \tag{7.7b}$$

We observe that the assignment of the constant multipliers in the value function above is quite similar to that in the risk-sensitive case. The control law, however, is somewhat different; it does not include the high-order nonlinear terms that are part of the risk-sensitive design. This result corroborates the earlier statement that the risk neutral control design does not put weight on the higher (than second) order moments in the system output. Apart from the missing high-order nonlinear terms, the control design for this risk-neutral case is quite similar to that of the risk-sensitive case.

Finally, we design a control law for the worst-case disturbance attenuation problem (7.4a), (7.4b) and (7.5). In this case the value function and the control law are given as follows:

$$\text{(7.8a)} \quad V_{rb}(x) = 0.5x_1^2 + 0.5\left(x_2 + 1.75x_1 + x_1^2\right)^2,$$

$$\text{(7.8b)} \quad \mu_{rb}(x) = -3.26563x_2 - 4.52734x_1 - 9.35938x_1^2 - 2.25x_1^2x_2 - 4.62500x_1x_2 \\ -8.56250x_1^3 - 2.25x_1^4.$$

We observe that both the value function and the controller are quite different from their counterparts in the risk-sensitive and risk-neutral cases. Here, the gains of the linear control terms are relatively small compared with those of their stochastic counterparts, yet the gains on high-order nonlinear terms are significantly larger than those of the stochastic controllers.

The mesh plots of the value functions and the control laws for the three designs are depicted in Figures 7.1–7.3. Comparing these figures, we observe that both stochastic designs lead to smaller value functions than the worst-case design. The stochastic control laws are much softer than the worst-case control laws. In particular, in the region of the state space $x_1 < 0$, the stochastic control laws are much softer than the worst-case (nonlinear H^∞) control law.

We simulated the closed-loop system under these three controllers subject to two different types of disturbance inputs. The simulations were carried out using the Matlab Simulink software package. We chose the disturbance inputs to be sampled Gaussian white noise inputs with a sample rate of 10 Hz. Each of the noise channels had power 1. The simulation results are plotted in Figures 7.4–7.6. We observe from these figures that the stochastic control designs lead to smaller system outputs than the worst-case controller. A simple calculation yields that the $\mathcal{L}_2$ norm of the output for the risk-sensitive control design is the smallest among the three, which is 5.0058. Second best is the risk-neutral controller, whose output norm is 6.0840. The worst-case controller yields a norm of 7.0492 for its output. Regarding the control effort, we observe that the stochastic controllers use more control effort when the system states are small in norm. Yet the worst-case controller leads to larger peak control actions. A simple calculation yields that the $\mathcal{L}_2$ norms of the stochastic controllers are bounded by twice the amount of the worst-case controller.

Finally, we simulated the system response under the following sinusoidal disturbance inputs:

$$w_1(t) = 8\,\sin(t), \qquad w_2(t) = 8\,\sin(2t+1).$$

We tried various levels for the magnitudes of the sine waves. For smaller-magnitude sine waves, the stochastic controllers showed better performance than the worst-case controller. When the magnitudes were fixed at 8, the performances of the two stochastic controllers were very different. These simulation results are plotted in Figures 7.7–7.9. We observe from these figures that the risk-sensitive controller again leads to the

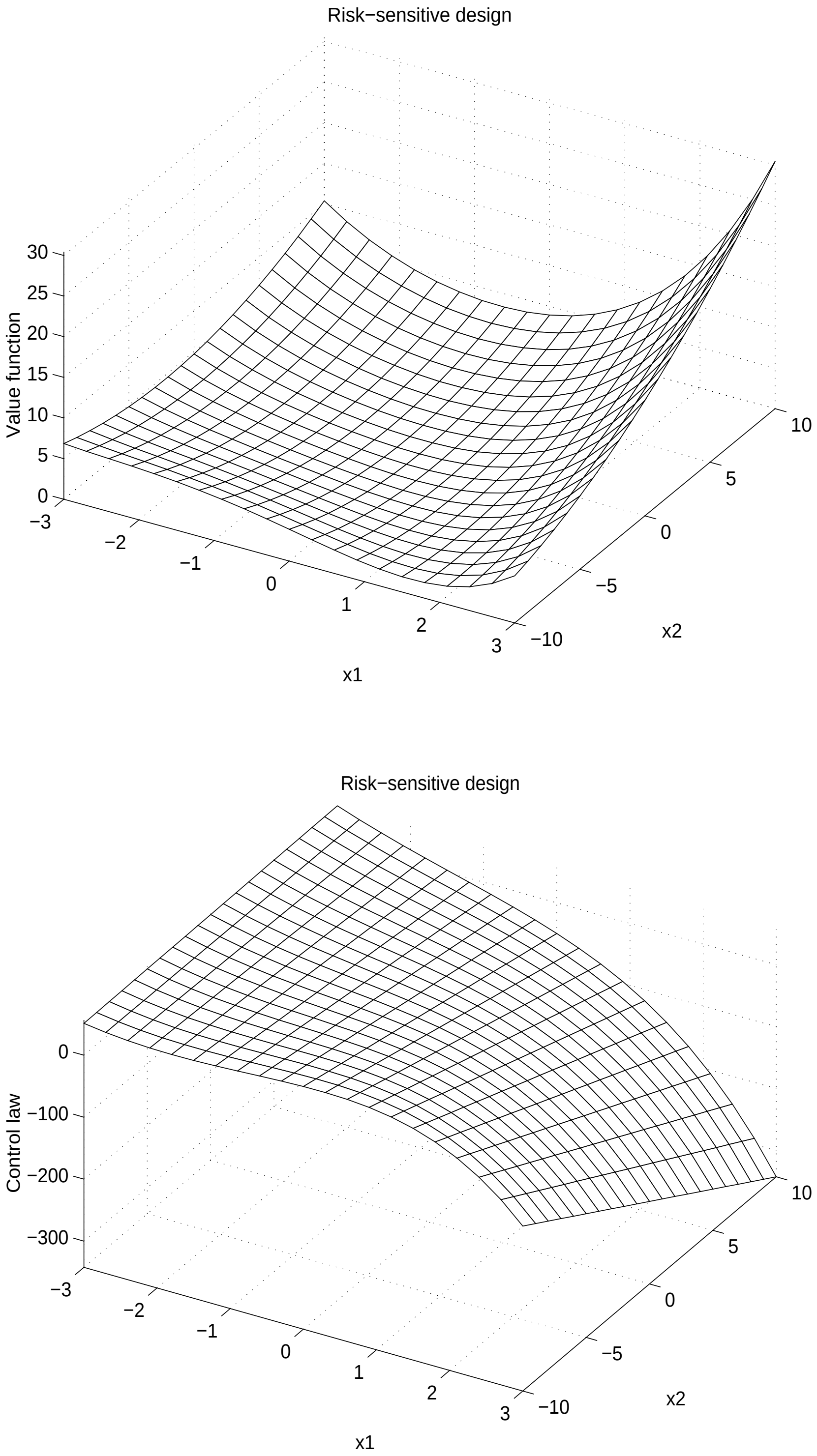


Fig. 7.1. *Value function and control law for risk-sensitive design.*

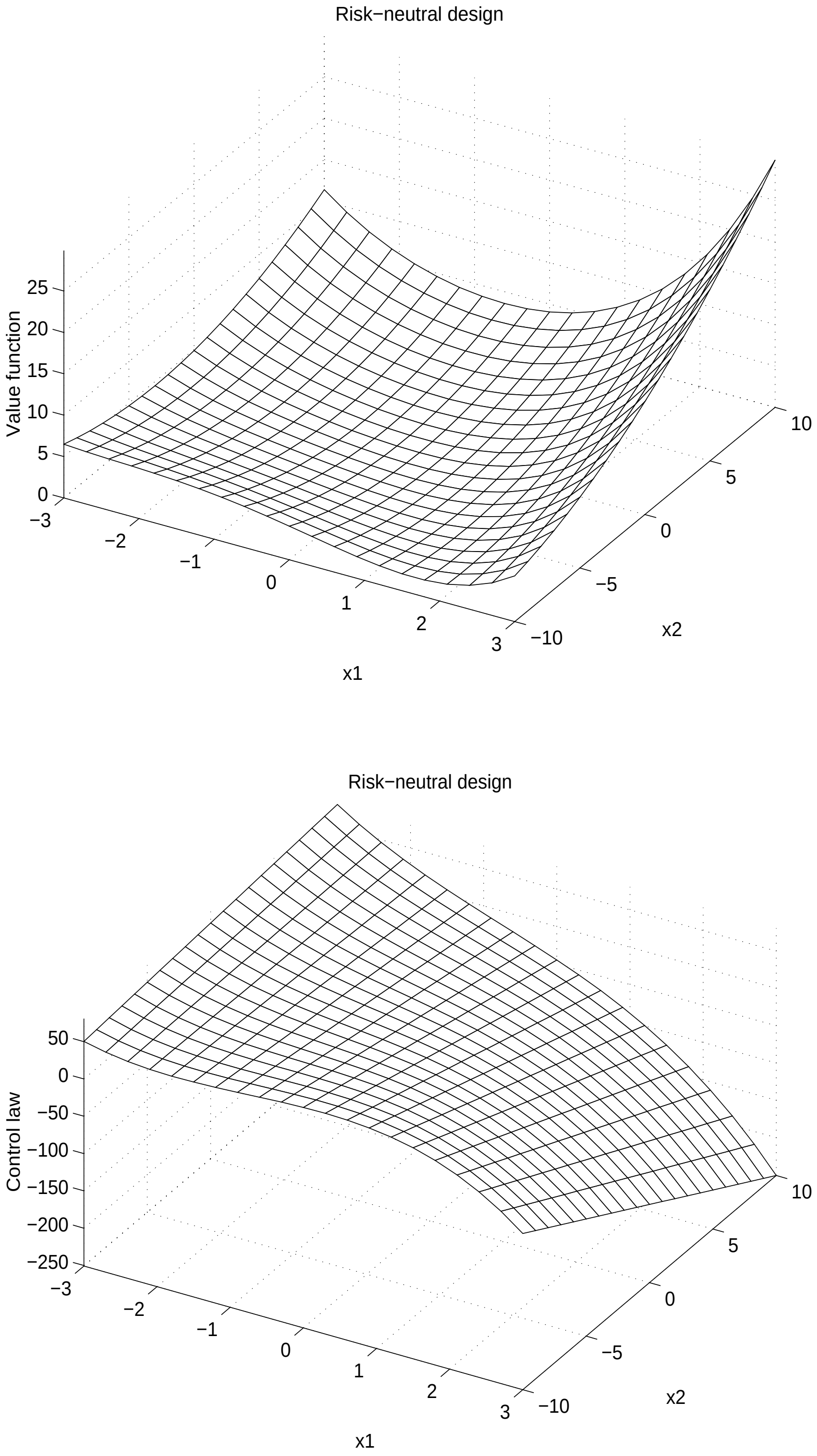


FIG. 7.2. *Value function and control law for risk-neutral design.*

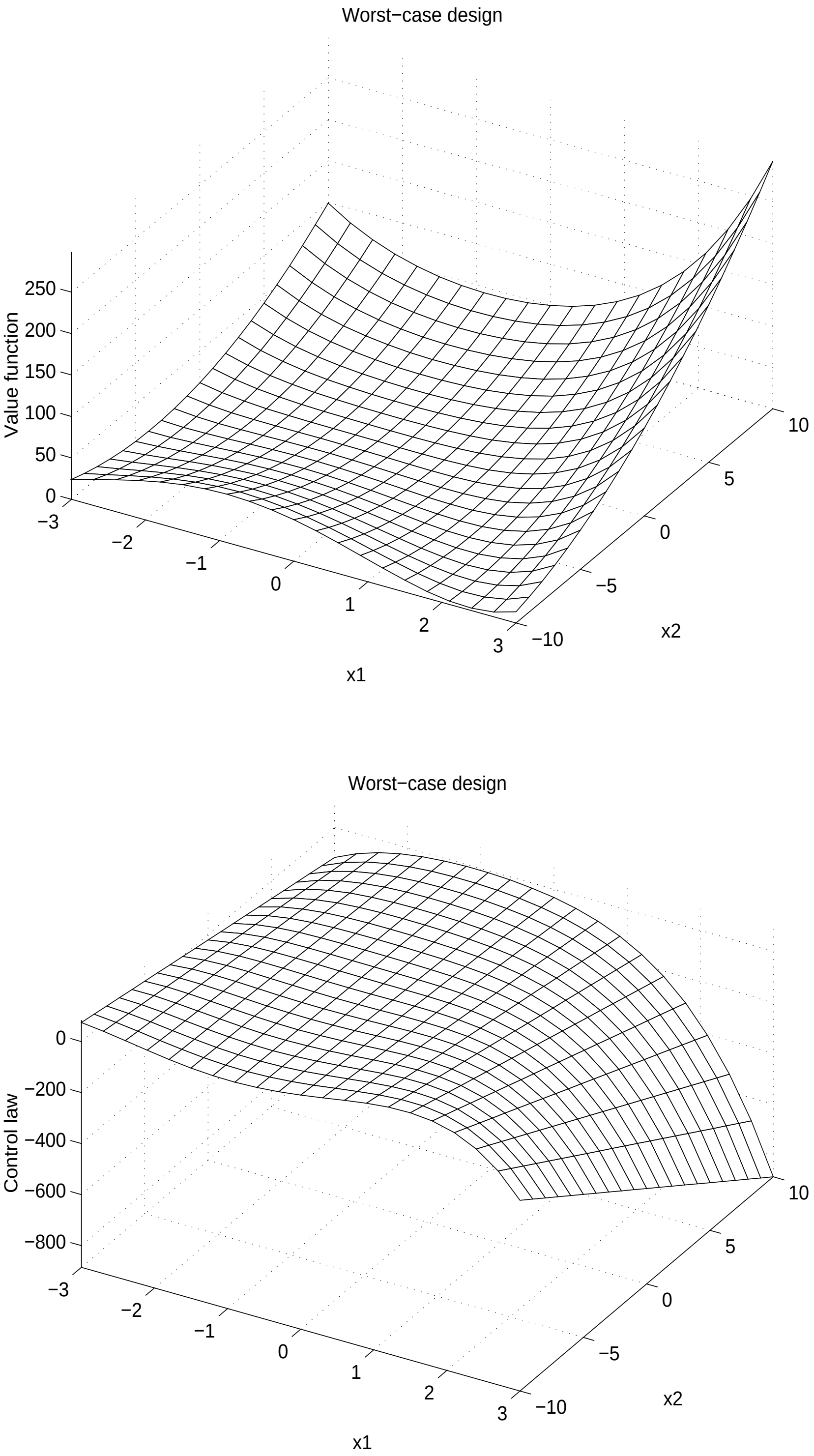


Fig. 7.3. *Value function and control law for worst-case design.*

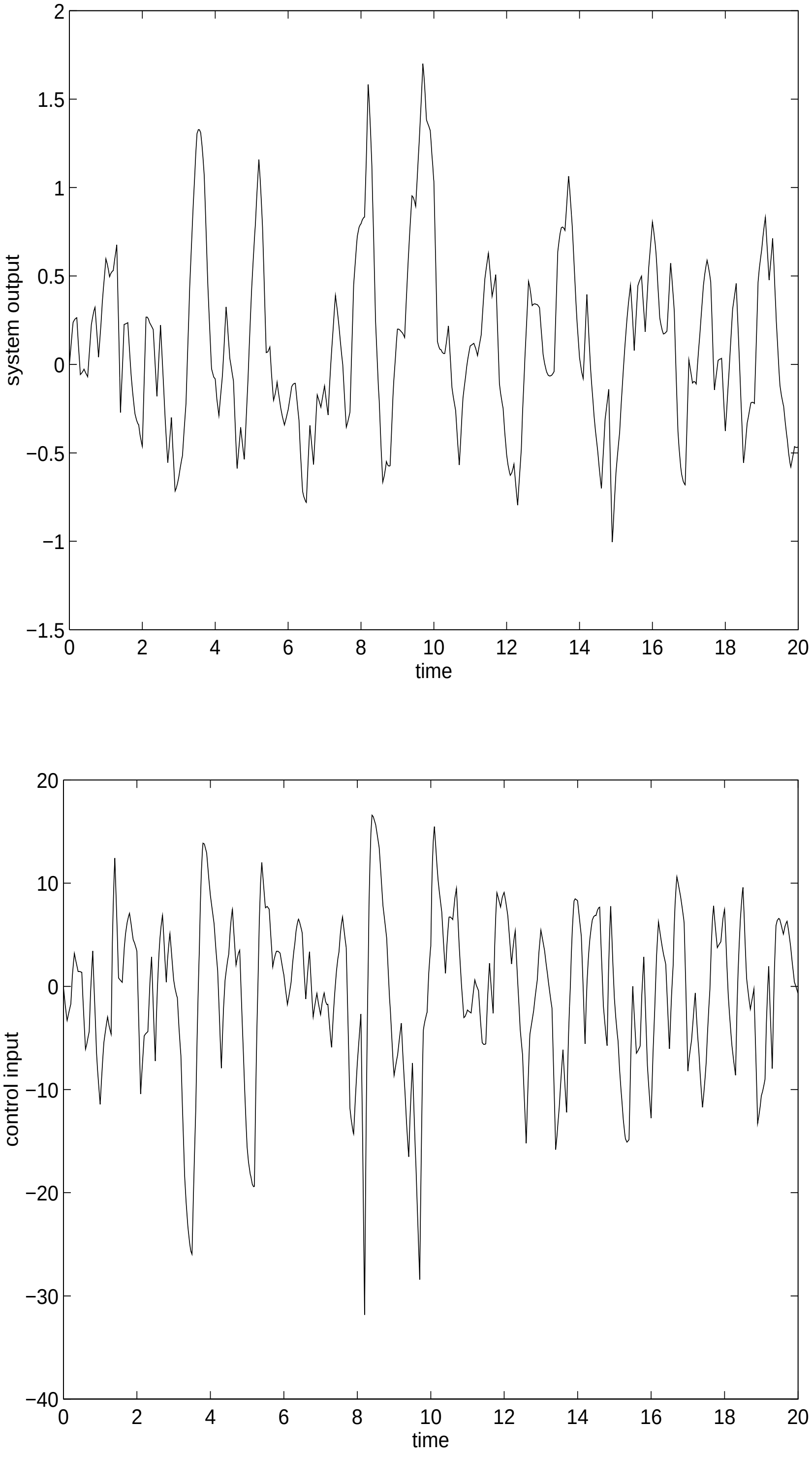


FIG. 7.4. *Risk-sensitive controller under white noise input.*

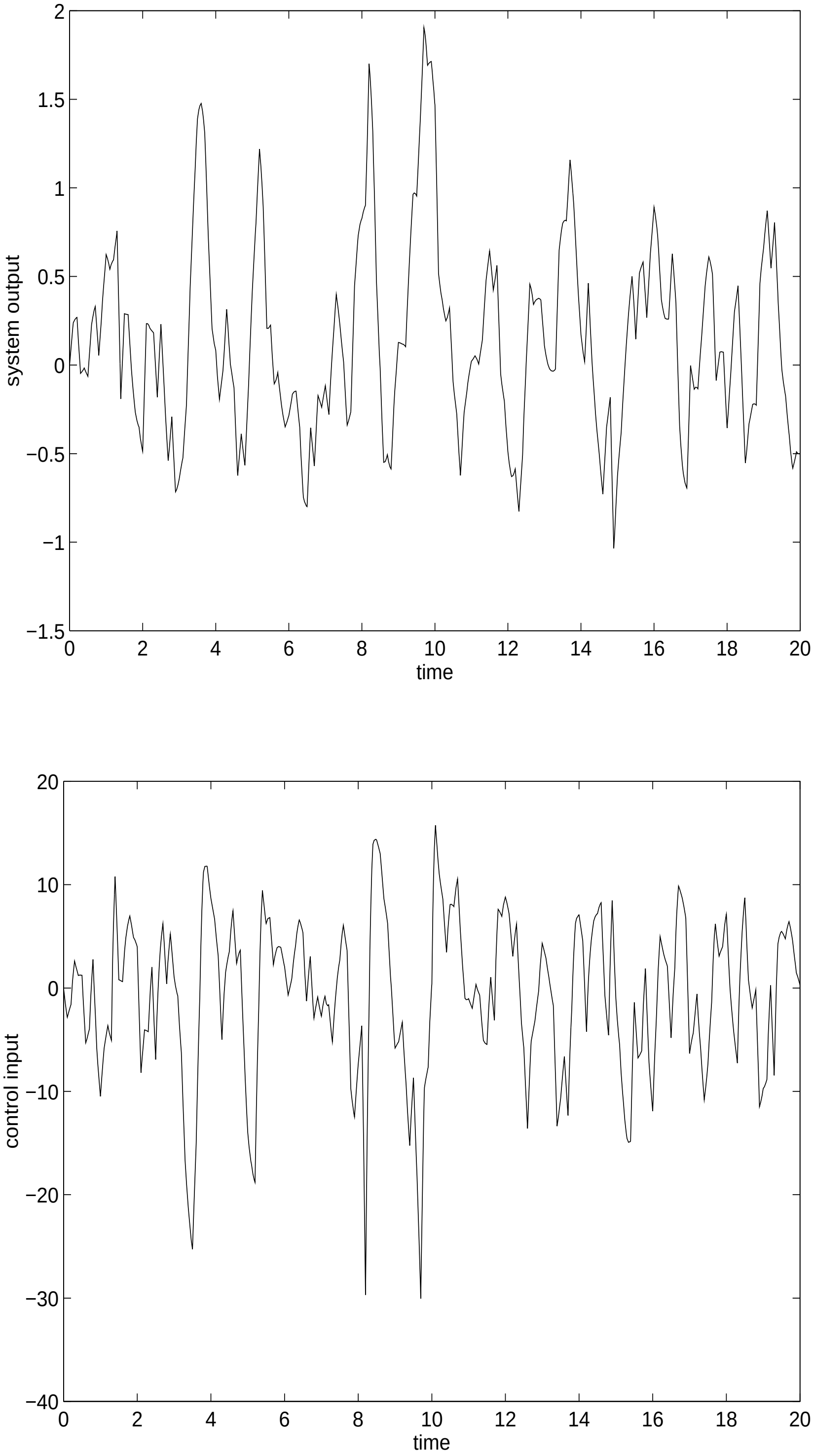


FIG. 7.5. *Risk-neutral controller under white noise input.*

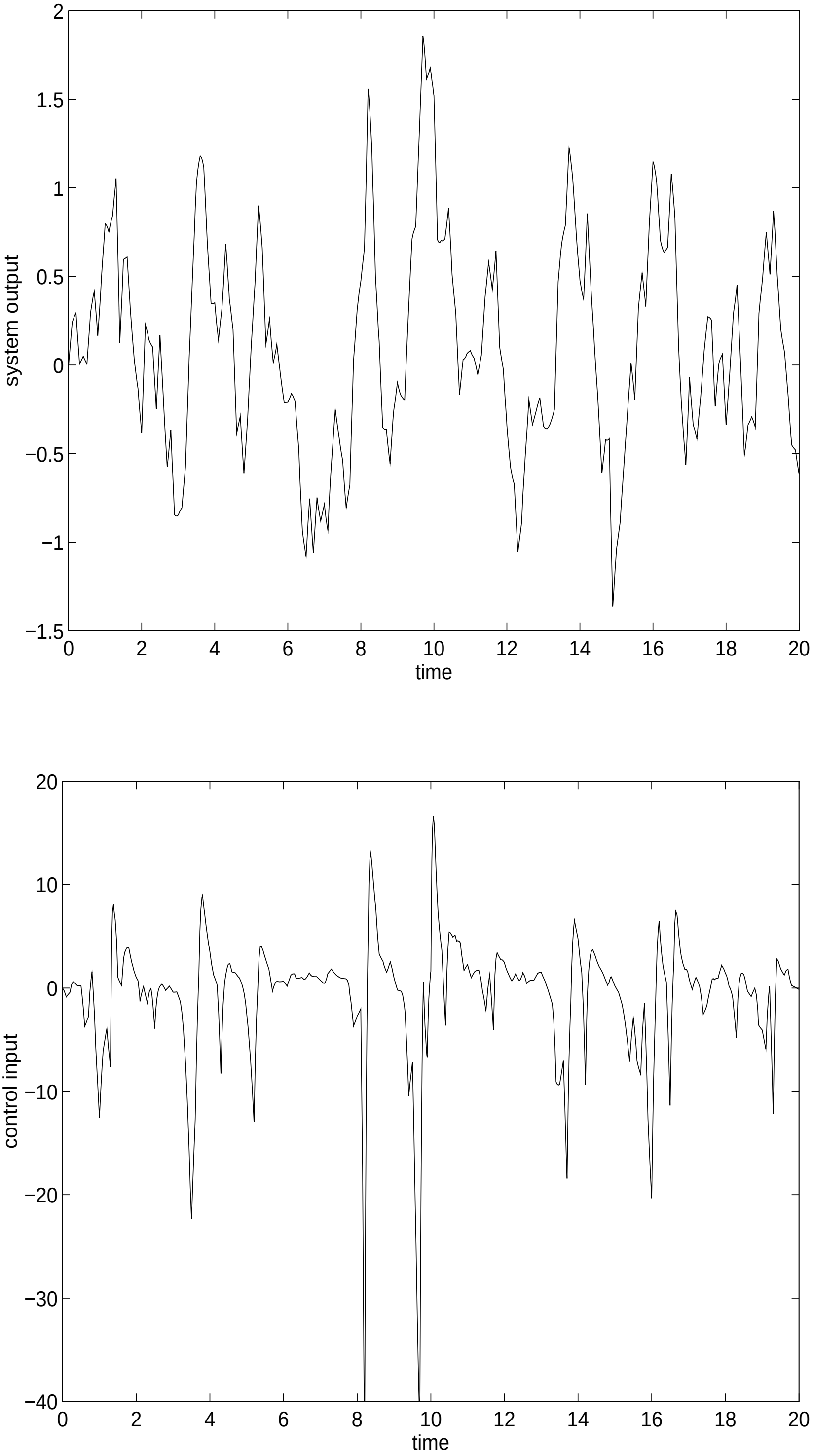


FIG. 7.6. *Worst-case controller under white noise input.*

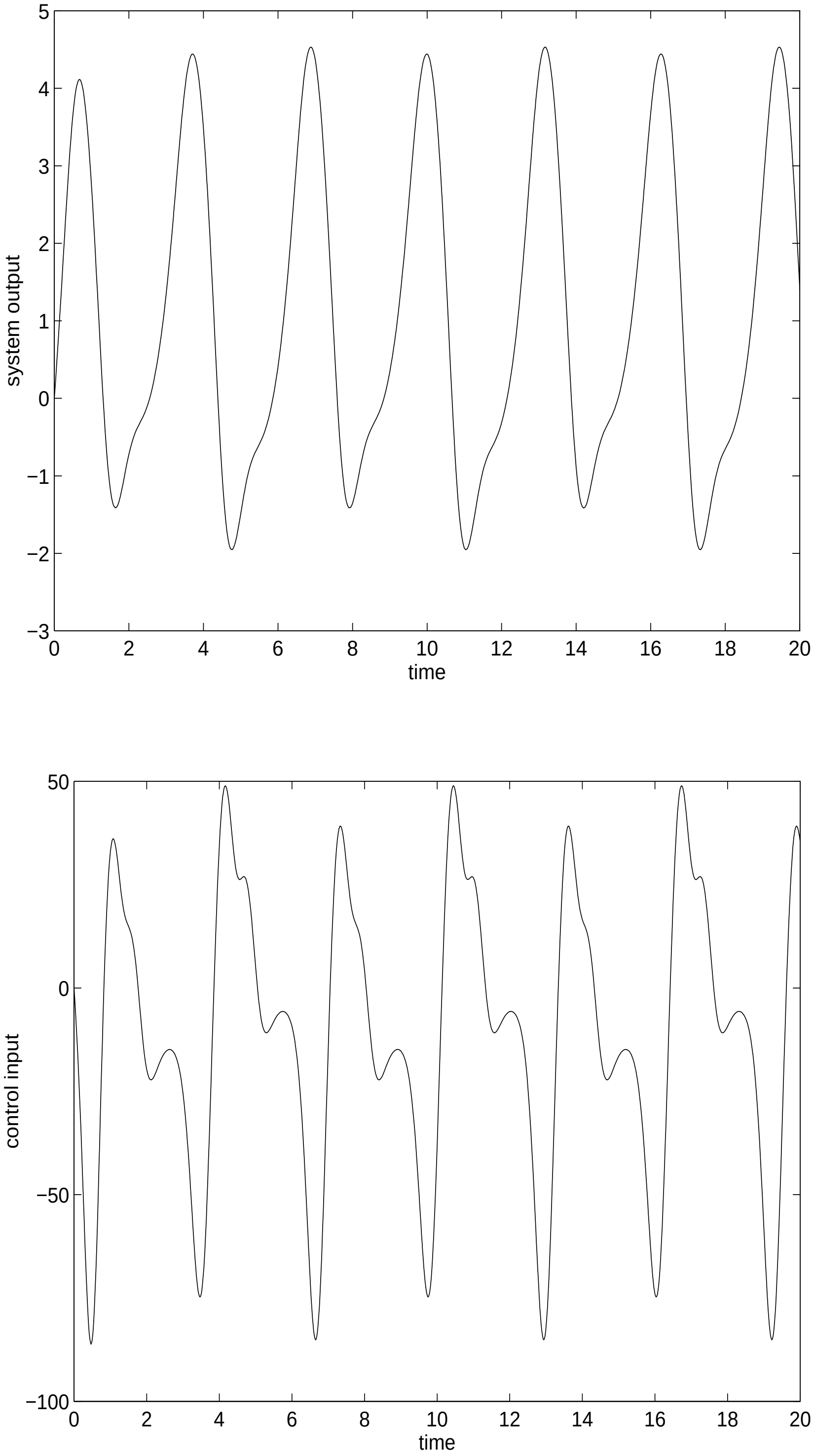


Fig. 7.7. *Risk-sensitive controller under sinusoidal noise input.*

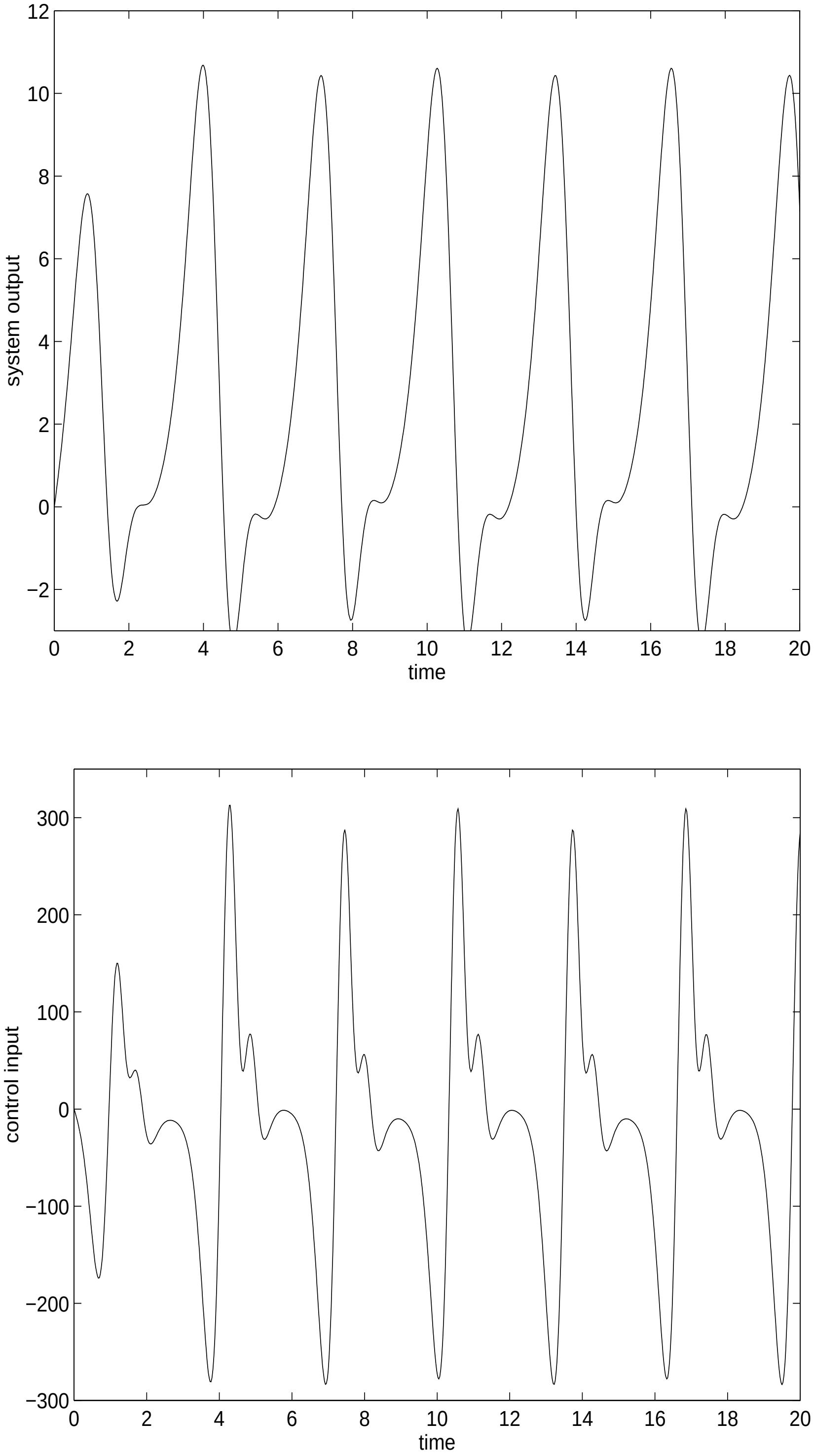


FIG. 7.8. *Risk-neutral controller under sinusoidal noise input.*

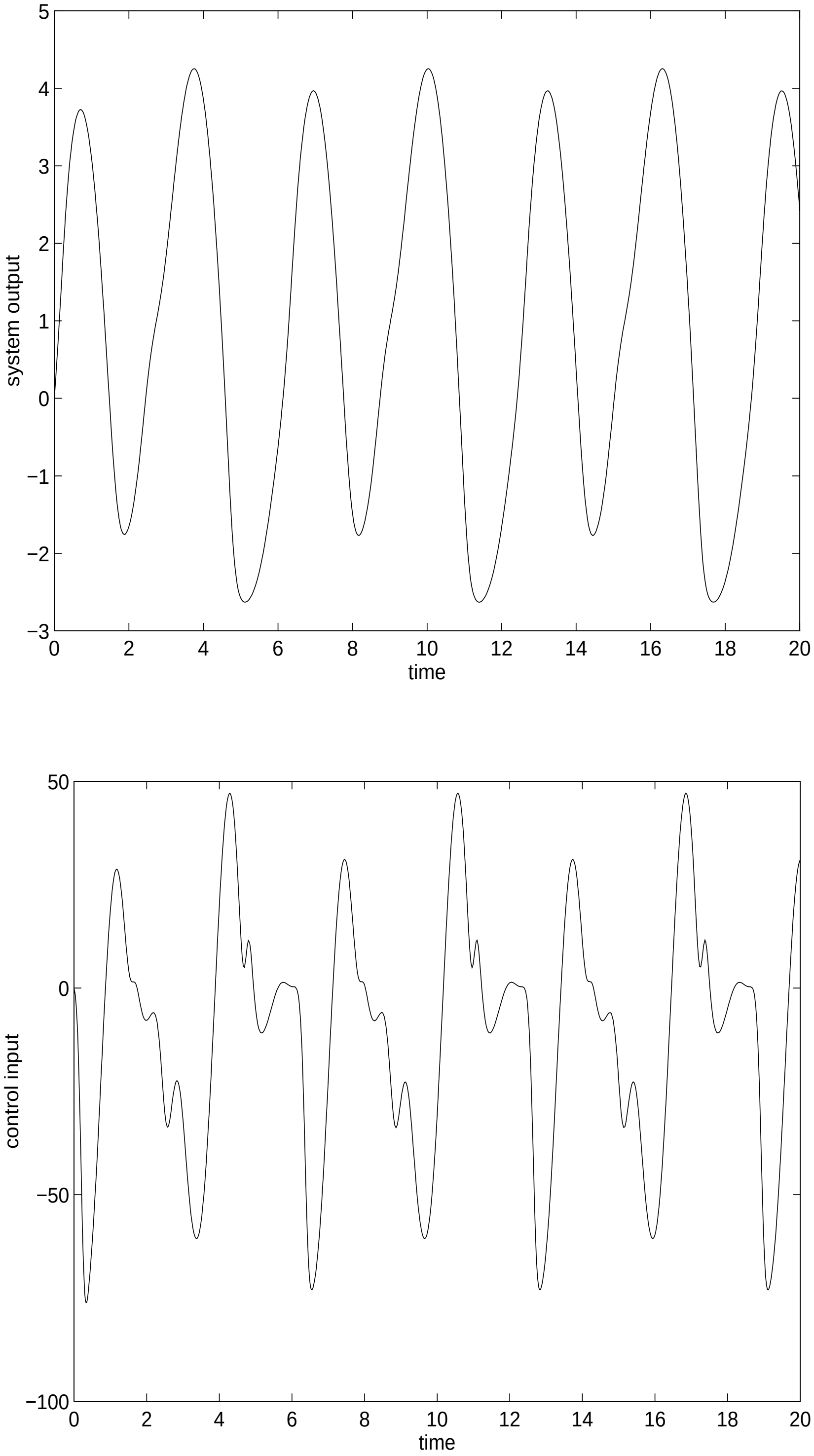


Fig. 7.9. *Worst-case controller under sinusoidal noise input.*

best overall system performance. A simple calculation yields that the $\mathcal{L}_2$ norm of the output for the risk-sensitive control design is the smallest among the three, which is 107.3. Second best is the worst-case controller, whose output norm is 122.3. The risk-neutral controller yields a norm of 479.4 for its output. Regarding the control effort, we observe that the worst-case controller uses the least amount of control effort. The risk-sensitive controller uses 33% more control effort when measured in $\mathcal{L}_2$ norm. The system performance for the risk-neutral controller is not satisfactory under this disturbance.

Overall, we observe that the risk-sensitive controller and the worst-case controller are quite satisfactory for this simple example. The risk-sensitive controller leads to a smaller output $\mathcal{L}_2$ norm in both of the simulations, at the expense of more control effort. This does not contradict the observation that the risk-sensitive controller has a smaller controller gain. It is generally observed that the behavior of the state variable x_2 is much better for the worst-case controller. Smaller control gain guarantees a smaller control effort only if the system states behave the same way.

8. Conclusion. In this paper, we have presented a recursive control design for strict-feedback stochastic nonlinear systems. Under an exponential (risk-sensitive) cost function, we have obtained controllers that guarantee any desired positive level of long-term average cost for a fixed risk-sensitivity parameter. Such an objective was feasible because (i) no weighting was attached to the control variable, (ii) the system was in the strict-feedback form, and (iii) there was weighting only on the output of the system. The construction presented here for the risk-sensitive control problem differs substantially from that of the H^∞ problem [23], which is to be expected since the equivalence between the two problems holds only as the noise intensity diminishes to zero. As a special case, the solution to the risk-neutral control problem has also been obtained by simply taking the limit as the risk-sensitivity parameter approaches zero. Another special case treated was that in which the vector fields for the disturbance vanish at the origin of the system; in this case, the control design can actually guarantee a *zero* long-term average cost. Furthermore, the closed-loop system becomes asymptotically stable in the large. These results have then been generalized to a class of nonlinear systems involving zero dynamics that are strongly stabilizable. The theoretical findings of the paper have been further illustrated in section 7 on a second-order nonlinear system.

Future research opportunities on this topic lie in several directions. One of these pertains to the parameter identification problem under a risk-sensitive performance criterion, to generalize the results of [2] and [24] to the stochastic case. Another interesting topic is the stochastic adaptive control design for strict-feedback systems, to achieve a desired risk-sensitive performance level for a given risk-sensitivity parameter; this would be a generalization of the results of [23] to the stochastic case.

Appendix.

In this appendix, we provide a proof of the fact (see Girsanov [9]) that the stochastic process $\zeta(t)$ defined in the proof of Theorem 2.5 is a supermartingale. The precise statement is as follows.

THEOREM A.1. *Let $w(t)$ be an m-dimensional vector-valued standard Wiener process on the time interval $[0, t_f]$ adapted to a filtration $\mathcal{F}$ on a probability space $(\Omega, \mathcal{F}, \boldsymbol{P})$. Further let a vector-valued stochastic process $f : [0, t_f] \times \Omega \to \mathbb{R}^m$ be adapted to the filtration $\mathcal{F}$. Define the scalar process*

$$\zeta_0^t(f) := \exp\left(\int_0^t f'(s)\,dw - \frac{1}{2}\int_0^t |f(s)|^2\,ds\right).$$

Then, we have $\boldsymbol{E}\left(\zeta_0^t(f)\right) \leq 1 \ \forall t \in [0, t_f]$.

Proof. Let each component of the mapping f be given by

$$f = \begin{bmatrix} f^1 & \cdots & f^m \end{bmatrix}'.$$

For each $n = 1, 2, \ldots$, define the function $f_n : [0, t_f] \times \Omega \to \mathbb{R}^m$ by

$$f_n = \begin{bmatrix} \max\{\min\{f^1, n\}, -n\} & \cdots & \max\{\min\{f^m, n\}, -n\} \end{bmatrix}'.$$

Clearly, we have

$$\lim_{n\to\infty} f_n = f \qquad \text{almost everywhere.}$$

It is easy to check that

$$\boldsymbol{E}\left(\exp\left(\frac{1}{2}\int_0^t |f_n(s)|^2\, ds\right)\right) \leq \exp\left(\frac{1}{2}n^2 t\right) < \infty.$$

By the result of [21], we have

$$\boldsymbol{E}\left(\zeta_0^t(f_n)\right) = 1 \qquad \forall\ t \in [0, t_f];\ n = 1, 2, \ldots .$$

By Fatou's lemma, we have

$$1 = \liminf_{n\to\infty} \boldsymbol{E}\left(\zeta_0^t(f_n)\right) \geq \boldsymbol{E}\left(\liminf_{n\to\infty} \zeta_0^t(f_n)\right) = \boldsymbol{E}\left(\zeta_0^t(f)\right).$$

This completes the proof of the theorem. □

Next, we present a lemma which is used in the proof of Proposition 2.6.

LEMMA A.2. *Let $w(t)$ be an m-dimensional vector-valued standard Wiener process on the time interval $[0, t_f]$ adapted to a filtration $\mathcal{F}$ on a probability space $(\Omega, \mathcal{F}, \boldsymbol{P})$. Let x and z be two constant m-dimensional vectors satisfying*

$$x'\, dw = z'\, dw$$

for some time instance $t_0 \geq 0$. Then, $x = z$.

Proof. Let x, z, and $w(t)$ be given by

$$x = \begin{bmatrix} x_1 & \cdots & x_m \end{bmatrix}' \quad z = \begin{bmatrix} z_1 & \cdots & z_m \end{bmatrix}' \quad w(t) = \begin{bmatrix} w_1(t) & \cdots & w_m(t) \end{bmatrix}'.$$

At t_0, we have

$$\sum_{i=1}^m x_i\, dw_i(t_0) = \sum_{i=1}^m z_i\, dw_i(t_0).$$

For each $i = 1, 2, \ldots, m$, we multiply both sides of this equation by $dw_i(t_0)$ and then take expectations on both sides of the equation. This leads to

$$x_i\, dt = z_i\, dt, \qquad i = 1, 2, \ldots, m \qquad \Rightarrow \qquad x_i = z_i, \qquad i = 1, 2, \ldots, m,$$

from which it follows that $x = z$. This completes the proof of the lemma. □

REFERENCES

[1] T. Başar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, 2nd ed., Birkhäuser, Boston, MA, 1995.

[2] G. Didinsky, Z. Pan, and T. Başar, *Parameter identification for uncertain plants using H^∞ methods*, Automatica J. IFAC, 31 (1995), pp. 1227–1250.

[3] W. H. Fleming and M. R. James, *The risk-sensitive index and the H_2 and H^∞ norms for nonlinear systems*, Math. Control Signals Systems, 8 (1995), pp. 199–221.

[4] W. H. Fleming and W. M. McEneaney, *Risk sensitive control with ergodic cost criteria*, in Proceedings of the 31st Conference on Decision and Control, Tucson, AZ, 1992, pp. 2048–2052.

[5] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Appl. Math. 25, Springer-Verlag, Berlin, New York, 1993.

[6] R. A. Freeman and P. V. Kokotović, *Design of "softer" robust control laws*, Automatica J. IFAC, 29 (1993), pp. 1425–1437.

[7] R. A. Freeman and P. V. Kokotović, *Inverse optimality in robust stabilization*, SIAM J. Control Optim., 34 (1996) pp. 1365–1391.

[8] I. I. Gihman and A. V. Skorohod, *Stochastic Differential Equations*, Springer-Verlag, New York, 1972.

[9] I. V. Girsanov, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theory Probab. Appl., 5 (1960), pp. 285–301.

[10] A. Isidori, *Nonlinear Control Systems*, 3rd ed., Springer-Verlag, London, 1995.

[11] D. H. Jacobson, *Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games*, IEEE Trans. Automat. Control, 18 (1973), pp. 167–172.

[12] S. Jain and F. Khorrami, *Application of a decentralized adaptive output feedback based on backstepping to power systems*, in Proceedings of the 34th IEEE Conference on Decision and Control, New Orleans, LA, 1995, pp. 1585–1590.

[13] M. R. James, J. Baras, and R. J. Elliott, *Risk-sensitive control and dynamic games for partially observed discrete-time nonlinear systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 780–792.

[14] Z. P. Jiang and J.-B. Pomet, *Backstepping-based adaptive controllers for uncertain nonholonomic systems*, in Proceedings of the 34th IEEE Conference on Decision and Control, New Orleans, LA, 1995, pp. 1573–1578.

[15] I. Kanellakopoulos, P. V. Kokotović, and A. S. Morse, *Systematic design of adaptive controllers for feedback linearizable systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 1241–1253.

[16] H. K. Khalil, *Nonlinear Systems*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, 1996.

[17] R. Z. Khas'minskii, *Stochastic Stability of Differential Equations*, S&N International Publishers, Rockville, MD, 1980.

[18] M. Krstić, I. Kanellakopoulos, and P. V. Kokotović, *Nonlinear design of adaptive controllers for linear systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 738–752.

[19] M. Krstić, I. Kanellakopoulos, and P. V. Kokotović, *Nonlinear and Adaptive Control Design*, Wiley, New York, 1995.

[20] H. Nagai, *Bellman equations of risk-sensitive control*, SIAM J. Control Optim., 34 (1996), pp. 74–101.

[21] A. A. Novikov, *On an identity for stochastic integrals*, Theory Probab. Appl., 16 (1972), pp. 717–720.

[22] Z. Pan, *Canonical forms for stochastic nonlinear systems*, in Proceedings of the 36th IEEE Conf. on Decision and Control, San Diego, CA, 1997, pp. 24-29.

[23] Z. Pan and T. Başar, *Adaptive controller design for tracking and disturbance attenuation in parametric-strict-feedback nonlinear systems*, in Proceedings of the 13th IFAC Congress of Automatic Control, San Francisco, CA, 1996.

[24] Z. Pan and T. Başar, *Parameter identification for uncertain linear systems with partial state measurements under an H^∞ criterion*, IEEE Trans. Automat. Control, to appear.

[25] T. Runolfsson, *The equivalence between infinite horizon control of stochastic systems with exponential-of-integral performance index and stochastic differential games*, IEEE Trans.

Automat. Control, 39 (1994), pp. 1551–1563.

[26] D. Seto, A. M. Annaswamy, and J. Baillieul, *Adaptive control of nonlinear systems with triangular structure*, IEEE Trans. Automat. Control, 39 (1994), pp. 1411–1428.

[27] M. W. Spong and Vidyasagar, *Robot Dynamics and Control*, John Wiley and Sons, New York, 1989.

[28] P. Whittle, *Risk-sensitive linear-quadratic-Gaussian control*, Adv. in Appl. Probab., 13 (1981), pp. 764–777.

[29] P. Whittle, *Risk-Sensitive Optimal Control*, John Wiley & Sons, Chichester, NY, 1990.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 4, pp. 997–1010

AVERAGING RESULTS AND THE STUDY OF UNIFORM ASYMPTOTIC STABILITY OF HOMOGENEOUS DIFFERENTIAL EQUATIONS THAT ARE NOT FAST TIME-VARYING*

JOAN PEUTEMAN[†] AND DIRK AEYELS[†]

Abstract. Within the Liapunov framework, a sufficient condition for uniform asymptotic stability of ordinary differential equations is proposed. Unlike with classical Liapunov theory, the time derivative of the V-function, taken along solutions of the system, may have positive and negative values. It is shown that the proposed condition is useful for the study of uniform asymptotic stability of homogeneous systems with order $\tau > 0$. In particular, it is established that asymptotic stability of the averaged homogeneous system implies local uniform asymptotic stability of the original time-varying homogeneous system. This shows that averaging techniques play a prominent role in the study of homogeneous—not necessarily fast time-varying—systems.

Key words. nonlinear systems, homogeneous systems, averaging, asymptotic stability

AMS subject classifications. 34, 34D, 34D20, 93D20

PII. S0363012997323862

1. Introduction.

The classical Liapunov approach to uniform asymptotic stability of the null solution of a dynamical system $\dot{x}(t) = f(x,t)$ requires the existence of a positive definite, decrescent Liapunov function $V(x,t)$ whose derivative along the solutions of the system is negative definite. When this derivative is negative semidefinite, stability rather than asymptotic stability follows; in case the differential equation is autonomous, the Barbashin–Krasovskii theorem or LaSalle's invariance principle may be helpful in proving asymptotic stability. Extensions to periodic differential equations are possible. For nonperiodic systems, Narendra and Annaswamy [12] show that with $\dot{V}(x,t) \leq 0$ uniform asymptotic stability can be proven if there exists a $T \in \mathbb{R}^+$ such that $\forall t : V(x(t+T), t+T) - V(x(t),t) \leq -\gamma(\|x(t)\|) < 0$, where $\gamma(\cdot)$ is a strictly increasing continuous function on $\mathbb{R}^+$ which is zero at the origin and where $x(t+T)$ is the solution of the system at $t+T$ with initial condition $x(t)$ at t. Weaker conditions than the Narendra–Annaswamy conditions also leading to asymptotic stability have recently been obtained [1, 2, 3, 4]. The present paper and [3, 4] have been inspired by the result of Narendra–Annaswamy. We claim that in the asymptotic stability theorem of Narendra–Annaswamy, *the negative semidefiniteness condition on the time-derivative of the V-function can be dispensed with*: the null solution of a differential equation is uniformly asymptotically stable under the condition that for a positive definite decrescent $V(x,t)$, $\exists T > 0$, and a strictly increasing sequence of times t_k^* such that $V(x(t_{k+1}^*), t_{k+1}^*) - V(x(t_k^*), t_k^*) \leq -\gamma(\|x(t_k^*)\|)$ with $\gamma(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$ continuous, strictly increasing and passing through the origin and $t_{k+1}^* - t_k^* \leq T \ \forall k \in \mathbb{Z}$ and $t_k^* \to \infty$ as $k \to \infty$ and $t_k^* \to -\infty$ as $k \to -\infty$. Compared to [12], $\dot{V}(x,t) \leq 0$ is no longer required. Compared to [3], the condition on V needs to be satisfied only for a sequence t_k^*, not for all t. Unlike [4], this paper is focused upon uniform asymptotic stability, not exponential stability.

[*] Received by the editors July 2, 1997; accepted for publication (in revised form) August 17, 1998; published electronically April 7, 1999. This paper presents research results of the Belgian Programme on Interuniversity Poles of Attraction initiated by the Belgian Prime Minister's Office for Science, Technology and Culture. The scientific responsibility rests with its authors.

http://www.siam.org/journals/sicon/37-4/32386.html

[†] SYSTeMS, Universiteit Gent, Technologiepark-Zwijnaarde 9, 9052 Gent (Zwijnaarde), Belgium (joan.peuteman@rug.ac.be, dirk.aeyels@rug.ac.be).

In section 2 of this paper, the classical theorem of Liapunov for discrete-time nonlinear systems is recalled. In section 3, we propose a sufficient condition which guarantees uniform asymptotic stability of a continuous-time nonlinear system. In section 4 and section 5, a proposition and a theorem are stated concerning uniform asymptotic stability of time-varying homogeneous systems $\dot{x}(t) = f(x, \alpha t)$ $(\forall \alpha \in \mathbb{R}_0^+)$ with order $\tau > 0$. It is shown that asymptotic stability of the averaged (*time-invariant* homogeneous) system guarantees local uniform asymptotic stability of the original *time-varying* homogeneous system. For stability results on time-invariant homogeneous systems, the reader is referred to [7, 9]. An important—perhaps surprising—feature of our result is that it is valid *independent of* α. This shows that averaging techniques play a prominent role in the study of time-varying—not necessarily fast time-varying—homogeneous systems of order $\tau > 0$.

As illustrated in section 6 and studied in [11], it is worthwhile mentioning that the averaging results for homogeneous systems with order $\tau = 0$ are different in the sense that they are valid only for α sufficiently large.

The region of attraction of the homogeneous system $\dot{x}(t) = f(x, \alpha t)$ $(\alpha \in \mathbb{R}_0^+)$ with order $\tau > 0$ depends on α. This region of attraction increases when α increases and grows unbounded as α goes to infinity.

2. Theorem of Liapunov for discrete-time systems. Consider the time-varying discrete-time system

$$x_{k+1} = g(x_k, k) \tag{1}$$

with $g : W_d \times \mathbb{Z} \to \mathbb{R}^n$, W_d open, $W_d \subset \mathbb{R}^n$. Let $g(0, k) = 0\ \forall\ k \in \mathbb{Z}$ and $0 \in W_d$.

PROPOSITION 1. *Consider a function $V : U_d \times \mathbb{Z} \to \mathbb{R}$, with U_d an open neighborhood of* 0. *We assume the following.*

Condition 1. *$V(x, k)$ is positive definite and decrescent; i.e., $V(0, k) = 0 \quad \forall k$ and $\forall x \in U_d : \alpha_V(\|x\|) \leq V(x, k) \leq \beta_V(\|x\|)$. The functions $\alpha_V(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$ and $\beta_V(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$ are strictly increasing continuous functions passing through the origin.*

Condition 2. *There exists a function $\gamma(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$, which is continuous, strictly increasing, and passing through the origin, and an open set $U'_d \subset U_d \cap W_d$, which contains the origin such that $\forall k \in \mathbb{Z}\ \forall x_k \in U'_d \setminus \{0\}$:*

$$V(x_{k+1}, k+1) - V(x_k, k) \leq -\gamma(\|x_k\|) < 0, \tag{2}$$

where $x_{k+1} = g(x_k, k)$.

Then the equilibrium point $x = 0$ of (1) *is locally uniformly asymptotically stable.*

Proof. The present proposition is the classical theorem of Liapunov for uniform asymptotic stability for discrete-time nonlinear time-varying systems. For the proof of this proposition, the reader is referred to [5, 8]. □

Remark 1. When the closed ball with radius ν centered at 0, $\overline{B}_\nu(0) \subset U'_d$ and $x_{k_0} \in B_{\nu'}(0)$ with $\nu' = \beta_V^{-1}(\alpha_V(\nu))$, then by (2), x_k exists and $x_k \in B_\nu(0) \subset U'_d\ \forall k_0$ and $\forall k \geq k_0$. The proof of Proposition 1, which is entirely analogous to the proof of the Liapunov theorem in the continuous-time case, implies that the open ball with radius ν' centered at 0, $B_{\nu'}(0)$ belongs to the region of attraction.

3. A sufficient condition for uniform asymptotic stability. In this section, a sufficient condition for uniform asymptotic stability of a continuous-time system is proposed. In classical Liapunov theory, the time derivative of the Liapunov function V along the solutions of the system is required to be negative definite. In the present case, the derivative of the V-function may have positive and negative values. The theorem requires only the existence of a positive definite and decrescent V-function and a sequence of times t_k^* such that the values of this V-function decrease when evaluated along the solutions at t_k^*.

Consider

$$\dot{x}(t) = f(x,t) \tag{3}$$

with $f : W_c \times \mathbb{R} \to \mathbb{R}^n$, W_c open, $W_c \subset \mathbb{R}^n$. Let $f(0,t) = 0 \ \forall\ t \in \mathbb{R}$ and $0 \in W_c$. Furthermore we assume that conditions are imposed on (3) such that existence and uniqueness of its solutions are secured. These conditions are imposed on all the differential equations mentioned in the present paper, and of these conditions we single out the local Lipschitz condition. This local Lipschitz condition will be used in the course of the proof of the propositions and the theorems hereafter: f is locally Lipschitz, i.e., for $\forall x \in W_c$, $\exists$ a neighborhood $\mathcal{N}(x) \subset W_c$, such that the restriction $f|_{\mathcal{N}(x)\times\mathbb{R}} \overset{not}{=} f|_{\mathcal{N}(x)}$ is Lipschitz with Lipschitz function $l_x(t)$. We assume that $l_x(t)$ is bounded $\forall\ t \in \mathbb{R}$. We are now ready to state a lemma and Proposition 2.

LEMMA 1. *Let $U \subset W_c$ be an open neighborhood of 0. Consider a closed ball $\overline{B}_\mu(0) \subset U$; then $\forall\ T > 0$, $\exists \mu' > 0$ such that $(x_0, t_0) \in B_{\mu'}(0) \times \mathbb{R}$ implies that $x(t) \in B_\mu(0)$ for $t \in [t_0, t_0 + T]$. Here $x(t)$ is the solution of (3) evaluated at t with initial condition x_0 at t_0. (The solutions are assumed to exist over the considered time interval.)*

Proof. The proof of the lemma is omitted. It can be found in [3]. □

Remark 2. The proof of the lemma shows that $\mu' = \mu e^{-KT}$ is an appropriate choice of μ'. K is the Lipschitz constant of (3) on $\overline{B}_\mu(0)$ [10, p. 70].

PROPOSITION 2. *Consider a function $V : U \times \mathbb{R} \to \mathbb{R}$, with U an open neighborhood of 0. We assume that the following additional conditions are satisfied.*

1. *Condition 1. $V(x,t)$ is positive definite and decrescent; i.e., $V(0,t) = 0 \ \ \forall t$ and $\forall x \in U : \alpha_V(\|x\|) \le V(x,t) \le \beta_V(\|x\|)$. The functions $\alpha_V(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$ and $\beta_V(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$ are strictly increasing continuous functions passing through the origin.*

2. *Condition 2. There exists an increasing sequence of times t_k^* ($k \in \mathbb{Z}$) with $t_k^* \to \infty$ as $k \to \infty$ and $t_k^* \to -\infty$ as $k \to -\infty$, $\exists$ finite $T > 0 : t_{k+1}^* - t_k^* \le T$ ($\forall k \in \mathbb{Z}$); there exists a function $\gamma(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$ which is continuous, strictly increasing, and passing through the origin and an open set $U' \subset U \cap W_c$ which contains the origin such that $\forall k \in \mathbb{Z}\ \forall x(t_k^*) \in U' \setminus \{0\}$:*

$$V(x(t_{k+1}^*), t_{k+1}^*) - V(x(t_k^*), t_k^*) \le -\gamma(\|x(t_k^*)\|) < 0, \tag{4}$$

where $x(t_{k+1}^*)$ is the solution of (3) at t_{k+1}^* with initial condition $x(t_k^*)$ at t_k^*.

Then the equilibrium point $x = 0$ of (3) is locally uniformly asymptotically stable.

Proof. By the continuous-time system $\dot{x} = f(x,t)$ and the sequence t_k^*, we define the discrete-time system $x_{k+1} = g(x_k, k)$ with $g(x_k, k) = x(t_{k+1}^*)\ \forall x_k \in U' = W_d$. Here $x(t_{k+1}^*)$ is the solution of (3) at t_{k+1}^* with initial condition $x(t_k^*) = x_k$ at t_k^*. The continuous times t_k^* are identified with the discrete times k ($\forall k \in \mathbb{Z}$) and the state $x(t_k^*)$ is equal to x_k ($\forall k \in \mathbb{Z}$).

Uniform asymptotic stability of $x_{k+1} = g(x_k, k)$. Condition 1 and Condition 2 of Proposition 2 imply that Condition 1 and Condition 2 of Proposition 1 are fulfilled with $W_d = U_d = U_d' = U'$, which implies local uniform asymptotic stability of $x_{k+1} = g(x_k, k)$.

We are set to prove local uniform asymptotic stability of (3). First uniform stability will be established and then uniform convergence will be established.

Uniform stability of $\dot{x} = f(x,t)$. Consider a closed ball $\overline{B}_\epsilon(0)$ centered at 0 and radius ϵ small enough such that $\overline{B}_\epsilon\ (0) \subset U'$. Let K be the Lipschitz constant of (3) on $\overline{B}_\epsilon(0)$. Define $\epsilon' := \epsilon\ e^{-KT}$. Define $\delta' := \beta_V^{-1}(\alpha_V(\epsilon'))$ and $\delta'' := e^{-KT}\delta'$.

Consider the open ball $B_{\delta''}(0)$. For all $(x(t_0), t_0) \in B_{\delta''}(0) \times \mathbb{R}$, there exists a $k_0 \in \mathbb{Z}$ such that $t^*_{k_0} - t_0 < T$. By Lemma 1 and Remark 2 with $\mu = \delta'$ and $\mu' = \delta''$, one obtains that $\forall t \in [t_0, t^*_{k_0}]$, $t - t_0 < T$ which implies that

$$\|x(t)\| < \delta' = \delta'' e^{KT} < \epsilon \qquad \forall t \in [t_0, t^*_{k_0}] \tag{5}$$

and $\|x(t^*_{k_0})\| < \delta'$.

By Proposition 1 and Remark 1 with $\nu = \epsilon'$ and $\nu' = \delta'$, it is clear that $\forall n \in \mathbb{N}$: x_{k_0+n} exists and $\|x_{k_0+n}\| < \epsilon'$, where x_{k_0+n} is the solution of $x_{k+1} = g(x_k, k)$ at $k_0 + n$ with initial condition $x_{k_0} = x(t^*_{k_0})$ at k_0.

For all $n \in \mathbb{N}$, x_{k_0+n} equals $x(t^*_{k_0+n})$, where $x(t^*_{k_0+n})$ is the solution of (3) at $t^*_{k_0+n}$ with initial condition $x(t^*_{k_0}) = x_{k_0}$ at $t^*_{k_0}$. This implies that $\forall n \in \mathbb{N} : \|x(t^*_{k_0+n})\| = \|x_{k_0+n}\| < \epsilon'$.

Notice that $\forall\ t \geq t^*_{k_0}$, $\exists n \in \mathbb{N}$ such that $t - t^*_{k_0+n} < T$. Since $\|x(t^*_{k_0+n})\| < \epsilon' = \epsilon e^{-KT}$ $\forall n \in \mathbb{N}$, Lemma 1 and Remark 2 with $\mu = \epsilon$ and $\mu' = \epsilon'$ imply that

$$\|x(t)\| < \epsilon \qquad \forall\ t \geq t^*_{k_0}. \tag{6}$$

By (5) and (6), $\|x(t)\| < \epsilon$ $\forall t \geq t_0$ when $(x(t_0), t_0) \in B_{\delta''}(0) \times \mathbb{R}$.

Uniform stability of (3) is established when

$$\forall \epsilon_c > 0, \exists \delta(\epsilon_c) > 0 : \|x(t_0)\| < \delta(\epsilon_c) \Rightarrow \|x(t)\| < \epsilon_c\ \ \forall t, t_0\ \text{ such that }\ t \geq t_0.$$

If $\overline{B}_{\epsilon_c} \subset U'$, then take $\epsilon = \epsilon_c$. If $\overline{B}_{\epsilon_c}(0) \not\subset U'$, then take ϵ small enough such that $\overline{B}_\epsilon(0) \subset U'$. Simply take $\delta(\epsilon_c) = \delta'' = e^{-KT}\beta_V^{-1}(\alpha_V(\epsilon e^{-KT}))$.

Uniform convergence of $\dot{x} = f(x,t)$. We prove the existence of an $\epsilon_1 > 0$ such that $\forall \epsilon_2 > 0$, there exists a $T(\epsilon_2) \geq 0$ such that $\|x(t_0)\| < \epsilon_1$ (t_0 arbitrary) implies $\|x(t)\| < \epsilon_2$ $\forall\ t \geq t_0 + T(\epsilon_2)$. Here $x(t)$ is the solution of (3) with initial condition $x(t_0)$ at t_0.

Take ϵ small enough such that $\overline{B}_\epsilon(0) \subset U'$. Let K be the Lipschitz constant of (3) on $\overline{B}_\epsilon(0)$. Take $\epsilon_1 = e^{-KT}\beta_V^{-1}(\alpha_V(\epsilon e^{-KT}))$. For all $(x(t_0), t_0) \in B_{\epsilon_1}(0) \times \mathbb{R}$, there exists a $k_0 \in \mathbb{Z}$ such that $t^*_{k_0} - t_0 < T$. Since $\|x(t_0)\| < \epsilon_1$, by Lemma 1 and Remark 2 with $\mu = \epsilon_1 e^{KT}$ and $\mu' = \epsilon_1$, one obtains that

$$\|x(t^*_{k_0})\| < \epsilon_1 e^{KT} = \beta_V^{-1}(\alpha_V(\epsilon e^{-KT})). \tag{7}$$

Take $\epsilon_{1d} := \beta_V^{-1}(\alpha_V(\epsilon))$. By the convergence property of $x_{k+1} = g(x_k, k)$, Proposition 1 and Remark 1 with $\nu = \epsilon$ and $\nu' = \epsilon_{1d}$ imply that $\forall \epsilon_{2d} : \exists k(\epsilon_{2d})$ such that $\|x_{k_0}\| < \epsilon_{1d}$ implies that $\|x_k\| < \epsilon_{2d}$ $\ \forall k \geq k_0 + k(\epsilon_{2d})$.

Since by (7) $\|x_{k_0}\| = \|x(t^*_{k_0})\| < \epsilon_{1d}$, one obtains by taking $\epsilon_{2d} = \epsilon_2 e^{-KT}$ that $\forall k \geq k_0 + k(\epsilon_{2d})$ that

$$\|x_k\| = \|x(t^*_k)\| < \epsilon_2 e^{-KT}. \tag{8}$$

By Lemma 1 and Remark 2 with $\mu = \epsilon_2$ and $\mu' = \epsilon_2 e^{-KT}$

$$\|x(t)\| < \epsilon_2 \qquad \forall\ t \geq t^*_{k_0+k(\epsilon_{2d})}. \tag{9}$$

Since $t^*_{k_0} - t_0 < T$ and since $t^*_{k_0+k(\epsilon_{2d})} - t^*_{k_0} \leq k(\epsilon_{2d})T$, $t^*_{k_0+k(\epsilon_{2d})} - t_0 < (k(\epsilon_{2d})+1)T$. Therefore $\|x(t)\| < \epsilon_2 \ \forall t \geq t_0 + T(\epsilon_2)$ with $T(\epsilon_2) := (k(\epsilon_2 e^{-KT}) + 1)T$. This implies uniform convergence to the origin and therefore also uniform asymptotic stability of (3). □

Remark 3. The proof of Proposition 2 shows that when $\overline{B}_\epsilon(0) \subset U'$ that $B_{\delta(\epsilon)}(0)$ with $\delta(\epsilon) = e^{-KT}\beta_V^{-1}(\alpha_V(\epsilon e^{-KT}))$ belongs to the region of attraction of (3).

Remark 4. If Condition 2 of Proposition 2 is replaced by the condition that there exists an increasing sequence of times t^*_k $(k \in \mathbb{Z})$ with $t^*_k \to \infty$ as $k \to \infty$ and $t^*_k \to -\infty$ as $k \to -\infty$, $\exists$ finite $T > 0 : t^*_{k+1} - t^*_k \leq T$ $(\forall k \in \mathbb{Z})$, there exists an open set $U' \subset U \cap W_c$ which contains the origin such that $\forall k \in \mathbb{Z}\ \forall x(t^*_k) \in U' \setminus \{0\}$,

$$V(x(t^*_{k+1}), t^*_{k+1}) - V(x(t^*_k), t^*_k) \leq 0,$$

then the equilibrium point $x = 0$ of (3) is locally uniformly stable.

4. Homogeneous systems. Proposition 2 introduces a sufficient condition for uniform asymptotic stability of a dynamical system. Because of Condition 2, it may be hard in general to verify uniform asymptotic stability by means of this proposition. This section and section 5 may be seen as an elaboration of the previous one. When considering homogeneous systems, we show that Condition 2 of Proposition 2 may be replaced by a condition *independent of* the flow.

DEFINITION 1. *The system $\dot{x}(t) = f(x,t)$ with $x = (x_1, ..., x_n)^T$ is homogeneous of order τ and with dilation $\delta(s,x) = (s^{r_1}x_1, ..., s^{r_n}x_n)^T$ $(\forall i \in \{1,...,n\} : 0 < r_i < \infty)$ if for each $i \in \{1,...,n\}$*

$$\forall x \in \mathbb{R}^n, \forall t, \forall s \geq 0 : f_i(s^{r_1}x_1, ..., s^{r_n}x_n, t) = s^{\tau+r_i} f_i(x_1, ..., x_n, t). \tag{10}$$

DEFINITION 2. *The homogeneous p-norm ρ_p with dilation $\delta(s,x)$ is a continuous map from $\mathbb{R}^n$ to $\mathbb{R}^+$, $x \to \rho_p(x)$ such that*

$$\rho_p(x) := \left(\sum_{i=1}^n |x_i|^{\frac{p}{r_i}}\right)^{\frac{1}{p}} \tag{11}$$

with $p \in \mathbb{R}^+_0$.

Remark 5. Calling the function ρ_p a "norm" is a misnomer. In general ρ_p does not satisfy the triangle inequality or the scale property.

PROPOSITION 3. *Consider the homogeneous system $\dot{x}(t) = f(x,t)$ of order $\tau > 0$ and with dilation $\delta(s,x) = (s^{r_1}x_1, ..., s^{r_n}x_n)^T$. f is locally Lipschitz, i.e., $\forall x$, $\exists$ neighborhood $\mathcal{N}(x)$ such that the restriction $f|_{\mathcal{N}(x)}$ is Lipschitz with Lipschitz function $l_x(t)$ and $l_x(t)$ is bounded $\forall t \in \mathbb{R}$.*

If there exists an increasing sequence of times t^*_k $(k \in \mathbb{Z})$ with $t^*_k \to \infty$ as $k \to \infty$ and $t^*_k \to -\infty$ as $k \to -\infty$, $\exists$ finite $T > 0 : t^*_{k+1} - t^*_k \leq T$ $(\forall k \in \mathbb{Z})$ and $\exists$ $K_1 > 0$ such that $\forall k$ and $\forall x$ with $\rho_r(x) = 1$ and $r > \max\{r_1, ..., r_n\}$

$$\frac{\partial V}{\partial x}(x) \int_{t^*_k}^{t^*_{k+1}} f(x,t)dt \leq -K_1 T, \tag{12}$$

where

1. *$V(x)$ is a positive definite continuous homogeneous function, i.e.,*

$$\forall x \in \mathbb{R}^n,\ \forall s \geq 0 :\ V(s^{r_1}x_1, ..., s^{r_n}x_n) = s^l V(x_1, ..., x_n) \tag{13}$$

for some $l > 0$, and

2. $\frac{\partial V}{\partial x}(x) := (\frac{\partial V}{\partial x_1}(x), ..., \frac{\partial V}{\partial x_n}(x))$ *is locally Lipschitz on* $\mathbb{R}^n$;
then $\dot{x} = f(x,t)$ *is locally uniformly asymptotically stable.*

Proof. Before starting the main part of the proof, we introduce some notation and we derive a number of inequalities to be used later on in the proof.

Define $S_\beta := \{x \in \mathbb{R}^n : \ |x_i| \leq \beta, \forall i \in \{1, ..., n\}\}$ $(\beta \in \mathbb{R}_0^+)$. If we use the norm $\|x\|_{\max} = \max_{1\leq i\leq n} |x_i|$, then $\forall x \in S_\beta : \ \|x\|_{\max} \leq \beta$. The max-norm is denoted as $\|\cdot\|_{\max}$ whereas the Euclidean norm is denoted as $\|\cdot\|$.

By the local Lipschitz property of $\dot{x} = f(x,t)$, it is Lipschitz on S_β with a Lipschitz constant which we denote as $K_{f\beta}$. Therefore, $\forall x, y \in S_\beta$ and $\forall\ i \in \{1, ..., n\}$

$$|f_i(x,t) - f_i(y,t)| \leq \|f(x,t) - f(y,t)\|_{\max} \leq K_{f\beta}\|x - y\|_{\max}. \tag{14}$$

The set $\{x \in \mathbb{R}^n | \rho_r(x) = 1\} \subset S_1$ and therefore if $i \in \{1, ..., n\}$ and $\rho_r(x) = 1$, then

$$|f_i(x,t)| \leq \|f(x,t)\|_{\max} \leq K_f\|x\|_{\max} \leq K_f \tag{15}$$

with $K_f := K_{f1}$. By the same argument, we obtain that $\forall x, y \in S_\beta$

$$\left\|\frac{\partial V}{\partial x}(x) - \frac{\partial V}{\partial x}(y)\right\|_{\max} \leq K_{V\beta}\|x - y\|_{\max}, \tag{16}$$

where $K_{V\beta}$ is the Lipschitz constant of $\frac{\partial V}{\partial x}$ on S_β. When $\rho_r(x) = 1$, then

$$\left\|\frac{\partial V}{\partial x}(x)\right\|_{\max} \leq K_V \tag{17}$$

with $K_V := K_{V1}$. The estimates (14), (15), (16), and (17) will be used in the following.

I. Evolution of the *V*-function with respect to the sequence t_k^*. The time-derivative of $V(x)$ along the solutions of the system $\dot{x}(t) = f(x,t)$ is given by

$$\dot{V}(x,t) = \frac{\partial V}{\partial x}(x)\dot{x} = \frac{\partial V}{\partial x}(x)f(x,t). \tag{18}$$

Consider

$$\Delta V(t_{k+1}^*, t_k^*) := \int_{t_k^*}^{t_{k+1}^*} \dot{V}(x,t)dt = \int_{t_k^*}^{t_{k+1}^*} \frac{\partial V}{\partial x}(x(t))f(x(t),t)dt, \tag{19}$$

which may be rewritten as

$$\int_{t_k^*}^{t_{k+1}^*} \frac{\partial V}{\partial x}(x(t_k^*))f(x(t_k^*),t)dt \tag{20}$$

$$+ \int_{t_k^*}^{t_{k+1}^*} \left(\frac{\partial V}{\partial x}(x(t))f(x(t),t) - \frac{\partial V}{\partial x}(x(t_k^*))f(x(t_k^*),t)\right) dt. \tag{21}$$

In order to evaluate (19), we successively estimate (20) and (21).

II. Estimate of (20). In order to invoke the homogeneity properties of $\dot{x} = f(x,t)$, we spell out (20) as

$$\sum_{i=1}^{n} \int_{t_k^*}^{t_{k+1}^*} \frac{\partial V}{\partial x_i}(x(t_k^*))f_i(x(t_k^*),t)dt. \tag{22}$$

By partially differentiating each member of (13) with respect to x_i, one obtains that $\forall x \in \mathbb{R}^n \setminus \{0\}$, $\forall i \in \{1, ..., n\}$, and $\forall s \geq 0$

$$\frac{\partial V}{\partial x_i}(s^{r_1}x_1, ..., s^{r_n}x_n) = s^{l-r_i}\frac{\partial V}{\partial x_i}(x_1, ..., x_n). \tag{23}$$

By (23) and the homogeneity of $f(x,t)$, one obtains that (22) can be written as

$$\rho_r^{\tau+l}(x(t_k^*))\int_{t_k^*}^{t_{k+1}^*}\sum_{i=1}^{n}\frac{\partial V}{\partial x_i}(\delta(\rho_r^{-1}(x(t_k^*)), x(t_k^*)))f_i(\delta(\rho_r^{-1}(x(t_k^*)), x(t_k^*)), t)dt \tag{24}$$

or as

$$\rho_r^{\tau+l}(x(t_k^*))\frac{\partial V}{\partial x}(\delta(\rho_r^{-1}(x(t_k^*)), x(t_k^*)))\int_{t_k^*}^{t_{k+1}^*} f(\delta(\rho_r^{-1}(x(t_k^*)), x(t_k^*)), t)dt. \tag{25}$$

Since $\delta(\rho_r^{-1}(x(t_k^*)), x(t_k^*))$ has a homogeneous norm equal to 1, (25) implies by (12) that

$$\int_{t_k^*}^{t_{k+1}^*}\frac{\partial V}{\partial x}(x(t_k^*))f(x(t_k^*), t)dt \leq -K_1 T\rho_r^{\tau+l}(x(t_k^*)). \tag{26}$$

The inequality (26) will force $\Delta V(t_{k+1}^*, t_k^*)$ to be negative definite. We show that (21) cannot account for a sign change when the initial state $x(t_k^*)$ is taken close enough to the origin. To prove this, we estimate an upper bound for the absolute value of (21).

III. Estimate of (21). By (23) and the homogeneity of $f(x,t)$, (21) can be written as

$$\rho_r^{\tau+l}(x(t_k^*))\int_{t_k^*}^{t_{k+1}^*} L_f V(\delta(\rho_r^{-1}(x(t_k^*)), x(t)), t) - L_f V(\delta(\rho_r^{-1}(x(t_k^*)), x(t_k^*)), t)dt, \tag{27}$$

where $\frac{\partial V}{\partial x}(x)f(x,t) =: L_f V(x,t)$.

In order to evaluate an upper bound for the absolute value of (27), we need the max-norm of

$$\delta(\rho_r^{-1}(x(t_k^*)), x(t)) - \delta(\rho_r^{-1}(x(t_k^*)), x(t_k^*)) = \delta(\rho_r^{-1}(x(t_k^*)), x(t) - x(t_k^*)) \tag{28}$$

since we use the Lipschitz property of $L_f V$. In III.1 we estimate the norm of (28) and in III.2 we estimate the Lipschitz constant corresponding to $L_f V$. In III.3 we use these results to estimate the absolute value of (21).

III.1. Estimate of the norm of (28). Since

$$x_i(t) - x_i(t_k^*) = \int_{t_k^*}^{t} f_i(x(s), s)ds, \tag{29}$$

one obtains by the homogeneity of $f(x,s)$ and by (15) that

$$|x_i(t) - x_i(t_k^*)| \leq \int_{t_k^*}^{t}\rho_r^{\tau+r_i}(x(s))K_f ds. \tag{30}$$

Since $\frac{\partial \rho_r}{\partial x_i}(s^{r_1}x_1, ..., s^{r_n}x_n) = s^{1-r_i}\frac{\partial \rho_r}{\partial x_i}(x_1, ..., x_n)$ and

$$\begin{aligned}\frac{d}{d\sigma}\rho_r(x(\sigma)) &= \sum\nolimits_{i=1}^{n}\frac{\partial \rho_r}{\partial x_i}(x(\sigma))f_i(x(\sigma), \sigma)\\ &= \rho_r^{\tau+1}(x(\sigma))\sum\nolimits_{i=1}^{n}\frac{\partial \rho_r}{\partial x_i}(y(\sigma))f_i(y(\sigma), \sigma)\end{aligned} \tag{31}$$

with $y(\sigma) = \delta(\rho_r^{-1}(x(\sigma)), x(\sigma))$. By defining $g(\sigma) := \sum_{i=1}^n \frac{\partial \rho_r}{\partial x_i}(y(\sigma)) f_i(y(\sigma), \sigma)$,

$$\frac{d}{d\sigma}\rho_r(x(\sigma)) = \rho_r^{\tau+1}(x(\sigma)) g(\sigma). \tag{32}$$

By direct substitution [6], it is clear that the solution of (32) equals

$$\rho_r(x(s)) = \frac{\rho_r(x(t_k^*))}{(1 - \tau\rho_r^\tau(x(t_k^*)) \int_{t_k^*}^s g(\sigma) d\sigma)^{\frac{1}{\tau}}} \tag{33}$$

under the assumption that

$$1 - \tau\rho_r^\tau(x(t_k^*)) \int_{t_k^*}^s g(\sigma) d\sigma > 0. \tag{34}$$

Since $r_i < r$, $\frac{\partial \rho_r}{\partial x_i}$ is continuous on the set $\{x \in \mathbb{R}^n | \rho_r(x) = 1\}$. Since $y(\sigma)$ belongs to this compact set ($\rho_r(y(\sigma)) = 1$), the continuity of $\frac{\partial \rho_r}{\partial x_i}$ and (15) imply boundedness of $g(\sigma)$. There exists a $g_m > 0$ such that $\forall \sigma$: $g(\sigma) \leq g_m$. This implies that

$$\rho_r(x(s)) \leq \frac{\rho_r(x(t_k^*))}{(1 - \tau\rho_r^\tau(x(t_k^*))(s - t_k^*) g_m)^{\frac{1}{\tau}}} < 2^{\frac{1}{\tau}} \rho_r(x(t_k^*)) \tag{35}$$

when $t \in [t_k^*, t_{k+1}^*]$ with $t_{k+1}^* - t_k^* \leq T$ $\forall k \in \mathbb{Z}$ and $\rho_r(x(t_k^*)) < (2\tau g_m T)^{-\frac{1}{\tau}} =: \rho'$. By (35), it is obvious that (30) implies that

$$|x_i(t) - x_i(t_k^*)| \leq 2^{\frac{\tau + r_i}{\tau}} K_f T \rho_r^{\tau + r_i}(x(t_k^*)). \tag{36}$$

Recall that

$$\|\delta(\rho_r^{-1}(x(t_k^*)), x(t)) - \delta(\rho_r^{-1}(x(t_k^*)), x(t_k^*))\|_{\max} = \max_{1 \leq i \leq n} \frac{|x_i(t) - x_i(t_k^*)|}{\rho_r^{r_i}(x(t_k^*))}. \tag{37}$$

Therefore, (36) and (37) imply the existence of a $\tilde{K} > 0$ such that

$$\|\delta(\rho_r^{-1}(x(t_k^*)), x(t)) - \delta(\rho_r^{-1}(x(t_k^*)), x(t_k^*))\|_{\max} \leq \tilde{K} T \rho_r^\tau(x(t_k^*)) \tag{38}$$

when $\rho_r(x(t_k^*)) < \rho'$ and $t \in [t_k^*, t_{k+1}^*]$.

Having estimated the norm of (28), we estimate the Lipschitz constant of $L_f V$ which will be used in the calculation of an upper bound for (21).

III.2. Estimate of the Lipschitz constant. Notice that

$$\begin{aligned} \delta(\rho_r^{-1}(x(t_k^*)), x(t)) = {} & \delta(\rho_r^{-1}(x(t_k^*)), x(t_k^*)) \\ & + \delta(\rho_r^{-1}(x(t_k^*)), x(t)) - \delta(\rho_r^{-1}(x(t_k^*)), x(t_k^*)) \end{aligned} \tag{39}$$

and

$$\begin{aligned} \|\delta(\rho_r^{-1}(x(t_k^*)), x(t))\|_{\max} \leq {} & \|\delta(\rho_r^{-1}(x(t_k^*)), x(t_k^*))\|_{\max} \\ & + \|\delta(\rho_r^{-1}(x(t_k^*)), x(t)) - \delta(\rho_r^{-1}(x(t_k^*)), x(t_k^*))\|_{\max} \end{aligned} \tag{40}$$

such that by (38)

$$\|\delta(\rho_r^{-1}(x(t_k^*)), x(t))\|_{\max} \leq 1 + \tilde{K} T \rho_r^\tau(x(t_k^*)) \leq 1 + \frac{\tilde{K}}{2\tau g_m} =: \tilde{\beta} \tag{41}$$

when $t \in [t_k^*, t_{k+1}^*]$ and $\rho_r(x(t_k^*)) < \rho' = (2\tau g_m T)^{-\frac{1}{\tau}}$. Therefore, (41) implies that $\delta(\rho_r^{-1}(x(t_k^*)), x(t)) \in S_{\tilde{\beta}}$.

Since $\forall x, y \in S_{\tilde{\beta}}$ and $\forall t$

$$\begin{aligned} L_f V(x,t) - & L_f V(y,t) \\ & = \frac{\partial V}{\partial x}(x)(f(x,t) - f(y,t)) + \left(\frac{\partial V}{\partial x}(x) - \frac{\partial V}{\partial x}(y)\right) f(y,t), \end{aligned} \tag{42}$$

the boundedness of $\frac{\partial V}{\partial x}$ and f on $S_{\tilde{\beta}}$, implied by (14) and (16), and the Lipschitz properties of $\frac{\partial V}{\partial x}$ and f imply the existence of a Lipschitz constant $K_{fV\tilde{\beta}}$ for $L_f V$ on $S_{\tilde{\beta}}$. Therefore, $\forall x, y \in S_{\tilde{\beta}}$, and $\forall t$

$$|L_f V(x,t) - L_f V(y,t)| \le K_{fV\tilde{\beta}} \|x - y\|_{\max}. \tag{43}$$

III.3. Estimate of (21). By (38) and (43), one obtains that

$$\begin{aligned} |L_f V(\delta(\rho_r^{-1}(x(t_k^*)), x(t)), t) - & L_f V(\delta(\rho_r^{-1}(x(t_k^*)), x(t_k^*)), t)| \\ & \le K_{fV\tilde{\beta}} \tilde{K} T \rho_r^{\tau}(x(t_k^*)). \end{aligned} \tag{44}$$

By (44), the absolute value of (27)–(21) is less than or equal to

$$K_{fV\tilde{\beta}} \tilde{K} T^2 \rho_r^{2\tau + l}(x(t_k^*)) \tag{45}$$

when $\rho_r(x(t_k^*)) < \rho'$.

IV. Estimate of (19). By (26) and (45), (19) is less than or equal to

$$\rho_r^{\tau + l}(x(t_k^*)) T(-K_1 + \tilde{K} K_{fV\tilde{\beta}} T \rho_r^{\tau}(x(t_k^*))). \tag{46}$$

Define $\rho := \min\{\rho', (\frac{K_1}{2K_{fV\tilde{\beta}}\tilde{K}T})^{\frac{1}{\tau}}\}$. This implies by (46) that $\forall x(t_k^*) \neq 0$ with $\rho_r(x(t_k^*)) < \rho$:

$$\Delta V(t_{k+1}^*, t_k^*) = V(x(t_{k+1}^*)) - V(x(t_k^*)) \le -\frac{K_1 T}{2} \rho_r^{\tau + l}(x(t_k^*)), \tag{47}$$

where $x(t_{k+1}^*)$ is the solution of $\dot{x}(t) = f(x,t)$ at t_{k+1}^* with initial condition $x(t_k^*)$ at t_k^*.

V. Uniform asymptotic stability. By (13), it is obvious that Condition 1 of Proposition 2 is fulfilled with $U = \mathbb{R}^n$. By (47), it is clear that Condition 2 of Proposition 2 is fulfilled with $U' = \{x | \rho_r(x) < \rho\}$. Therefore, Proposition 2 may be applied, which implies local uniform asymptotic stability of the homogeneous system $\dot{x} = f(x,t)$. □

Remark 6. Notice that r is taken to be larger than $\max\{r_1, ..., r_n\}$ (and not equal to $\max\{r_1, ..., r_n\}$) in order to avoid technical difficulties when taking the derivative of $\sum_{i=1}^n |y_i|^{\frac{r}{r_i}}$ with respect to time.

5. Uniform asymptotic stability of time-varying homogeneous systems. Having Proposition 3 available, it is now possible to establish that asymptotic stability of the averaged system of a time-varying homogeneous system implies local uniform asymptotic stability of the original time-varying homogeneous system. Because of the homogeneity and the order condition $\tau > 0$, this result is valid even when the system is not fast time-varying.

THEOREM 1. *Consider the homogeneous system $\dot{x}(t) = f(x,t)$ of order $\tau > 0$ and with dilation $\delta(s,x) = (s^{r_1}x_1, ..., s^{r_n}x_n)^T$. f is locally Lipschitz, i.e., $\forall x$, $\exists$ neighborhood $\mathcal{N}(x)$ such that the restriction $f|_{\mathcal{N}(x)}$ is Lipschitz with Lipschitz function $l_x(t)$ and $l_x(t)$ is bounded $\forall t \in \mathbb{R}$. If the following conditions hold:*

Condition 1. *The averaged system $\dot{x}(t) = \bar{f}(x)$ is asymptotically stable, where*

$$\bar{f}(x) := \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} f(x,t)dt \tag{48}$$

is continuous in x;

Condition 2. *There exists a continuous function $M : [0, +\infty[\to [0, +\infty[$ with $\lim_{\sigma\to\infty} \sigma^{-1}M(\sigma) = 0$ such that $\forall t_1, t_2 \in \mathbb{R}$ $(t_2 > t_1)$*

$$\left\| \int_{t_1}^{t_2} (f(x,t) - \bar{f}(x))dt \right\| \leq M(t_2 - t_1) \tag{49}$$

when $\rho_r(x) = 1$ with $r > \max\{r_1, ..., r_n\}$.

Then $\dot{x}(t) = f(x,t)$ is locally uniformly asymptotically stable.

Proof. The proof of the theorem is based on Proposition 3. Definition (48) of the averaged system $\dot{x}(t) = \bar{f}(x)$ implies its homogeneity of order τ with dilation $\delta(s,x) = (s^{r_1}x_1, ..., s^{r_n}x_n)^T$. Definition (48) also implies that $\bar{f}(0) = 0$. By Condition 1, the homogeneous system $\dot{x} = \bar{f}(x)$ is asymptotically stable. Let p be a positive integer and let l be a real number larger than $p(\max_{1\leq i\leq n} r_i)$. Following Rosier [13], there exists a Liapunov function $V : \mathbb{R}^n \to \mathbb{R}$ such that

P1. $V(x)$ is of class C^∞ in $\mathbb{R}^n \setminus \{0\}$ and of class C^p in $\mathbb{R}^n$;

P2. $V(0) = 0$, $V(x) > 0$ $\forall$ $x \neq 0$ and $V(x) \to +\infty$ as $\|x\| \to +\infty$;

P3. V is homogeneous: $\forall x \in \mathbb{R}^n \setminus \{0\} : \forall s > 0 : V(s^{r_1}x_1, ..., s^{r_n}x_n) = s^l V(x_1, ..., x_n)$;

P4. $\forall x \neq 0 : \frac{\partial V}{\partial x}(x)\bar{f}(x) < 0$.

By P1, $\frac{\partial V}{\partial x}$ is continuous in $\mathbb{R}^n$ and the continuity of $\bar{f}(x)$ implies continuity of $\frac{\partial V}{\partial x}(x)\bar{f}(x)$ on the compact set $\{x|\rho_r(x) = 1\}$. This implies by P4 that $\exists\beta > 0$ such that $\forall x$ with $\rho_r(x) = 1$,

$$\frac{\partial V}{\partial x}(x)\bar{f}(x) \leq -\beta. \tag{50}$$

Take a $t_0^\circ \in \mathbb{R}$ and $\forall T'$ and $\forall x$ with $\rho_r(x) = 1$:

$$\frac{\partial V}{\partial x}(x) \int_{t_0^\circ}^{t_0^\circ+T'} f(x,t)dt = T'\frac{\partial V}{\partial x}(x)\bar{f}(x) + \frac{\partial V}{\partial x}(x) \int_{t_0^\circ}^{t_0^\circ+T'} (f(x,t) - \bar{f}(x))dt. \tag{51}$$

By the continuity of $\frac{\partial V}{\partial x}$ on the compact set $\{x|\rho_r(x) = 1\}$, $\exists M_V > 0$ such that

$$\left\| \frac{\partial V}{\partial x}(x) \right\| \leq M_V \tag{52}$$

when $\rho_r(x) = 1$. One obtains by Condition 2 that

$$\frac{\partial V}{\partial x}(x) \int_{t_0^\circ}^{t_0^\circ+T'} f(x,t)dt \leq T'\frac{\partial V}{\partial x}(x)\bar{f}(x) + M_V M(T') \tag{53}$$

$$\leq -\beta T' + M_V M(T') \tag{54}$$

when $\rho_r(x) = 1$. By Condition 2, $\lim_{T' \to \infty} \frac{M(T')}{T'} = 0$, which implies that $\exists T''$ such that $\forall T' \geq T'' : \frac{M(T')}{T'} < \frac{\beta}{M_V}$. Take such a T' and define a sequence of times $t_k^* := t_0^\circ + kT'$, then $\forall k \in \mathbb{Z}$ and $\forall x$ with $\rho_r(x) = 1$,

$$\frac{\partial V}{\partial x}(x) \int_{t_k^*}^{t_{k+1}^*} f(x,t)dt \leq -K_1 T' \tag{55}$$

with $K_1 = \beta - M_V \frac{M(T')}{T'} > 0$.

With $T = T'$ and $t_k^* = t_0^\circ + kT'$, (55) implies that (12) is satisfied and P3 implies that (13) is satisfied. With $p \geq 2$, $\frac{\partial V}{\partial x}$ is continuously differentiable and therefore also locally Lipschitz. By Proposition 3, one obtains local uniform asymptotic stability of the homogeneous system $\dot{x} = f(x,t)$. □

Remark 7. Theorem 1 is a result on asymptotic stability of a homogeneous system under a classical averaging condition, *without* requirements on the time-scale—typical for averaging results.

Remark 8. The proof of Theorem 1 is based on Proposition 3 and the crucial part of the proof of Proposition 3 is the negative definiteness of (19). This negative definiteness of (19) is guaranteed by the negative definiteness of (20) when (21) is sufficiently small. The expression (21) is small when, roughly, the variation of the flow $x(t) - x(t_k^*)$ is small in comparison with $x(t_k^*)$. This proportional variation can be reduced by reducing the size of the right-hand side $\rho_r^\tau(x(t_k^*))\tilde{K}T$ of (38). A first way of achieving this reduction is by a time-scale transformation. It is clear that when the homogeneous system $\dot{x} = f(x,t)$ satisfies the conditions of Theorem 1 with $T = T_f$, then $\forall \alpha > 0$: $\dot{x} = f(x, \alpha t)$ satisfies the conditions of Theorem 1 with $T = \frac{T_f}{\alpha}$. By increasing α, i.e., by decreasing $T = \frac{T_f}{\alpha}$, $\rho_r^\tau(x(t_k^*))\tilde{K}T$ can be made arbitrarily small. For any fixed time-scale, there is still a second possibility of reducing the size of $\rho_r^\tau(x(t_k^*))\tilde{K}T$ by reducing $\rho_r(x(t_k^*))$, i.e., by starting at initial conditions close enough to the origin. This is the technique that leads to Theorem 1.

In [4], where exponential stability is considered, and in [11], where homogeneous systems of order $\tau = 0$ are considered, there is a different situation. The effect of a decreasing proportional variation of the flow when starting close enough to the origin is not available. The only way to obtain a small proportional variation is by making α sufficiently large. This explains why the averaging results in [4] and [11] imply uniform asymptotic stability of *fast* time-varying systems.

6. A counterexample and an example.

The averaging result of Theorem 1, for not necessarily fast time-varying systems, is valid for homogeneous systems of order $\tau > 0$. As explained in Remark 8, the averaging result is not valid for homogeneous systems with an order $\tau = 0$.

This is easily illustrated. Consider the linear time-varying system $\dot{x}(t) = A(t)x(t)$ [10, p. 144] with

$$A(t) = \begin{pmatrix} -1 + 1.5\,\cos^2 t & 1 - 1.5\,\sin t\,\cos t \\ -1 - 1.5\,\sin t\,\cos t & -1 + 1.5\,\sin^2 t \end{pmatrix}. \tag{56}$$

The averaged system $\dot{x}(t) = \bar{A}x(t)$ with

$$\bar{A} = \begin{pmatrix} -0.25 & 1 \\ -1 & -0.25 \end{pmatrix} \tag{57}$$

is asymptotically stable but the original time-varying system $\dot{x}(t) = A(t)x(t)$ is unstable with transition matrix

$$\Phi(t,0) = \begin{pmatrix} e^{0.5t}\text{cos}t & e^{-t}\text{sin}t \\ -e^{0.5t}\text{sin}t & e^{-t}\text{cos}t \end{pmatrix}. \tag{58}$$

Since $\bar{A}$ is Hurwitz, there exists a positive definite matrix P such that $\bar{A}^TP + P\bar{A}$ is negative definite. Consider the homogeneous system $\dot{x}(t) = \|x(t)\|A(t)x(t)$ with positive order. The averaged system $\dot{x}(t) = \|x(t)\|\bar{A}x(t)$ is asymptotically stable since along its flow the derivative of x^TPx equals $\|x\|x^T(\bar{A}^TP + P\bar{A})x$, which is negative definite. By Theorem 1, asymptotic stability of $\dot{x}(t) = \|x(t)\|\bar{A}x(t)$ implies local uniform asymptotic stability of $\dot{x}(t) = \|x(t)\|A(t)x(t)$.

7. Semiglobal uniform asymptotic stability. The conditions of Theorem 1 imply local uniform asymptotic stability of the homogeneous time-varying system $\dot{x} = f(x,t)$ with order $\tau > 0$. These conditions also imply that $\forall\alpha > 0$: $\dot{x} = f(x,\alpha t)$ is locally uniformly asymptotically stable. No global stability property is obtained. In the present section, we prove that the region of attraction of the homogeneous system $\dot{x} = f(x,\alpha t)$ depends on α. We show that the bounded region of attraction increases when α increases and grows unbounded as α goes to infinity.

THEOREM 2. *The homogeneous system $\dot{x} = f(x,\alpha t)$ of order $\tau > 0$, where $\dot{x} = f(x,t)$ satisfies all the conditions of Theorem 1, is semiglobally uniformly asymptotically stable, i.e., $\forall R > 0$, $\exists\alpha_R > 0$, and also a class $\mathcal{KL}$-function $\beta_R(\cdot,\cdot)$ such that $\forall x_0$ with $\rho_r(x_0) < R$, $\forall t_0$, $\forall t \geq t_0$*

$$\rho_r(x_{\alpha_R}(t,t_0,x_0)) \leq \beta_R(\rho_r(x_0), t - t_0). \tag{59}$$

Here $x_{\alpha_R}(t,t_0,x_0)$ is the solution at t of the homogeneous system $\dot{x} = f(x,\alpha_R t)$ with initial condition x_0 at t_0.

Proof. The conditions imposed by Theorem 1 on $\dot{x} = f(x,t)$ are satisfied, which implies local uniform asymptotic stability of $\dot{x} = f(x,t)$. This is equivalent to the existence of a $\rho > 0$ and a class $\mathcal{KL}$-function $\beta_\rho(\cdot,\cdot)$ such that $\forall t_0$, $\forall t \geq t_0$, $\forall x_0$ with $\rho_r(x_0) < \rho$

$$\rho_r(x_1(t,t_0,x_0)) \leq \beta_\rho(\rho_r(x_0), t - t_0), \tag{60}$$

where $x_1(t,t_0,x_0)$ denotes the solution at t of $\dot{x} = f(x,t)$ with initial condition x_0 at t_0. We denote the solution of $\dot{x} = f(x,\alpha t)$ with initial condition x_0 at t_0 as $x_\alpha(t,t_0,x_0)$. The solutions of $\dot{x} = f(x,t)$ and $\dot{x} = f(x,\alpha t)$ are related, i.e., $\forall\alpha > 0$, $\forall t_0$, and $\forall t \geq t_0$

$$x_\alpha(t,t_0,x_0) = \delta(\sqrt[\tau]{\alpha}, x_1(\alpha t, \alpha t_0, \delta^{-1}(\sqrt[\tau]{\alpha}, x_0))). \tag{61}$$

Define $\alpha_R := (\frac{R}{\rho})^\tau$. For all x_0 with $\rho_r(x_0) < R$, $\forall t_0$, $\forall t \geq t_0$, by (61)

$$\rho_r(x_{\alpha_R}(t,t_0,x_0)) = \sqrt[\tau]{\alpha_R}\ \rho_r(x_1(\alpha_R t, \alpha_R t_0, \delta^{-1}(\sqrt[\tau]{\alpha_R}, x_0))). \tag{62}$$

Since $\rho_r(\delta^{-1}(\sqrt[\tau]{\alpha_R}, x_0)) < \rho$, applying (60) to the right-hand side of (62) implies that for all x_0 with $\rho_r(x_0) < R$, $\forall t_0$, $\forall t \geq t_0$

$$\rho_r(x_{\alpha_R}(t,t_0,x_0)) \leq \sqrt[\tau]{\alpha_R}\beta_\rho(\rho_r(\delta^{-1}(\sqrt[\tau]{\alpha_R}, x_0)), \alpha_R(t - t_0)) = \beta_R(\rho_r(x_0), t - t_0) \tag{63}$$

with the obvious definition of $\beta_R(\cdot,\cdot)$. □

8. Conclusions.

Proposition 2 gives a sufficient condition for uniform asymptotic stability of a differential equation. This result is related to the result of Narendra and Annaswamy [12], but negative semidefiniteness on $\dot{V}(x,t)$ is dispensed with.

This result is useful for the investigation of uniform asymptotic stability of homogeneous sytems with order $\tau > 0$. More precisely, averaging becomes a useful tool for studying uniform asymptotic stability: asymptotic stability of the averaged system implies local uniform asymptotic stability of the original time-varying system. It is important that this result is *not restricted to fast time-varying systems.* The region of attraction of $\dot{x} = f(x, \alpha t)$ increases with increasing α. The uniform asymptotic stability is semiglobal since by taking α large enough, every bounded region of attraction can be guaranteed.

Comparing these results with the results of M'Closkey and Murray [11], the following should be noted:

1. We are dealing with homogeneous systems of order $\tau > 0$ while M'Closkey and Murray are dealing with homogeneous systems of order $\tau = 0$.

2. Because of the order $\tau > 0$, asymptotic stability of the averaged system implies asymptotic stability, not exponential stability, of the original time-varying system. M'Closkey and Murray consider homogeneous systems of order $\tau = 0$ and therefore they are able to conclude exponential stability, with respect to the homogeneous norm, of the original time-varying system.

3. We obtain local asymptotic stability results for the homogeneous system $\dot{x} = f(x, \alpha t)$ with order $\tau > 0$ for every $\alpha > 0$. By setting $\sigma = \alpha t$ and $\epsilon = \frac{1}{\alpha}$, $\dot{x} = f(x, \alpha t)$ is equivalent to $\dot{x} = \epsilon f(x, \sigma)$. M'Closkey and Murray deal with homogeneous systems $\dot{x} = \epsilon f(x, \sigma, \epsilon)$ of order $\tau = 0$ and therefore, the stability results are valid only for ϵ sufficiently small.

Acknowledgments. The authors acknowledge the constructive comments of the anonymous reviewers.

REFERENCES

[1] D. Aeyels, *Asymptotic stability of nonautonomous systems by Liapunov's direct method*, Systems Control Lett., 25 (1995), pp. 273–280.

[2] D. Aeyels and R. Sepulchre, *On the convergence of a time-variant linear differential equation arising in identification*, Kybernetika, 30 (1994), pp. 715–723.

[3] D. Aeyels and J. Peuteman, *A new asymptotic stability criterion for nonlinear time-variant differential equations*, IEEE Trans. Automat. Control, 43 (1998), pp. 968–971.

[4] D. Aeyels and J. Peuteman, *On exponential stability of nonlinear time-variant differential equations*, Automatica, to appear.

[5] W. Hahn, *On the application of the method of Lyapunov to difference equations*, Math. Ann., 136 (1958), pp. 430–441 (in German).

[6] W. Hahn, *Stability of Motion*, Springer-Verlag, New York, 1967.

[7] H. Hermes, *Asymptotic stabilization via homogeneous approximation*, in Geometry of Feedback and Optimal Control, B. Jakubczyk and W. Respondek, eds., Marcel Dekker, New York, 1998, pp. 205–218.

[8] R.E. Kalman and J.E. Bertram, *Control system analysis and design via the "second method" of Lyapunov.* II *Discrete-time systems*, Trans. ASME Ser. D. J. Basic Engrg., 82 (1960), pp. 394–400.

[9] M. Kawski, *Families of dilations and asymptotic stability*, in Analysis of Controlled Dynamical Systems, Progr. Systems Control Theory 8, B. Bonnard, B. Bride, J.P. Gauthier, and I. Kupka, eds., Birkhäuser, Boston, MA, 1991, pp. 285–294.

[10] H.K. Khalil, *Nonlinear Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1996.

[11] R.T. M'Closkey and R.M. Murray, *Nonholonomic systems and exponential convergence: Some analysis tools*, in Proceedings of the 32nd Conference on Decision and Control, San Antonio, TX, 1993, pp. 943–948.

[12] K.S. Narendra and A.M. Annaswamy, *Persistent excitation in adaptive systems*, Internat. J. Control, 45 (1987), pp. 127–160.

[13] L. Rosier, *Homogeneous Lyapunov function for homogeneous continuous vector fields*, Systems Control Lett., 19 (1992), pp. 467–473.

[14] J.L. Willems, *Stability Theory of Dynamical Systems*, Nelson, London, 1970.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 4, pp. 1011–1032

NECESSARY OPTIMALITY CONDITIONS FOR CONTROL PROBLEMS AND THE STONE–ČECH COMPACTIFICATION[*]

NADIR ARADA[†] AND JEAN-PIERRE RAYMOND[†]

Abstract. This paper deals with optimal control problems of parabolic equations in the presence of pointwise state constraints. We consider bounded controls which act in the initial condition of the state equation. The state variable is a bounded continuous function on the domain $Q = \Omega\times]0,T[$ but is not continuous on $\overline{Q}$. In this case, the multiplier associated with state constraints is a regular bounded finitely additive measure on Q (but not a σ-additive one). Using some properties of the Stone–Čech compactification, we prove a decomposition theorem for this measure which allows us to interpret the adjoint equation in a classical sense. We obtain new optimality conditions for these kinds of problems, and we apply these results to the case of bilateral constraints.

Key words. optimal control, Stone–Čech compactification, semilinear parabolic equation, pointwise state constraint

AMS subject classification. 49K20

PII. S036301299733035X

1. Introduction.

Consider an optimal control problem governed by a partial differential equation, with state constraints of the form $y \in \mathcal{C}$, where $\mathcal{C}$ is a closed convex subset of a Banach space Z. It is well known that some conditions are needed on $\mathcal{C}$ to establish first order optimality conditions. For example, $\mathcal{C}$ must be of finite codimension in Z (which is satisfied if the interior of $\mathcal{C}$ is nonempty in Z). In this case, the adjoint state satisfies a partial differential equation in which appears a multiplier ζ belonging to Z', the topological dual of Z. Different cases where Z is separable are treated in [3], [4], [5], [9], [11], [14], [16]. There are situations in which Z is not separable. In such problems, the main drawback is that the adjoint equation cannot be interpreted in the sense of distributions and classical methods for partial differential equations cannot be used (to prove existence and uniqueness, and to study approximation). In this paper, we use some properties of the Stone–Čech compactification to derive optimality conditions.

Since the use of these tools is completely new in the context of optimal control problems, we have chosen a simple problem to illustrate such situations: we consider a control problem for a semilinear parabolic equation with a control in the initial condition, the cost functional depends only on the final observation and on the control, and pointwise state constraints are imposed on the whole domain. The problem is simple because there is only one control and because the control acts only in the initial condition. However, the same method may be applied to other control problems. In particular, the Stone–Čech compactification may be used to obtain new optimality conditions for problems with Dirichlet boundary controls in the presence of pointwise state constraints. (Such problems are considered in [10], [12], [13].) Some results have been presented in [1], and a complete analysis will be carried out in a forthcoming paper [2].

[*] Received by the editors November 19, 1997; accepted for publication (in revised form) July 16, 1998; published electronically April 7, 1999.
http://www.siam.org/journals/sicon/37-4/33035.html

[†] UMR-CNRS 5640, Université Paul Sabatier, 31062 Toulouse Cedex 4, France (arada@mip.ups-tlse.fr, raymond@mip.ups-tlse.fr).

Consider the following semilinear parabolic equation:

$$(1)\qquad \frac{\partial y}{\partial t}+Ay+\Phi(\cdot,y)=0 \quad \text{in } Q,\qquad \frac{\partial y}{\partial n_A}+\Psi(\cdot,y)=0 \quad \text{on } \Sigma,\qquad y(\cdot,0)=w \quad \text{in } \Omega,$$

where $Q=\Omega\times]0,T[$, Ω is a bounded domain in $\mathbb{R}^N$, $\Sigma=\Gamma\times]0,T[$, Γ is the boundary of Ω, $T>0$, $w\in W_{ad}\subset L^\infty(\Omega)$ is a control in the initial condition, A is a second order elliptic operator of the form $Ay(x)=-\sum_{i,j=1}^N D_i(a_{ij}(x)D_jy(x))$ (D_i denotes the partial derivative with respect to x_i), $\frac{\partial y}{\partial n_A}$ denotes the conormal derivative with respect to A, and Φ and Ψ are Carathéodory functions (assumptions are specified in section 2.1). Constraints of the form

$$(2)\qquad g(y)\in\mathcal{C}$$

are imposed on the state variable y. (Here g is a continuous mapping from $C_b(\overline{Q}\setminus\overline{\Omega}_0)$ into $C_b(\overline{Q}\setminus\overline{\Omega}_0)$, $\mathcal{C}$ is a closed convex subset with a nonempty interior in $C_b(\overline{Q}\setminus\overline{\Omega}_0)$, and $\overline{\Omega}_0=\overline{\Omega}\times\{0\}$.) The paper is concerned with the control problem

$$(\mathrm{P})\qquad \inf\{J(y,w)\ \mid\ y\in C_b(\overline{Q}\setminus\overline{\Omega}_0),\ w\in W_{ad},\ \ (y,w) \text{ satisfies (1) and (2)}\},$$

where the cost functional is given by

$$J(y,w)=\int_\Omega L(x,y(x,T),w(x))\,dx.$$

Since the initial condition w belongs to $L^\infty(\Omega)$ (and not to $C(\overline{\Omega})$), the solution y_w of (1) belongs to $C_b(\overline{Q}\setminus\overline{\Omega}_0)$ (and not to $C(\overline{Q})$). Since $C_b(\overline{Q}\setminus\overline{\Omega}_0)$ is not separable, we are in the situation described at the beginning. One way to obtain optimality conditions is to proceed as in [12], [3]. We can associate with (P) a problem in which the state constraint (2) is penalized, we can write optimality conditions for the penalized problem, and, by passing to the limit with respect to the penalization parameter, we recover optimality conditions in which the adjoint state p obeys:

$$\int_Q\Big(-p\frac{\partial z}{\partial t}+\sum_{i,j=1}^N a_{ij}D_jpD_iz+\Phi'_y(x,t,\bar y)pz\Big)\,dxdt+\int_\Sigma\Psi'_y(s,t,\bar y)pz\,dsdt$$
$$+\int_\Omega\bar\alpha L'_y(x,\bar y(T),\bar w)z\,dx+\Big\langle g'(\bar y)^*\bar\zeta,z\Big\rangle_{(C_b(\overline{Q}\setminus\overline{\Omega}_0))'\times C_b(\overline{Q}\setminus\overline{\Omega}_0)}=0$$

for all $z\in C^1(\overline{Q})$ such that $z(0)=0$, where $(\bar y,\bar w)$ is an optimal solution for (P), $\bar\alpha$ is a multiplier of the cost functional, and $\bar\zeta\in(C_b(\overline{Q}\setminus\overline{\Omega}_0))'$ is a multiplier associated with the state constraint. As mentioned above, because of the term $g'(\bar y)^*\bar\zeta$, such an equation cannot be studied in the classical framework of partial differential equations. To overcome this difficulty, we identify $C_b(\overline{Q}\setminus\overline{\Omega}_0)$ with the space $C((\overline{Q}\setminus\overline{\Omega}_0)^\#)$, where $(\overline{Q}\setminus\overline{\Omega}_0)^\#$ is the Stone–Čech compactification of $\overline{Q}\setminus\overline{\Omega}_0$. Next, following an idea of DiPerna and Majda [7], we can associate with any $\zeta\in(C_b(\overline{Q}\setminus\overline{\Omega}_0))'\equiv\mathcal{M}((\overline{Q}\setminus\overline{\Omega}_0)^\#)$ an element of $\mathcal{M}(\overline{Q}\times(\overline{Q}\setminus\overline{\Omega}_0)^\#)$. In this way, we prove a decomposition theorem for elements of $(C_b(\overline{Q}\setminus\overline{\Omega}_0))'$ (Corollary 4.8). Each $\zeta\in(C_b(\overline{Q}\setminus\overline{\Omega}_0))'$ is represented by a "regular part," which is a bounded Radon measure on $\overline{Q}\setminus\overline{\Omega}_0$ and an additional part, which acts only on $\overline{\Omega}_0$. Due to this decomposition, we prove that only the regular part intervenes in the adjoint equation. The additional part intervenes in the optimality conditions. We state optimality conditions for (P) in Theorem 2.1. We give applications of these optimality conditions to the case of bilateral constraints in section 6. Notice that optimality conditions obtained in Corollary 6.2 are particularly simple.

The rest of the paper is organized as follows. Assumptions and the main result (Theorem 2.1) are stated in section 2. We have collected some results for the state and the adjoint equations in section 3. Results related to the Stone–Čech compactification are proved in section 4. The proof of the optimality conditions is given in section 5.

2. Assumptions. Main results. Throughout the paper, Ω denotes a bounded open subset in $\mathbb{R}^N$ $(N \geq 2)$ of class $C^{2+\gamma}$ for some $0 < \gamma \leq 1$. The coefficients a_{ij} of the operator A belong to $C^{1+\gamma}(\overline{\Omega})$ and satisfy the conditions

$$a_{ij}(x) = a_{ji}(x) \qquad \text{for every} \quad i,j \in \{1,\dots,N\}, \qquad m_0|\xi|^2 \leq \sum_{i,j=1}^{N} a_{ij}(x)\xi_i\xi_j$$

for all $\xi \in \mathbb{R}^N$ and all $x \in \overline{\Omega}$, with $m_0 > 0$. In (1), $\frac{\partial y}{\partial n_A}$ denotes the conormal derivative of y with respect to A, that is,

$$\frac{\partial y}{\partial n_A}(s,t) = \sum_{i,j} a_{ij}(s) D_j y(s,t) n_i(s),$$

where $n = (n_1,\dots,n_N)$ is the unit normal to Γ outward Ω.

For every $1 \leq \ell \leq \infty$, the usual norms in the spaces $L^\ell(\Omega)$, $L^\ell(Q)$, $L^\ell(\Sigma)$ will be denoted by $||\cdot||_{\ell,\Omega}$, $||\cdot||_{\ell,Q}$, $||\cdot||_{\ell,\Sigma}$. The Hilbert space

$$W(0,T;H^1(\Omega),(H^1(\Omega))') = \left\{ y \in L^2(0,T;H^1(\Omega)) \ \mid \ \frac{dy}{dt} \in L^2(0,T;(H^1(\Omega))') \right\}$$

will be denoted by $W(0,T)$. We introduce the space

$$Y_\infty = \left\{ y \in W(0,T) \ \mid \ \frac{\partial y}{\partial t} + Ay \in L^\infty(Q), \ \ \frac{\partial y}{\partial n_A} \in L^\infty(\Sigma), \ \text{ and } \ y(\cdot,0) \in L^\infty(\Omega) \right\}.$$

Endowed with the norm

$$y \longrightarrow ||y||_{W(0,T)} + ||\frac{\partial y}{\partial t} + Ay||_{\infty,Q} + ||\frac{\partial y}{\partial n_A}||_{\infty,\Sigma} + ||y(\cdot,0)||_{\infty,\Omega},$$

Y_∞ is a Banach space. Throughout what follows, $\langle \cdot,\cdot \rangle_{*,\overline{Q}\setminus\overline{\Omega}_0}$ denotes the duality pairing between the spaces $(C_b(\overline{Q} \setminus \overline{\Omega}_0))'$ and $C_b(\overline{Q} \setminus \overline{\Omega}_0)$. If $\mu \in \mathcal{M}_b(\overline{Q} \setminus \overline{\Omega}_0)$ (the space of bounded Radon measures on $\overline{Q}\setminus\overline{\Omega}_0$) and $y \in C_b(\overline{Q}\setminus\overline{\Omega}_0)$, we set $\langle \mu, y \rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} = \int_{\overline{Q}\setminus\overline{\Omega}_0} y(x,t)d\mu(x,t)$.

2.1. Assumptions. A1. For every $y \in \mathbb{R}$, $\Phi(\cdot,y)$ and $\Psi(\cdot,y)$ are, respectively, measurable on Q and Σ. For almost all $(x,t) \in Q$, $\Phi(x,t,\cdot)$ is of class C^1, and for almost all $(s,t) \in \Sigma$, $\Psi(s,t,\cdot)$ is of class C^1. Moreover, the following estimates hold:

$$|\Phi(x,t,0)| + |\Phi'_y(x,t,y)| \leq C_0, \quad |\Psi(s,t,0)| + |\Psi'_y(s,t,y)| \leq C_0 \ \text{ for some } C_0 \geq 0.$$

A2. For every $(y,w) \in \mathbb{R}^2$, $L(\cdot,y,w)$ is measurable on Ω. For almost all $x \in \Omega$, $L(x,\cdot,\cdot)$ is of class C^1. The following estimate holds:

$$|L(x,y,w)| + |L'_y(x,y,w)| + |L'_w(x,y,w)| \leq L_1(x,t)\eta(|w|)\eta(|y|),$$

where $L_1 \in L^1(\Omega)$ and η is a nondecreasing function from $\mathbb{R}^+$ into $\mathbb{R}^+$.

A3. W_{ad} is a closed convex subset in $L^\infty(\Omega)$.

A4. $g : C_b(\overline{Q} \setminus \overline{\Omega}_0) \longrightarrow C_b(\overline{Q} \setminus \overline{\Omega}_0)$ is of class C^1.

2.2. Statement of the main result. Define the following Hamiltonian function:

$$H(x,y,w,p,\alpha) = \alpha L(x,y,w) - pw \qquad \text{for all } (x,y,w,p,\alpha) \in \Omega \times \mathbb{R}^3 \times \mathbb{R}^+.$$

As mentioned in the introduction, every $\zeta \in (C_b(\overline{Q}\setminus\overline{\Omega}_0))'$ is identified with a measure $\hat{\hat{\zeta}} \in \mathcal{M}(\overline{Q}\times(\overline{Q}\setminus\overline{\Omega}_0)^{\#})$ (see section 4). For notational simplicity, $\hat{\hat{\zeta}}$ will also be denoted by ζ. The canonical projection from $\mathcal{M}(\overline{Q}\times(\overline{Q}\setminus\overline{\Omega}_0)^{\#})$ onto $\mathcal{M}(\overline{Q})$ is denoted by π and defined by

$$\pi : \zeta \in \mathcal{M}(\overline{Q}\times(\overline{Q}\setminus\overline{\Omega}_0)^{\#}) \longrightarrow \pi_\zeta \in \mathcal{M}(\overline{Q}),$$

$$\Big\langle \pi_\zeta, \phi \Big\rangle_{\mathcal{M}(\overline{Q})\times C(\overline{Q})} = \Big\langle \zeta, \phi \Big\rangle_{\mathcal{M}(\overline{Q}\times(\overline{Q}\setminus\overline{\Omega}_0)^{\#})\times C(\overline{Q}\times(\overline{Q}\setminus\overline{\Omega}_0)^{\#})} \qquad \text{for all } \phi \in C(\overline{Q}).$$

Throughout the following, for any $\zeta \in (C_b(\overline{Q}\setminus\overline{\Omega}_0))'$, $|\zeta|$ stands for the total variation of ζ. If $\mu \in \mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)$, we denote by $g'(\bar{y})^*\mu$ the Radon measure on $\overline{Q}\setminus\overline{\Omega}_0$ defined by

$$z \longrightarrow \Big\langle \mu, g'(\bar{y})z \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} \qquad \text{for all } z \in C_0(\overline{Q}\setminus\overline{\Omega}_0).$$

Moreover, $[g'(\bar{y})^*\mu]_{|Q}$ denotes the restriction of $g'(\bar{y})^*\mu$ to Q, $[g'(\bar{y})^*\mu]_{|\Sigma}$ denotes the restriction of $g'(\bar{y})^*\mu$ to Σ, and $[g'(\bar{y})^*\mu]_{|\overline{\Omega}_T}$ denotes the restriction of $g'(\bar{y})^*\mu$ to $\overline{\Omega}\times\{T\}$.

Theorem 2.1. *If* A1–A4 *are fulfilled and if* $(\bar{y},\bar{w})$ *is a solution of* (P)*, then there exist* $\bar{\alpha} \geq 0$, $\bar{p} \in L^1(0,T;W^{1,1}(\Omega))$, $\bar{\zeta} \in (C_b(\overline{Q}\setminus\overline{\Omega}_0))'$, *and a bounded linear transformation* $\Lambda_{\bar{\zeta}} : C_b(\overline{Q}\setminus\overline{\Omega}_0) \longrightarrow L^\infty_{\pi_{|\bar{\zeta}|}}(\overline{Q})$, *such that the following conditions hold:*

Nontriviality condition:

$$(\bar{\zeta},\bar{\alpha}) \neq 0. \tag{3}$$

Complementary condition:

$$\Big\langle \bar{\zeta}, z - g(\bar{y}) \Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} = \Big\langle \pi_{\bar{\zeta}}, z - g(\bar{y}) \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} + \int_{\overline{\Omega}_0} \Lambda_{\bar{\zeta}}\Big(z - g(\bar{y})\Big)\, d\pi_{|\bar{\zeta}|} \leq 0 \tag{4}$$

for all $z \in \mathcal{C}$.

Adjoint equation:

$$\begin{cases} -\frac{\partial \bar{p}}{\partial t} + A\bar{p} + \Phi'_y(x,t,\bar{y})\bar{p} + [g'(\bar{y})^*\pi_{\bar{\zeta}}]_{|Q} = 0 & \text{in } Q, \\ \frac{\partial \bar{p}}{\partial n_A} + \Psi'_y(s,t,\bar{y})\bar{p} + [g'(\bar{y})^*\pi_{\bar{\zeta}}]_{|\Sigma} = 0 & \text{on } \Sigma, \\ \bar{p}(x,T) + \bar{\alpha}L'_y(x,\bar{y}(T),\bar{w}) + [g'(\bar{y})^*\pi_{\bar{\zeta}}]_{|\overline{\Omega}_T} = 0 & \text{on } \overline{\Omega}. \end{cases} \tag{5}$$

Optimality condition for $\bar{w}$:

$$\int_\Omega H'_w\Big(x,\bar{y}(T),\bar{w},\bar{p}(0),\bar{\alpha}\Big)(\bar{w}-w)\,dx + \int_{\overline{\Omega}_0} \Lambda_{\bar{\zeta}}\Big(g'(\bar{y})(z_{\bar{w}} - z_w)\Big) d\pi_{|\bar{\zeta}|} \leq 0 \tag{6}$$

for all $w \in W_{ad}$, *where* z_w *is the solution of:*

$$\frac{\partial z}{\partial t} + Az = 0 \quad \text{in } Q, \qquad \frac{\partial z}{\partial n_A} = 0 \quad \text{on } \Sigma, \qquad z(\cdot,0) = w \quad \text{in } \Omega. \tag{7}$$

Decomposition property of $\bar\zeta$*:*

$$\left\langle \bar\zeta, h\right\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} = \left\langle \pi_{\bar\zeta}, h\right\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} + \int_{\overline{\Omega}_0} \Lambda_{\bar\zeta}(h)\, d\pi_{|\bar\zeta|} \quad \text{for all } h \in C_b(\overline{Q}\setminus\overline{\Omega}_0). \tag{8}$$

Property of the operator $\Lambda_{\bar\zeta}$*:*

$$\int_{\overline{\Omega}_0} \Lambda_{\bar\zeta}(h)\, d\pi_{|\bar\zeta|} = \left\langle \pi_{\bar\zeta}, h\right\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)} \quad \text{for all } h \in C(\overline{Q}). \tag{9}$$

Remark 2.2. *In the case when* $W_{ad} \subset C(\overline{\Omega})$ ($\mathcal{C}$ *is a closed convex subset of* $C(\overline{Q})$ *and* $g : C(\overline{Q}) \longrightarrow C(\overline{Q})$), *by using the property* (9) *satisfied by the operator* $\Lambda_{\bar\zeta}$, *the complementary condition is written as*

$$\left\langle \pi_{\bar\zeta}, z - g(\bar y)\right\rangle_{\mathcal{M}(\overline{Q})\times C(\overline{Q})} \le 0 \quad \text{for all } z \in \mathcal{C},$$

the optimality condition for $\bar w$ *is written as*

$$\int_\Omega H'_w\Big(x, \bar y(T), \bar w, \bar p(0), \bar\alpha\Big)(\bar w - w)\, dx + \left\langle [g'(\bar y)^*\pi_{\bar\zeta}]_{|\overline{\Omega}_0}, \bar w - w\right\rangle_{\mathcal{M}(\overline{\Omega})\times C(\overline{\Omega})} \le 0$$

for all $w \in W_{ad}$, *and the condition* $(\bar\zeta, \bar\alpha) \ne 0$ *is equivalent to* $(\pi_{\bar\zeta}, \bar\alpha) \ne 0$ (*it is a direct consequence of* (9)). *In this case, we recover optimality conditions proved in* [14] *and* [16].

Remark 2.3. *As in* [16], *we can recover optimality conditions in qualified form* (*that is, with* $\bar\alpha = 1$ *in* (5) *and* (6)) *under a strong stability condition.*

3. State equation and adjoint equation.

3.1. Existence, uniqueness, and regularity of the state variable. The following result is proved in [16, Proposition 3.4].

Proposition 3.1. *Let* (a, b) *be in* $L^\infty(Q) \times L^\infty(\Sigma)$ *satisfying* $||a||_{\infty,Q} \le C_0$ *and* $||b||_{\infty,\Sigma} \le C_0$. *For every* $(f, h, y_0) \in L^q(Q) \times L^r(\Sigma) \times L^\infty(\Omega)$ *(with* $q > \frac{N}{2} + 1$ *and* $r > N + 1$*), the solution* y *of*

$$\frac{\partial y}{\partial t} + Ay + ay = f \quad \text{in } Q, \quad \frac{\partial y}{\partial n_A} + by = h \quad \text{on } \Sigma, \quad y(0) = y_0 \quad \text{in } \Omega$$

belongs to $W(0, T) \cap C_b(\overline{Q}\setminus\overline{\Omega}_0)$. *Moreover, for every* $\tau \in]0, T]$, y *satisfies*

$$||y||_{\infty,Q} \le C\Big(||f||_{q,Q} + ||h||_{r,\Sigma} + ||y_0||_{\infty,\Omega}\Big),$$

$$||y||_{C(\overline{Q}_{\tau T})} \le C(\tau)\Big(||f||_{q,Q} + ||h||_{r,\Sigma} + ||y_0||_{2,\Omega}\Big)$$

for all $q > \frac{N}{2} + 1$ *and all* $r > N + 1$, *where* $C \equiv C(T, \Omega, N, C_0, q, r)$ *and* $C(\tau) \equiv C(T, \Omega, N, C_0, q, r, \tau)$.

Theorem 3.2 (see [15, Theorem 3.1]). *Let* w *be in* $L^\infty(\Omega)$*; then* (1) *admits a unique weak solution* y_w *in* $W(0, T) \cap C_b(\overline{Q}\setminus\overline{\Omega}_0)$. *This solution satisfies*

$$||y_w||_{\infty,Q} \le C_1(1 + ||w||_{\infty,\Omega}),$$

where $C \equiv C(T, \Omega, N, C_0)$. *Moreover, the mapping* $w \longrightarrow y_w$ *is continuous from* $L^\infty(\Omega)$ *into* $C_b(\overline{Q}\setminus\overline{\Omega}_0)$.

Theorem 3.3 (see [16, Corollary 3.1]). *For every $M > 0$ and every $\tau > 0$, there exist $C \equiv C(T, \Omega, N, C_0, M, \tau)$ and $\bar{\theta} > 0$ such that, for every $w \in W_{ad}$ satisfying $||w||_{\infty,\Omega} \leq M$, the weak solution y_w of* (1)*, corresponding to w, is Hölder continuous on $\overline{\Omega} \times [\tau, T]$ and satisfies*

$$||y_w||_{C^{\bar{\theta},\bar{\theta}/2}(\overline{\Omega}\times[\tau,T])} \leq C.$$

3.2. Adjoint equation. Let (a, b) be in $L^\infty(Q)\times L^\infty(\Sigma)$ such that $||a||_{\infty,Q} \leq C_0$ and $||b||_{\infty,\Sigma} \leq C_0$. We consider the following terminal boundary value problem:

$$(10) -\frac{\partial p}{\partial t} + Ap + ap = \mu_Q \quad \text{in } Q, \quad \frac{\partial p}{\partial n_A} + bp = \mu_\Sigma \quad \text{on } \Sigma, \quad p(T) = \mu_{\overline{\Omega}_T} \quad \text{on } \overline{\Omega},$$

where $\mu = \mu_Q + \mu_\Sigma + \mu_{\overline{\Omega}_T}$ is a bounded Radon measure on $\overline{Q}\setminus\overline{\Omega}_0$, μ_Q is the restriction of μ to Q, μ_Σ is the restriction of μ to Σ, and $\mu_{\overline{\Omega}_T}$ the restriction of μ to $\overline{\Omega} \times \{T\}$.

Definition 3.4. *A function $p \in L^1(0, T; W^{1,1}(\Omega))$ is a weak solution of* (10) *if and only if*

$$\int_Q \Big(p\frac{\partial z}{\partial t} + \sum_{i,j=1}^N a_{ij} D_j p D_i z + ayp\Big)\, dxdt + \int_\Sigma bpz\, dsdt = \Big\langle \mu, z\Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0}$$

for all $z \in C^1(\overline{Q})$ satisfying $z(0) = 0$ on $\overline{\Omega}$.

We recall an existence theorem for parabolic equations with measures as data stated in [14], [5].

Theorem 3.5. *Let (a, b) be in $L^\infty(Q)\times L^\infty(\Sigma)$ satisfying $||a||_{\infty,Q} \leq C_0$, $||b||_{\infty,\Sigma} \leq C_0$, and let μ be in $\mathcal{M}_b(\overline{Q} \setminus \overline{\Omega}_0)$. Equation* (10) *admits a unique solution p in $L^1(0, T; W^{1,1}(\Omega))$. For every (δ, d) satisfying $\delta > 1$, $d > 1$, $\frac{N}{2d} + \frac{1}{\delta} < \frac{1}{2}$, p belongs to $L^{\delta'}(0, T; W^{1,d'}(\Omega))$, and*

$$||p||_{L^{\delta'}(0,T;W^{1,d'}(\Omega))} \leq C||\mu||_{\mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)},$$

where $C \equiv C(\Omega, T, \delta, d, C_0)$ is a positive constant independent of a and b. Moreover, there exists a function in $L^1(\Omega)$, denoted by $p(0)$, such that

$$\int_Q \left(\frac{\partial z}{\partial t} + Az + az\right) p\, dxdt + \int_\Sigma \left(\frac{\partial z}{\partial n_A} + bz\right) p\, dsdt = \Big\langle \mu, z\Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} - \int_\Omega z(0)p(0)\, dx$$

for every $z \in Y_\infty = \{y \in W(0, T) \mid \frac{\partial y}{\partial t} + Ay \in L^\infty(Q),\ \frac{\partial y}{\partial n_A} \in L^\infty(\Sigma),$ and $y(0) \in L^\infty(\Omega)\}$.

Proof. The first part of the theorem is stated in [14], the only new result being the Green formula for test functions z belonging to Y_∞. (In [14], test functions belong to $Y_\infty \cap C(\overline{Q})$.) Let $z \in Y_\infty$ and set $f_z = \frac{\partial z}{\partial t} + Az + az$, $h_z = \frac{\partial z}{\partial n_A} + bz$, $k_z = z(0)$. Let $(z_{n0})_n$ be a sequence of regular functions converging to $z(0)$ for the weak-star topology of $L^\infty(\Omega)$ and for the strong topology of $L^2(\Omega)$. Let z_n be the solution of

$$\frac{\partial z_n}{\partial t} + Az_n + az_n = f_z \quad \text{in } Q, \quad \frac{\partial z_n}{\partial n_A} + bz_n = h_z \quad \text{on } \Sigma, \quad z_n(0) = z_{n0} \quad \text{in} \quad \Omega.$$

It is clear that $z_n \in C(\overline{Q})$. Moreover, the function $\phi = z_n - z$ is the solution of

$$\frac{\partial \phi}{\partial t} + A\phi + a\phi = 0 \quad \text{in } Q, \quad \frac{\partial \phi}{\partial n_A} + b\phi = 0 \quad \text{on } \Sigma, \quad \phi(0) = z(0) - z_{n0} \quad \text{in } \Omega$$

and satisfies (see Proposition 3.1)

$$||z_n - z||_{C_b(\overline{Q}\setminus\overline{\Omega}_0)} \le C||z_{n0} - z(0)||_{\infty,\Omega} \le C',$$

$$||z_n - z||_{C(\overline{Q}_{\tau T})} \le C(\tau)||z_{n0} - z(0)||_{2,\Omega} \quad \text{for every } \tau \in]0,T].$$

Therefore, $(z_n)_n$ converges to z uniformly on $\overline{Q}_{\tau T}$, for every $\tau \in]0,T]$. On the other hand, by using the Green formula stated in [14], it follows that

$$\int_Q \Big(\frac{\partial z_n}{\partial t} + Az_n + az_n\Big)p\,dxdt + \int_\Sigma \Big(\frac{\partial z_n}{\partial n_A} + bz_n\Big)p\,dsdt = \Big\langle \mu, z_n \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} - \int_\Omega z_{n0}p(0)\,dx.$$

We claim that

$$\lim_{n\to\infty} \Big\langle \mu, z_n \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} - \int_\Omega z_{n0}\ p(0)\,dx = \Big\langle \mu, z \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} - \int_\Omega z(0)p(0)\,dx.$$

Indeed, for every $\epsilon > 0$, there exists $\tau_\epsilon > 0$ such that $|\mu|(\overline{\Omega}\times]0,\tau_\epsilon[) \le \frac{\epsilon}{C'}$. Thus,

$$\begin{aligned}
&\Big|\Big\langle \mu, z - z_n \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} + \int_\Omega (z(0) - z_{n0})p(0)\,dx\Big| \\
&\le ||\mu||_{\mathcal{M}(\overline{Q}_{\tau_\epsilon})}\ ||z - z_n||_{C(\overline{Q}_{\tau_\epsilon})} + |\mu|\Big(\overline{\Omega}\times]0,\tau_\epsilon[\Big)\ ||z - z_n||_{C_b(\overline{\Omega}\times]0,\tau_\epsilon[)} \\
&+\Big|\int_\Omega (z(0) - z_{n0})p(0)\,dx\Big| \le C||z - z_n||_{C(\overline{Q}_{\tau_\epsilon})} + \epsilon + \Big|\int_\Omega (z(0) - z_{n0})p(0)\,dx\Big|.
\end{aligned}$$

By passing to the limit first when n tends to infinity, and next when ϵ tends to zero, we prove the above claim and the proof is complete. □

4. The Stone–Čech compactification. Let us introduce the compactification in the sense of Stone–Čech. Let $\mathcal{O}$ be a locally compact subset of $\overline{Q}$, and let $C_b(\mathcal{O})$ be the space of bounded continuous functions on $\mathcal{O}$. Denote by $(C_b(\mathcal{O}))'$ the topological dual of $C_b(\mathcal{O})$ and by B_0 the unit ball of $(C_b(\mathcal{O}))'$. For every $(x,t) \in \mathcal{O}$, consider $\nu_{(x,t)}$ in $(C_b(\mathcal{O}))'$ defined by

$$\Big\langle \nu_{(x,t)}, h \Big\rangle_{(C_b(\mathcal{O}))'\times C_b(\mathcal{O})} = h(x,t), \qquad h \in C_b(\mathcal{O}). \tag{11}$$

It is clear that

$$\sup_{(x,t)\in\mathcal{O}} \Big|\nu_{(x,t)}(h)\Big| = \sup_{(x,t)\in\mathcal{O}} \Big|\langle \nu_{(x,t)}, h\rangle_{(C_b(\mathcal{O}))'\times C_b(\mathcal{O})}\Big| = ||h||_{C_b(\mathcal{O})}.$$

Proposition 4.1 (see [6, p. 137]). *The mapping $(x,t) \longrightarrow \nu_{(x,t)}$ is a homeomorphism from $\mathcal{O}$ into a subset of B_0. (B_0 is endowed with the weak-star topology of $(C_b(\mathcal{O}))'$.)*

Let $\mathcal{O}^\#$ be the closure of the set $\{\nu_{(x,t)} \mid (x,t) \in \mathcal{O}\}$, for the weak-star topology of $(C_b(\mathcal{O}))'$. It is a Hausdorff compact space which, in view of the previous proposition, admits a dense subset homeomorphic to $\mathcal{O}$. This is called the Stone–Čech compactification of $\mathcal{O}$.

Proposition 4.2. *Each $h \in C_b(\mathcal{O})$ can be uniquely extended to a function $\tau(h) \in C(\mathcal{O}^\#)$, such that*

$$||\tau(h)||_{C(\mathcal{O}^\#)} = ||h||_{C_b(\mathcal{O})}.$$

Proof. With each $h \in C_b(\mathcal{O})$, we associate $\tau(h) \in C(\mathcal{O}^{\#})$ defined by

$$\tau(h)(q^{\#}) = \langle q^{\#}, h\rangle_{(C_b(\mathcal{O}))' \times C_b(\mathcal{O})} \qquad \text{for every } q^{\#} \in \mathcal{O}^{\#}.$$

We have $\sup_{q^{\#} \in \mathcal{O}^{\#}} |\tau(h)(q^{\#})| = \sup_{q^{\#} \in \mathcal{O}^{\#}} |\langle q^{\#}, h\rangle_{(C_b(\mathcal{O}))' \times C_b(\mathcal{O})}|$, and because the set $\mathcal{O}$ is identified with a dense subset of $\mathcal{O}^{\#}$, it follows that

$$\begin{aligned} ||\tau(h)||_{C(\mathcal{O}^{\#})} &= \sup_{q^{\#} \in \mathcal{O}^{\#}} \left|\tau(h)(q^{\#})\right| = \sup_{q^{\#} \in \mathcal{O}^{\#}} \left|\langle q^{\#}, h\rangle_{(C_b(\mathcal{O}))' \times C_b(\mathcal{O})}\right| \\ &= \sup_{(x,t) \in \mathcal{O}} \left|\nu_{(x,t)}(h)\right| = ||h||_{C_b(\mathcal{O})}. \end{aligned}$$

REMARK 4.3. *For every h_1, $h_2 \in C_b(\mathcal{O})$ and every real continuous function f, we have* (*see* [9, *pp.* 274 – 275])

$$\tau(h_1 h_2) = \tau(h_1)\tau(h_2), \qquad \tau(f(h)) = f(\tau(h)).$$

Moreover, there exists a continuous mapping i from $\mathcal{O}^{\#}$ into $\overline{\mathcal{O}}$, such that

$$\tau(\phi) = \phi \circ i \qquad \textit{for all } \phi \in C(\overline{\mathcal{O}}).$$

(*See* [8, *Theorem* 26, *p.* 278].)

The set $C(\overline{\mathcal{O}}) \otimes C(\mathcal{O}^{\#})$ of linear combinations of functions of the form $\phi\psi$ with $\phi \in C(\overline{\mathcal{O}})$ and $\psi \in C(\mathcal{O}^{\#})$ is a subspace of $C(\overline{\mathcal{O}} \times \mathcal{O}^{\#})$. The following result gives interesting properties for elements of $\mathcal{M}(\overline{\mathcal{O}} \times \mathcal{O}^{\#})$ (the topological dual of $C(\overline{\mathcal{O}} \times \mathcal{O}^{\#})$).

LEMMA 4.4. *Let η be a Radon measure on $\overline{\mathcal{O}} \times \mathcal{O}^{\#}$, let $\pi_\eta \in \mathcal{M}(\overline{\mathcal{O}})$ be the projection of η on $\overline{\mathcal{O}}$, and let $\pi_{|\eta|} \in \mathcal{M}^+(\overline{\mathcal{O}})$ be the projection of $|\eta|$ on $\overline{\mathcal{O}}$. There exists a bounded linear operator $\Lambda_\eta : C(\mathcal{O}^{\#}) \longrightarrow L^\infty_{\pi_{|\eta|}}(\overline{\mathcal{O}})$, such that*

$$\left\langle \eta, \phi\psi \right\rangle_{\#} = \int_{\overline{\mathcal{O}}} \phi \Lambda_\eta(\psi)\, d\pi_{|\eta|} \quad \textit{for all } (\phi, \psi) \in C(\overline{\mathcal{O}}) \times C(\mathcal{O}^{\#}), \tag{12}$$

where $\langle \cdot, \cdot \rangle_{\#}$ is the duality pairing between $\mathcal{M}(\overline{\mathcal{O}} \times \mathcal{O}^{\#})$ and $C(\overline{\mathcal{O}} \times \mathcal{O}^{\#})$.

Proof. Let η be in $\mathcal{M}(\overline{\mathcal{O}} \times \mathcal{O}^{\#})$. With notation of Lemma 4.4 we have

$$\left\langle \pi_\eta, \phi \right\rangle_{\mathcal{M}(\overline{\mathcal{O}}) \times C(\overline{\mathcal{O}})} = \left\langle \eta, \phi \right\rangle_{\#} \quad \text{for all } \phi \in C(\overline{\mathcal{O}}).$$

Let ψ be in $C(\mathcal{O}^{\#})$, and consider the Radon measure η_ψ on $\overline{\mathcal{O}}$ defined by

$$\left\langle \eta_\psi, \phi \right\rangle_{\mathcal{M}(\overline{\mathcal{O}}) \times C(\overline{\mathcal{O}})} = \left\langle \eta, \phi\psi \right\rangle_{\#}.$$

It satisfies

$$\left|\left\langle \eta_\psi, \phi \right\rangle_{\mathcal{M}(\overline{\mathcal{O}}) \times C(\overline{\mathcal{O}})}\right| \le ||\psi||_{C(\mathcal{O}^{\#})} \int_{\overline{\mathcal{O}}} |\phi|\, d\pi_{|\eta|}.$$

Thus, the measure η_ψ is absolutely continuous with respect to $\pi_{|\eta|}$ and, due to the Radon–Nikodým theorem, admits a representation of the form:

$$\left\langle \eta_\psi, \phi \right\rangle_{\mathcal{M}(\overline{\mathcal{O}}) \times C(\overline{\mathcal{O}})} = \int_{\overline{\mathcal{O}}} f_\psi \phi \, d\pi_{|\eta|} \quad \text{for all } \phi \in C(\overline{\mathcal{O}}), \quad \text{where } f_\psi \in L^1_{\pi_{|\eta|}}(\overline{\mathcal{O}}).$$

The integrand f_ψ linearly depends on ψ and defines a linear transformation Λ_η such that

$$\Big\langle \eta_\psi, \phi \Big\rangle_{\mathcal{M}(\overline{\mathcal{O}})\times C(\overline{\mathcal{O}})} = \int_{\overline{\mathcal{O}}} \Lambda_\eta(\psi)\phi \, d\pi_{|\eta|} \quad \text{for all } \phi \in C(\overline{\mathcal{O}}).$$

Observe that Λ_η is bounded in $L^\infty_{\pi_{|\eta|}}(\overline{\mathcal{O}})$. Indeed, since

$$\Big| \int_{\overline{\mathcal{O}}} \Lambda_\eta(\psi)\phi \, d\pi_{|\eta|} \Big| \le ||\psi||_{C(\mathcal{O}^\#)} \int_{\overline{\mathcal{O}}} |\phi| \, d\pi_{|\eta|} \quad \text{for all } \phi \in C(\overline{\mathcal{O}}),$$

and since $(L^1_{\pi_{|\eta|}}(\overline{\mathcal{O}}))' = L^\infty_{\pi_{|\eta|}}(\overline{\mathcal{O}})$, it follows that

$$||\Lambda_\eta(\psi)||_{L^\infty_{\pi_{|\eta|}}(\overline{\mathcal{O}})} = \sup_{\{\phi\in C(\overline{\mathcal{O}}) \mid \, ||\phi||_{L^1_{\pi_{|\eta|}}(\overline{\mathcal{O}})}=1\}} \Big| \int_{\overline{\mathcal{O}}} \Lambda_\eta(\psi)\phi \, d\pi_{|\eta|} \Big| \le ||\psi||_{C(\mathcal{O}^\#)}$$

for all $\psi \in C(\mathcal{O}^\#)$. The proof is complete. □

REMARK 4.5. *A measure $\zeta \in (C_b(\mathcal{O}))'$ can be identified with $\hat{\zeta} \in \mathcal{M}(\mathcal{O}^\#)$ via the formula*

$$(13) \quad \Big\langle \hat{\zeta}, \chi \Big\rangle_{\mathcal{M}(\mathcal{O}^\#)\times C(\mathcal{O}^\#)} = \Big\langle \zeta, \chi \circ \nu_{(\cdot)} \Big\rangle_{(C_b(\mathcal{O}))'\times C_b(\mathcal{O})} \qquad \textit{for all } \chi \in C(\mathcal{O}^\#)$$

(*where $\nu_{(\cdot)}$ is the evaluation measure defined by* (11)). *Additionally, elements of $(C_b(\mathcal{O}))'$ can be considered as elements of $\mathcal{M}(\overline{\mathcal{O}} \times \mathcal{O}^\#)$ (the topological dual of $C(\overline{\mathcal{O}} \times \mathcal{O}^\#)$). More precisely, with each $\hat{\zeta} \in \mathcal{M}(\mathcal{O}^\#)$ (defined by* (13)), *we associate $\hat{\hat{\zeta}} \in \mathcal{M}(\overline{\mathcal{O}} \times \mathcal{O}^\#)$ defined by*

$$(14) \quad \Big\langle \hat{\hat{\zeta}}, \psi \Big\rangle_{\#} = \Big\langle \hat{\zeta}, \psi \circ e \Big\rangle_{\mathcal{M}(\mathcal{O}^\#)\times C(\mathcal{O}^\#)} = \Big\langle \zeta, (\psi \circ e) \circ \nu_{(\cdot)} \Big\rangle_{(C_b(\mathcal{O}))'\times C_b(\mathcal{O})}$$

for all $\psi \in C(\overline{\mathcal{O}} \times \mathcal{O}^\#)$, where e is the continuous mapping from $\mathcal{O}^\#$ into $\overline{\mathcal{O}} \times \mathcal{O}^\#$, defined by

$$e(q^\#) = \Big(i(q^\#), q^\#\Big) \qquad q^\# \in \mathcal{O}^\#,$$

and where i is the continuous mapping defined in Remark 4.3.

LEMMA 4.6. *Let $\zeta \in (C_b(\mathcal{O}))'$, and let $\hat{\hat{\zeta}} \in \mathcal{M}(\overline{\mathcal{O}} \times \mathcal{O}^\#)$ be the associated measure defined by* (14). *Then*

$$(15) \quad \Big\langle \zeta, \phi h \Big\rangle_{(C_b(\mathcal{O}))'\times C_b(\mathcal{O})} = \Big\langle \hat{\hat{\zeta}}, \phi\tau(h) \Big\rangle_{\#} \qquad \textit{for all } (\phi, h) \in C(\overline{\mathcal{O}}) \times C_b(\mathcal{O}),$$

where τ is defined in Proposition 4.2. *Moreover, if ℓ is a continuous function with compact support in $\mathcal{O}$, then*

$$(16) \quad \Big\langle \hat{\hat{\zeta}}, \ell\phi\tau(h) \Big\rangle_{\#} = \Big\langle \pi_{\hat{\hat{\zeta}}}, \ell\phi h \Big\rangle_{\mathcal{M}(\overline{\mathcal{O}})\times C(\overline{\mathcal{O}})} \qquad \textit{for all } (\phi, h) \in C(\overline{\mathcal{O}}) \times C_b(\mathcal{O}).$$

Proof. Due to (13) and (14), for every $\phi \in C(\overline{\mathcal{O}})$ and every $h \in C_b(\mathcal{O})$, we have

$$\begin{aligned}&\Big\langle \zeta, \phi h \Big\rangle_{(C_b(\mathcal{O}))' \times C_b(\mathcal{O})} = \Big\langle \zeta, \tau(\phi h) \circ \nu_{(\cdot)} \Big\rangle_{(C_b(\mathcal{O}))' \times C_b(\mathcal{O})} = \Big\langle \hat{\zeta}, \tau(\phi h) \Big\rangle_{\mathcal{M}(\mathcal{O}^{\#}) \times C(\mathcal{O}^{\#})} \\ &= \Big\langle \hat{\zeta}, \tau(\phi)\tau(h) \Big\rangle_{\mathcal{M}(\mathcal{O}^{\#}) \times C(\mathcal{O}^{\#})} = \Big\langle \hat{\zeta}, (\phi \circ i)\ \tau(h) \Big\rangle_{\mathcal{M}(\mathcal{O}^{\#}) \times C(\mathcal{O}^{\#})} \\ &= \Big\langle \hat{\zeta}, (\phi\tau(h)) \circ e \Big\rangle_{\mathcal{M}(\mathcal{O}^{\#}) \times C(\mathcal{O}^{\#})} = \Big\langle \hat{\hat{\zeta}}, \phi\tau(h) \Big\rangle_{\#}.\end{aligned}$$

Therefore, we have proved (15). To prove (16), observe that, since $\ell\phi$ belongs to $C(\overline{\mathcal{O}})$, with (15) we have

$$\Big\langle \zeta, \ell\phi h \Big\rangle_{(C_b(\mathcal{O}))' \times C_b(\mathcal{O})} = \Big\langle \hat{\hat{\zeta}}, \ell\phi\tau(h) \Big\rangle_{\#}.$$

On the other hand, since $\ell\phi h$ belongs to $C(\overline{\mathcal{O}})$, from the definition of $\hat{\zeta}$ and $\hat{\hat{\zeta}}$, it follows that

$$\begin{aligned}&\Big\langle \zeta, \ell\phi h \Big\rangle_{(C_b(\mathcal{O}))' \times C_b(\mathcal{O})} = \Big\langle \zeta, \tau(\ell\phi h) \circ \nu_{(\cdot)} \Big\rangle_{(C_b(\mathcal{O}))' \times C_b(\mathcal{O})} \\ &= \Big\langle \hat{\zeta}, \tau(\ell\phi h) \Big\rangle_{\mathcal{M}(\mathcal{O}^{\#}) \times C(\mathcal{O}^{\#})} = \Big\langle \hat{\zeta}, (\ell\phi h) \circ i \Big\rangle_{\mathcal{M}(\mathcal{O}^{\#}) \times C(\mathcal{O}^{\#})} \\ &= \Big\langle \hat{\zeta}, (\ell\phi h) \circ e \Big\rangle_{\mathcal{M}(\mathcal{O}^{\#}) \times C(\mathcal{O}^{\#})} = \Big\langle \hat{\hat{\zeta}}, \ell\phi h \Big\rangle_{\#} = \Big\langle \pi_{\hat{\hat{\zeta}}}, \ell\phi h \Big\rangle_{\mathcal{M}(\overline{\mathcal{O}}) \times C(\overline{\mathcal{O}})}.\end{aligned}$$

The proof is complete. □

A very useful result is given in the following theorem.

THEOREM 4.7. *Let $\zeta \in (C_b(\mathcal{O}))'$, and let $\hat{\hat{\zeta}} \in \mathcal{M}(\overline{\mathcal{O}} \times \mathcal{O}^{\#})$ be the associated measure defined by* (14)*. There exists a bounded linear transformation $\Lambda_{\hat{\hat{\zeta}}} : C(\mathcal{O}^{\#}) \longrightarrow L^{\infty}_{\pi_{|\hat{\hat{\zeta}}|}}(\overline{Q})$, such that*

$$\Big\langle \zeta, h\phi \Big\rangle_{(C_b(\mathcal{O}))' \times C_b(\mathcal{O})} = \Big\langle \pi_{\hat{\hat{\zeta}}}, h\phi \Big\rangle_{\mathcal{M}_b(\mathcal{O}) \times C_b(\mathcal{O})} + \int_{\overline{\mathcal{O}} \setminus \mathcal{O}} \Lambda_{\hat{\hat{\zeta}}}\Big(\tau(h)\Big)\phi\, d\pi_{|\hat{\hat{\zeta}}|}$$

for all $(\phi, h) \in C(\overline{\mathcal{O}}) \times C_b(\mathcal{O})$. (In $\langle \pi_{\hat{\hat{\zeta}}}, \phi h\rangle_{\mathcal{M}_b(\mathcal{O}) \times C_b(\mathcal{O})}$, $\pi_{\hat{\hat{\zeta}}}$ denotes the restriction of $\pi_{\hat{\hat{\zeta}}}$ to $\mathcal{O}$.)

Proof. Due to (15) and Lemma 4.4, by setting $\psi = \tau(h)$ in (12), we obtain

$$\Big\langle \zeta, h\phi \Big\rangle_{(C_b(\mathcal{O}))' \times C_b(\mathcal{O})} = \Big\langle \hat{\hat{\zeta}}, \phi\tau(h) \Big\rangle_{\#} = \int_{\overline{\mathcal{O}}} \phi\Lambda_{\hat{\hat{\zeta}}}\Big(\tau(h)\Big) d\pi_{|\hat{\hat{\zeta}}|} \tag{17}$$

for all $(\phi, h) \in C(\overline{\mathcal{O}}) \times C_b(\mathcal{O})$. For $\phi \in C(\overline{\mathcal{O}})$, the integrals $\int_{\mathcal{O}} \phi\Lambda_{\hat{\hat{\zeta}}}(\tau(h))\, d\pi_{|\hat{\hat{\zeta}}|}$ and $\int_{\mathcal{O}} \phi h\, d\pi_{\hat{\hat{\zeta}}}$ are obtained by passing to the limit in $\int_{\mathcal{O}} \phi\ell_k\Lambda_{\hat{\hat{\zeta}}}(\tau(h))\, d\pi_{|\hat{\hat{\zeta}}|}$ and in $\int_{\mathcal{O}} \phi\ell_k h\ \, d\pi_{\hat{\hat{\zeta}}}$, where $(\ell_k)_k$ is a sequence of continuous functions with compact support in $\mathcal{O}$. From (16) and (17), we have

$$\Big\langle \hat{\hat{\zeta}}, \ell_k\phi\tau(h) \Big\rangle_{\#} = \Big\langle \pi_{\hat{\hat{\zeta}}}, \ell_k\phi h \Big\rangle_{\mathcal{M}(\overline{\mathcal{O}}) \times C(\overline{\mathcal{O}})} = \int_{\overline{\mathcal{O}}} \phi\ell_k\Lambda_{\hat{\hat{\zeta}}}\Big(\tau(h)\Big)\ \, d\pi_{|\hat{\hat{\zeta}}|}.$$

Therefore, we deduce that

$$\int_{\mathcal{O}} \phi \ell_k \Lambda_{\hat{\hat{\zeta}}}\Big(\tau(h)\Big)\ d\pi_{|\hat{\hat{\zeta}}|} = \int_{\mathcal{O}} \phi \ell_k h\ d\pi_{\hat{\hat{\zeta}}}.$$

By passing to the limit, when k tends to infinity, we obtain

$$\int_{\mathcal{O}} \phi \Lambda_{\hat{\hat{\zeta}}}\Big(\tau(h)\Big)\ d\pi_{|\hat{\hat{\zeta}}|} = \int_{\mathcal{O}} \phi h\ d\pi_{\hat{\hat{\zeta}}} = \Big\langle \pi_{\hat{\hat{\zeta}}}, \phi h \Big\rangle_{\mathcal{M}_b(\mathcal{O})\times C_b(\mathcal{O})}.$$

Consequently,

$$\int_{\overline{\mathcal{O}}} \phi \Lambda_{\hat{\hat{\zeta}}}\Big(\tau(h)\Big)\, d\pi_{|\hat{\hat{\zeta}}|} = \int_{\mathcal{O}} \phi \Lambda_{\hat{\hat{\zeta}}}\Big(\tau(h)\Big)\, d\pi_{|\hat{\hat{\zeta}}|} + \int_{\overline{\mathcal{O}}\setminus\mathcal{O}} \phi \Lambda_{\hat{\hat{\zeta}}}\Big(\tau(h)\Big)\, d\pi_{|\hat{\hat{\zeta}}|}$$

$$= \int_{\mathcal{O}} \phi h\, d\pi_{\hat{\hat{\zeta}}} + \int_{\overline{\mathcal{O}}\setminus\mathcal{O}} \phi \Lambda_{\hat{\hat{\zeta}}}\Big(\tau(h)\Big)\, d\pi_{|\hat{\hat{\zeta}}|}$$

for every $\phi \in C(\overline{\mathcal{O}})$ and every $h \in C_b(\mathcal{O})$. The proof is complete. □

COROLLARY 4.8. *Let $\zeta \in (C_b(\overline{Q} \setminus \overline{\Omega}_0))'$. There exists a bounded linear transformation $\Lambda_\zeta :\ C_b(\overline{Q} \setminus \overline{\Omega}_0) \longrightarrow L^\infty_{\pi_{|\zeta|}}(\overline{Q})$, such that*

$$\Big\langle \zeta, h\phi \Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} = \Big\langle \pi_\zeta, h\phi \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} + \int_{\overline{\Omega}_0} \Lambda_\zeta(h)\phi\, d\pi_{|\zeta|}$$

for all $h \in C_b(\overline{Q} \setminus \overline{\Omega}_0)$ and all $\phi \in C(\overline{Q})$. Moreover, for every $\tilde{h}$ in $C(\overline{Q})$, we have

$$\int_{\overline{\Omega}_0} \Lambda_\zeta(\tilde{h})\, d\pi_{|\zeta|} = \Big\langle \pi_\zeta, \tilde{h} \Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)}.$$

Proof. The first assertion is a direct consequence of Theorem 4.7, by setting $\mathcal{O} = \overline{Q}\setminus\overline{\Omega}_0$, $\Lambda_\zeta = \Lambda_{\hat{\hat{\zeta}}} \circ \tau$, and by identifying ζ with the measure $\hat{\hat{\zeta}}$, defined on $\overline{\mathcal{O}} \times \mathcal{O}^\#$, by (14). If $\tilde{h}$ belongs to $C(\overline{Q})$, by definition of π_ζ and with Theorem 4.7, we have

$$\Big\langle \zeta, \tilde{h} \Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} = \Big\langle \pi_\zeta, \tilde{h} \Big\rangle_{\mathcal{M}(\overline{Q})\times C(\overline{Q})} = \Big\langle \pi_\zeta, \tilde{h} \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} + \int_{\overline{\Omega}_0} \Lambda_\zeta(\tilde{h})\, d\pi_{|\zeta|}.$$

It follows that $\int_{\overline{\Omega}_0} \Lambda_\zeta(\tilde{h})\, d\pi_{|\zeta|} = \langle \pi_\zeta, \tilde{h} \rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)}$. The proof is complete. □

COROLLARY 4.9. *Let ζ be a nonnegative measure belonging to $(C_b(\overline{Q} \setminus \overline{\Omega}_0))'$. There exists a bounded linear transformation $\Lambda_\zeta :\ C_b(\overline{Q} \setminus \overline{\Omega}_0) \longrightarrow L^\infty_{\pi_\zeta}(\overline{Q})$, such that*

$$\Big\langle \zeta, h\phi \Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} = \Big\langle \pi_\zeta, h\phi \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} + \int_{\overline{\Omega}_0} \Lambda_\zeta(h)\phi\, d\pi_\zeta$$

for all $(h, \phi) \in C_b(\overline{Q} \setminus \overline{\Omega}_0) \times C(\overline{Q})$. Moreover, if $\tilde{h}$ is a nonnegative function in $C_b(\overline{Q} \setminus \overline{\Omega}_0)$, then $\int_{\overline{\Omega}_0} \Lambda_\zeta(\tilde{h})\, d\pi_\zeta \geq 0$.

Proof. The equality $\pi_\zeta = \pi_{|\zeta|}$ is obvious. Let $\tilde{h}$ be a nonnegative function in $C_b(\overline{Q} \setminus \overline{\Omega}_0)$. Observe that the Radon measure $\zeta_{\tilde{h}}$ defined by

$$\Big\langle \zeta_{\tilde{h}}, \phi \Big\rangle_{\mathcal{M}(\overline{Q})\times C(\overline{Q})} = \Big\langle \zeta, \phi\tilde{h} \Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} \qquad \phi \in C(\overline{Q}),$$

is nonnegative and that

$$\zeta_{\tilde{h}}(E) = \int_E \Lambda_\zeta(\tilde{h})\, d\pi_\zeta \qquad \text{for all Borel sets } E \subset \overline{Q}.$$

Therefore, $\int_{\overline{\Omega}_0} \Lambda_\zeta(\tilde{h})\, d\pi_\zeta = \zeta_{\tilde{h}}(\overline{\Omega}_0) \geq 0$. The proof is complete. □

For $\zeta \in (C_b(\overline{Q} \setminus \overline{\Omega}_0))'$ we denote by $(\pi_\zeta, \pi_{|\zeta|}, \Lambda_\zeta)$ the triplet corresponding to ζ defined in Corollary 4.8. The following result gives another property of the operator Λ_ζ often used later.

Proposition 4.10. *Let (a, b) be in $L^\infty(Q) \times L^\infty(\Sigma)$, let w be in $L^\infty(\Omega)$, and let z_1 and z_2 be solutions of*

$$\frac{\partial z_1}{\partial t} + Az_1 + az_1 = 0 \quad \text{in } Q, \qquad \frac{\partial z_1}{\partial n_A} + bz_1 = 0 \quad \text{on } \Sigma, \qquad z_1(\cdot, 0) = w \quad \text{in } \Omega,$$

$$\frac{\partial z_2}{\partial t} + Az_2 = 0 \quad \text{in } Q, \qquad \frac{\partial z_2}{\partial n_A} = 0 \quad \text{on } \Sigma, \qquad z_2(\cdot, 0) = w \quad \text{in } \Omega.$$

Then, for every $h \in C_b(\overline{Q} \setminus \overline{\Omega}_0)$, we have $\int_{\overline{\Omega}_0} \Lambda_\zeta(hz_1)\, d\pi_{|\zeta|} = \int_{\overline{\Omega}_0} \Lambda_\zeta(hz_2)\, d\pi_{|\zeta|}$.

Proof. Notice that the function $z_1 - z_2$ is the solution of

$$\frac{\partial z}{\partial t} + Az + az = -az_2 \quad \text{in } Q, \qquad \frac{\partial z}{\partial n_A} = -bz_2 \quad \text{on } \Sigma, \qquad z(\cdot, 0) = 0 \quad \text{in } \Omega,$$

and it belongs to $C_0(\overline{Q} \setminus \overline{\Omega}_0)$. It follows that, for every $h \in C_b(\overline{Q} \setminus \overline{\Omega}_0)$, the function $(z_1 - z_2)h$ belongs to $C_0(\overline{Q} \setminus \overline{\Omega}_0)$. From Corollary 4.8, we obtain

$$\begin{aligned} \Big\langle \zeta, (z_1 - z_2)h \Big\rangle_{*, \overline{Q} \setminus \overline{\Omega}_0} &= \Big\langle \pi_\zeta, (z_1 - z_2)h \Big\rangle_{b, \overline{Q} \setminus \overline{\Omega}_0} + \int_{\overline{\Omega}_0} \Lambda_\zeta((z_1 - z_2)h)\, d\pi_{|\zeta|} \\ &= \Big\langle \pi_\zeta, (z_1 - z_2)h \Big\rangle_{b, \overline{Q} \setminus \overline{\Omega}_0}. \end{aligned}$$

Thus, $\int_{\overline{\Omega}_0} \Lambda_\zeta(z_1 h)\, d\pi_{|\zeta|} = \int_{\overline{\Omega}_0} \Lambda_\zeta(z_2 h)\, d\pi_{|\zeta|}$. □

5. Optimality conditions.

5.1. Metric space of controls. Taylor expansions.

We endow W_{ad} with the distance $d(w_1, w_2) = ||w_1 - w_2||_{\infty, \Omega}$.

Lemma 5.1. *The metric space (W_{ad}, d) is complete. Moreover, the mapping which associates $(y_w, J(y_w, w), d_{\mathcal{C}}(g(y_w)))$ with the control w is continuous from (W_{ad}, d) into $C_b(\overline{Q} \setminus \overline{\Omega}_0) \times \mathbb{R} \times \mathbb{R}$.*

Proof. It is a direct consequence of Theorem 3.2. □

In the following, we consider control problems in which the state constraints (2) are penalized. For a given solution $\bar{w}$ of (P), the penalization is chosen so that $\bar{w}$ is an ϵ^2-solution of the penalized problem. To obtain optimality conditions for the penalized problem, we first establish Taylor expansions in the theorem below.

Theorem 5.2 (see [16, Theorem 4.1]). *Let ρ be such that $0 < \rho < 1$. For every w_1, $w_2 \in L^\infty(\Omega)$, if $w_\rho = w_1 + \rho w_2$, if y_ρ and y_1 are the solutions of* (1) *corresponding respectively to w_ρ and w_1, then we have*

$$\begin{aligned} y_\rho &= y_1 + \rho z + r_\rho \qquad \textit{with } \lim_{\rho \to 0} \frac{1}{\rho} ||r_\rho||_{C(\overline{Q})} = 0, \\ J(y_\rho, w_\rho) &= J(y_1, w_1) + \rho \Delta J + o(\rho), \end{aligned}$$

where z is the weak solution of

$$\frac{\partial z}{\partial t} + Az + \Phi'_y(\cdot, y_1)z = 0 \quad \textit{in } Q, \quad \frac{\partial z}{\partial n_A} + \Psi'_y(\cdot, y_1)z = 0 \ \ \textit{on } \Sigma, \quad z(0) = w_2 \ \ \textit{in } \Omega,$$

and

$$\Delta J = \int_\Omega \Big(L'_w(x, y_1(T), w_1)w_2 + L'_y(x, y_1(T), w_1)z(T) \Big)\, dx.$$

5.2. Approximate optimality conditions. Let $(\bar y, \bar w)$ be a solution of problem (P). For every $\epsilon > 0$, define

$$J_\epsilon(y, w) = \{[(J(y, w) - J(\bar y, \bar w) + \epsilon^2)^+]^2 + (d_{\mathcal{C}}(g(y)))^2\}^{\frac{1}{2}},$$

where $d_{\mathcal{C}}$ denotes the distance to $\mathcal{C}$ in $C_b(\overline{Q} \setminus \overline{\Omega}_0)$.

Due to Lemma 5.1, the functional $w \longrightarrow J_\epsilon(y_w, w)$ is continuous and bounded on the metric space (W_{ad}, d). Moreover, we have

$$J_\epsilon(y_w, w) > 0 \quad \text{for every } w \in W_{ad} \ \text{ and } \ J_\epsilon(\bar y, \bar w) = \epsilon^2 \leq \inf_{W_{ad}} J_\epsilon(y_w, w) + \epsilon^2.$$

Due to the Ekeland variational principle, there exists $w_\epsilon \in W_{ad}$ such that

$$\text{(18)} \qquad d(\bar w, w_\epsilon) \leq \epsilon \quad \text{and} \quad J_\epsilon(y_\epsilon, w_\epsilon) \leq J_\epsilon(y_w, w) + \epsilon d(w_\epsilon, w) \quad \text{for every } w \in W_{ad},$$

where y_ϵ is the solution of (1) corresponding to w_ϵ. The following theorem gives an approximate optimality condition for (y_ϵ, w_ϵ). Optimality conditions for $(\bar y, \bar w)$ will be recovered by passing to the limit, when ϵ tends to zero, in this approximate optimality condition.

THEOREM 5.3. *Let $w_\epsilon \in W_{ad}$ satisfy* (18)*. There exist $\alpha_\epsilon \geq 0$, $p_\epsilon \in L^1(0, T; W^{1,1}(\Omega))$, $\zeta_\epsilon \in (C_b(\overline{Q} \setminus \overline{\Omega}_0))'$, and a bounded linear transformation $\Lambda_\epsilon \equiv \Lambda_{\zeta_\epsilon} : C_b(\overline{Q} \setminus \overline{\Omega}_0) \longrightarrow L^\infty_{\pi_{|\zeta_\epsilon|}}(\overline{Q})$ such that*

$$\text{(19)} \qquad ||\zeta_\epsilon||^2_{(C_b(\overline{Q}\setminus\overline{\Omega}_0))'} + (\alpha_\epsilon)^2 = 1,$$

$$\text{(20)} \quad \Big\langle \zeta_\epsilon, z - g(y_\epsilon) \Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} = \Big\langle \pi_{\zeta_\epsilon}, z - g(y_\epsilon) \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} + \int_{\overline{\Omega}_0} \Lambda_\epsilon \Big(z - g(y_\epsilon) \Big)\, d\pi_{|\zeta_\epsilon|} \leq 0$$

for all $z \in \mathcal{C}$,

$$\text{(21)} \qquad \begin{cases} -\frac{\partial p_\epsilon}{\partial t} + Ap_\epsilon + \Phi'_y(x, t, y_\epsilon)p_\epsilon + [g'(y_\epsilon)^* \pi_{\zeta_\epsilon}]_{|Q} = 0 & \textit{in } Q, \\ \frac{\partial p_\epsilon}{\partial n_A} + \Psi'_y(s, t, y_\epsilon)\bar p_\epsilon + [g'(y_\epsilon)^* \pi_{\zeta_\epsilon}]_{|\Sigma} = 0 & \textit{on } \Sigma, \\ p_\epsilon(T) = -\alpha_\epsilon L'_y(x, y_\epsilon(T), w_\epsilon) - [g'(y_\epsilon)^* \pi_{\zeta_\epsilon}]_{|\overline{\Omega}_T} & \textit{on } \overline{\Omega}, \end{cases}$$

$$\text{(22)} \ \int_\Omega H'_w\Big(x, y_\epsilon(T), w_\epsilon, p_\epsilon(0), \alpha_\epsilon\Big)(w_\epsilon - w)\, dx + \int_{\overline{\Omega}_0} \Lambda_\epsilon\Big(g'(y_\epsilon)(z_{\epsilon,w_\epsilon} - z_{\epsilon,w})\Big)\, d\pi_{|\zeta_\epsilon|}$$

$$\leq C\epsilon\ (1 + ||w - w_\epsilon||_{\infty,\Omega}) \qquad \textit{for all } w \in W_{ad},$$

where $z_{\epsilon,\tilde w}$ *(with* $\tilde w = w_\epsilon$ *or* $\tilde w = w$*) is the weak solution of*

$$\frac{\partial z}{\partial t}+Az+\Phi'_y(\cdot,y_\epsilon)z=0 \ \textit{in } Q,\ \ \frac{\partial z}{\partial n_A}+\Psi'_y(\cdot,y_\epsilon)z=0 \ \textit{on } \Sigma,\quad z(0)=\tilde w \ \textit{in } \Omega. \tag{23}$$

Proof. Let (y_ϵ, w_ϵ) satisfy (18). Let us set

$$\begin{aligned} A_\epsilon = \Big\{&(z,\lambda)\in C_b(\overline{Q}\setminus\overline{\Omega}_0)\times\mathbb{R} \mid \exists w\in W_{ad},\ z=g(y_\epsilon)+g'(y_\epsilon)(z_{\epsilon,w}-z_{\epsilon,w_\epsilon}) \\ &\text{and } \lambda\ge J'_y(y_\epsilon,w_\epsilon)\ (z_{\epsilon,w}-z_{\epsilon,w_\epsilon})+J'_w(y_\epsilon,w_\epsilon)(w-w_\epsilon)+\epsilon\ d(w_\epsilon,w)\Big\}, \\ B = &\ \text{int}\ \mathcal{C}\times]-\infty,0[, \end{aligned}$$

where int $\mathcal{C}$ denotes the interior of $\mathcal{C}$. The sets A_ϵ and B are convex, and B is open. Let us prove that $A_\epsilon\cap B=\emptyset$. Suppose that there exists $w_o\in W_{ad}$, $\lambda_o\in\mathbb{R}$, such that

$$g(y_\epsilon)+g'(y_\epsilon)(z_{\epsilon,w_o}-z_{\epsilon,w_\epsilon})\in \text{int}\ \mathcal{C}, \tag{24}$$

$$0>\lambda_o\ge J'_y(y_\epsilon,w_\epsilon)\ (z_{\epsilon,w_o}-z_{\epsilon,w_\epsilon})+J'_w(y_\epsilon,w_\epsilon)(w_o-w_\epsilon)+\epsilon\ d(w_\epsilon,w_o). \tag{25}$$

Set $w^\rho_\epsilon = w_\epsilon+\rho(w_o-w_\epsilon)$, $z^\rho_\epsilon = g(y_\epsilon)+\frac{1}{\rho}(g(y^\rho_\epsilon)-g(y_\epsilon))$, and denote by y^ρ_ϵ the solution of (1) corresponding to w^ρ_ϵ. From (24), (25), and Theorem 5.2, it follows that

$$\lim_{\rho\searrow 0} z^\rho_\epsilon \in \text{int}\ \mathcal{C} \quad\text{and}\quad 0>\lim_{\rho\searrow 0}\Big(\frac{J(y^\rho_\epsilon,w^\rho_\epsilon)-J(y_\epsilon,w_\epsilon)}{\rho}+\epsilon\ \frac{d(w_\epsilon,w^\rho_\epsilon)}{\rho}\Big).$$

Therefore, there exists $\rho_o>0$ such that, for every $0<\rho\le\rho_o<1$, we have

$$g(y^\rho_\epsilon)=\rho\ z^\rho_\epsilon+(1-\rho)\ g(y_\epsilon)\in\ \text{int}\ \mathcal{C} \quad\text{and}\quad J(y^\rho_\epsilon,w^\rho_\epsilon)<J(y_\epsilon,w_\epsilon)-\epsilon\ d(w_\epsilon,w^\rho_\epsilon).$$

Observe that

$$\begin{aligned} J_\epsilon(y^\rho_\epsilon,w^\rho_\epsilon) &= (J(y^\rho_\epsilon,w^\rho_\epsilon)-J(\bar y,\bar w)+\epsilon^2)^+ = J(y^\rho_\epsilon,w^\rho_\epsilon)-J(\bar y,\bar w)+\epsilon^2 \\ &< J(y_\epsilon,w_\epsilon)-J(\bar y,\bar w)+\epsilon^2-\epsilon\ d(w_\epsilon,w^\rho_\epsilon)\le J_\epsilon(y_\epsilon,w_\epsilon)-\epsilon\ d(w_\epsilon,w^\rho_\epsilon). \end{aligned}$$

This contradicts (18) and proves that $A_\epsilon\cap B=\emptyset$. From a geometric version of the Hahn–Banach theorem, there exists $(\alpha_\epsilon,\zeta_\epsilon)\in\mathbb{R}\times(C_b(\overline{Q}\setminus\overline{\Omega}_0))'$, such that

$$\alpha_\epsilon\ \lambda_\epsilon+\Big\langle\zeta_\epsilon,z_\epsilon\Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} > \alpha_\epsilon\ \lambda+\Big\langle\zeta_\epsilon,z\Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} \text{ for all } (z_\epsilon,\lambda_\epsilon)\in A_\epsilon,\ \text{for all } (z,\lambda)\in B. \tag{26}$$

We can easily check that α_ϵ is nonnegative and that $(\alpha_\epsilon,\zeta_\epsilon)\neq 0$. By a normalization procedure, we can suppose that $(\alpha_\epsilon,\zeta_\epsilon)$ satisfies (19). Due to (26), we obtain

$$\alpha_\epsilon\ \lambda_\epsilon+\Big\langle\zeta_\epsilon,z_\epsilon\Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} \ge \alpha_\epsilon\ \lambda+\Big\langle\zeta_\epsilon,z\Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} \text{ for all } (z_\epsilon,\lambda_\epsilon)\in A_\epsilon,\ \text{for all } (z,\lambda)\in \overline{B}. \tag{27}$$

Therefore, by taking $z_\epsilon=g(y_\epsilon)$, $z\in\mathcal{C}$, and $\lambda_\epsilon=\lambda=0$, we deduce that

$$\Big\langle\zeta_\epsilon,z-g(y_\epsilon)\Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0}\le 0 \qquad \text{for all } z\in\mathcal{C}. \tag{28}$$

Due to Corollary 4.8, there exists a bounded linear transformation $\Lambda_\epsilon\equiv\Lambda_{\zeta_\epsilon}:\ C_b(\overline{Q}\setminus\overline{\Omega}_0)\longrightarrow L^\infty_{\pi_{|\zeta_\epsilon|}}(\overline{Q})$, such that

$$\Big\langle\zeta_\epsilon,h\Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0}=\Big\langle\pi_{\zeta_\epsilon},h\Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0}+\int_{\overline{\Omega}_0}\Lambda_\epsilon(h)\,d\pi_{|\zeta_\epsilon|} \qquad \text{for all } h\in C_b(\overline{Q}\setminus\overline{\Omega}_0).$$

Thus, (20) follows from (28). Let $w \in W_{ad}$; by setting $z_\epsilon = g(y_\epsilon) + g'(y_\epsilon)(z_{\epsilon,w} - z_{\epsilon,w_\epsilon})$, $\lambda_\epsilon = J'_y(y_\epsilon, w_\epsilon)\ (z_{\epsilon,w} - z_{\epsilon,w_\epsilon}) + J'_w(y_\epsilon, w_\epsilon)(w - w_\epsilon) + \epsilon\ d(w_\epsilon, w)$, $\lambda = 0$, and $z = g(\bar{y})$, in (27), with (18) and Proposition 3.1, we obtain

$$\begin{aligned}
(29) \qquad & \alpha_\epsilon\Big(J'_y(y_\epsilon, w_\epsilon)\ (z_{\epsilon,w} - z_{\epsilon,w_\epsilon}) + J'_w(y_\epsilon, w_\epsilon)(w - w_\epsilon)\Big) \\
& + \Big\langle \pi_{\zeta_\epsilon}, g'(y_\epsilon)(z_{\epsilon,w} - z_{\epsilon,w_\epsilon})\Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} + \int_{\overline{\Omega}_0} \Lambda_\epsilon\Big(g'(y_\epsilon)(z_{\epsilon,w} - z_{\epsilon,w_\epsilon})\Big)\, d\pi_{|\zeta_\epsilon|}, \\
& \geq -\alpha_\epsilon\ \epsilon\ d(w_\epsilon, w) + \Big\langle \zeta_\epsilon, g(\bar{y}) - g(y_\epsilon)\Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} \geq -C\epsilon\ (1 + ||w - w_\epsilon||_{\infty,\Omega}),
\end{aligned}$$

where C is independent of ϵ. If p_ϵ is the solution of (21), with Theorem 3.5, we obtain

$$\begin{aligned}
(30) \qquad & \int_\Omega -\alpha_\epsilon L'_y(x, y_\epsilon(T), w_\epsilon)(z_{\epsilon,w} - z_{\epsilon,w_\epsilon})(T)\, dx - \Big\langle \zeta_\epsilon, g'(y_\epsilon)(z_{\epsilon,w} - z_{\epsilon,w_\epsilon})\Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} \\
& = \int_\Omega p_\epsilon(0)(w - w_\epsilon)\, dx - \int_{\overline{\Omega}_0} \Lambda_\epsilon\Big(g'(y_\epsilon)(z_{\epsilon,w} - z_{\epsilon,w_\epsilon})\Big)\, d\pi_{|\zeta_\epsilon|}.
\end{aligned}$$

The optimality condition (22) follows from (29) and (30). □

5.3. Proof of optimality conditions. Theorem 3.5 gives

$$\begin{aligned}
||p_\epsilon||_{L^{\delta'}(0,T;W^{1,d'}(\Omega))} & \leq C\Big(||L'_y(\cdot, y_\epsilon(T), w_\epsilon)||_{1,\Omega} + ||g'(y_\epsilon)||_{\mathcal{L}(C_b(\overline{Q}\setminus\overline{\Omega}_0))}||\pi_{\zeta_\epsilon}||_{\mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)}\Big) \\
& \leq C\Big(||L'_y(\cdot, y_\epsilon(T), w_\epsilon)||_{1,\Omega} + ||g'(y_\epsilon)||_{\mathcal{L}(C_b(\overline{Q}\setminus\overline{\Omega}_0))}||\zeta_\epsilon||_{(C_b(\overline{Q}\setminus\overline{\Omega}_0))'}\Big)
\end{aligned}$$

for every $\delta > 1$, $d > 1$ satisfying $\frac{N}{2d} + \frac{1}{\delta} < \frac{1}{2}$ ($\mathcal{L}(C_b(\overline{Q} \setminus \overline{\Omega}_0))$ is the space of linear continuous mappings from $C_b(\overline{Q} \setminus \overline{\Omega}_0)$ into $C_b(\overline{Q} \setminus \overline{\Omega}_0)$). Since the sequences $(\zeta_\epsilon)_\epsilon$, $(y_\epsilon)_\epsilon$, and $(w_\epsilon)_\epsilon$ are bounded in $(C_b(\overline{Q} \setminus \overline{\Omega}_0))'$, $C_b(\overline{Q} \setminus \overline{\Omega}_0)$, and $L^\infty(\Omega)$ respectively, it follows that the sequence $(p_\epsilon)_\epsilon$ is bounded in $L^{\delta'}(0,T;W^{1,d'}(\Omega))$ for every (δ, d) satisfying $\frac{N}{2d} + \frac{1}{\delta} < \frac{1}{2}$. Then there exists a subsequence, still indexed by ϵ, and $\bar{p}$ such that $(p_\epsilon)_\epsilon$ converges to $\bar{p}$ weakly in $L^{\delta'}(0,T;W^{1,d'}(\Omega))$ for every (δ, d) satisfying $\frac{N}{2d} + \frac{1}{\delta} < \frac{1}{2}$. Moreover, due to classical imbeddings, $(p_\epsilon)_\epsilon$ converges to $\bar{p}$ weakly in $L^{q'}(Q)$ for all $q > \frac{N}{2} + 1$, and the sequence of traces $(p_{\epsilon|\Sigma})_\epsilon$ converges to $\bar{p}_{|\Sigma}$ weakly in $L^{r'}(\Sigma)$ for all $r > N + 1$. The sequence $(\alpha_\epsilon)_\epsilon$ (or at least a subsequence) converges to $\bar{\alpha} \geq 0$, $(w_\epsilon)_\epsilon$ converges to $\bar{w}$ in $L^\infty(\Omega)$, and $(y_\epsilon)_\epsilon$ converges to $\bar{y}$ in $C_b(\overline{Q} \setminus \overline{\Omega}_0)$.

From assumptions on Φ, Ψ and L, we can prove that the sequence $(\Phi'_y(\cdot, y_\epsilon), \Psi'_y(\cdot, y_\epsilon), L'_y(\cdot, y_\epsilon(T), w_\epsilon))_\epsilon$ converges to $(\Phi'_y(\cdot, \bar{y}), \Psi'_y(\cdot, \bar{y}), L'_y(\cdot, \bar{y}(T), \bar{w}))$ in $L^\theta(Q) \times L^\theta(\Sigma) \times L^1(\Omega)$ for every $\theta < \infty$. The sequence $(\zeta_\epsilon)_\epsilon$ is bounded in $(C_b(\overline{Q} \setminus \overline{\Omega}_0))'$. Then there exists a generalized sequence, still indexed by ϵ, such that $(\zeta_\epsilon)_\epsilon$ converges to a limit $\bar{\zeta}$ for the weak-star topology of $(C_b(\overline{Q} \setminus \overline{\Omega}_0))'$. Due to Corollary 4.8, there exists a bounded linear transformation $\Lambda_{\bar{\zeta}} :\ C_b(\overline{Q} \setminus \overline{\Omega}_0) \longrightarrow L^\infty_{\pi_{|\bar{\zeta}|}}(\overline{Q})$, such that

$$\Big\langle \bar{\zeta}, h\Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} = \Big\langle \pi_{\bar{\zeta}}, h\Big\rangle_{b,\overline{Q}\setminus\bar{\Omega}_0} + \int_{\overline{\Omega}_0} \Lambda_{\bar{\zeta}}(h)\, d\pi_{|\bar{\zeta}|} \quad \text{for all } h \in C_b(\overline{Q} \setminus \overline{\Omega}_0).$$

For every $z \in C^1(\overline{Q})$ satisfying $z(0) = 0$, we have

$$\int_Q \left(p_\epsilon \frac{\partial z}{\partial t} + \sum_{i,j} a_{ij} D_i z D_j p_\epsilon + \Phi'_y(x,t,y_\epsilon) p_\epsilon z) \right) , dxdt + \int_\Sigma p_\epsilon \Psi'_y(s,t,y_\epsilon) z \, dsdt$$
$$+ \int_\Omega \alpha_\epsilon L'_y(x, y_\epsilon(T), w_\epsilon) z(x,T)\, dx = -\Big\langle \pi_{\zeta_\epsilon}, g'(y_\epsilon) z \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} = -\Big\langle \zeta_\epsilon, g'(y_\epsilon) z \Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0}.$$

By passing to the limit in this variational formulation, it follows that:

$$\int_Q \left(\bar{p} \frac{\partial z}{\partial t} + \sum_{i,j} a_{ij} D_i z D_j \bar{p} + \Phi'_y(x,t,\bar{y}) \bar{p} z \right) dxdt + \int_\Sigma \bar{p} \Psi'_y(s,t,\bar{y}) z \, dsdt$$
$$+ \int_\Omega \bar{\alpha} L'_y(x, \bar{y}(T), \bar{w}) z(x,T)\, dx = -\Big\langle \bar{\zeta}, g'(\bar{y}) z \Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} = -\Big\langle \pi_{\bar{\zeta}}, g'(\bar{y}) z \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0}$$

for every $z \in C^1(\overline{Q})$ satisfying $z(0) = 0$. Therefore, $\bar{p}$ is the weak solution of equation (5). On the other hand, by using the Green formulas satisfied by $\bar{p}$ and p_ϵ, we have

$$(31) \qquad -\int_\Omega \alpha_\epsilon L'_y(x, y_\epsilon(T), w_\epsilon)(z_{\epsilon,w} - z_{\epsilon,w_\epsilon})(T)\, dx - \int_\Omega p_\epsilon(0)(w - w_\epsilon)\, dx$$
$$= \Big\langle \zeta_\epsilon, g'(y_\epsilon)(z_{\epsilon,w} - z_{\epsilon,w_\epsilon}) \Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} - \int_{\overline{\Omega}_0} \Lambda_\epsilon\Big(g'(y_\epsilon)(z_{\epsilon,w} - z_{\epsilon,w_\epsilon}) \Big)\, d\pi_{|\zeta_\epsilon|}$$

and

$$(32) \qquad -\int_\Omega \bar{\alpha} L'_y(x, \bar{y}(T), \bar{w})(\bar{z}_w - \bar{z}_{\bar{w}})(T)\, dx - \int_\Omega \bar{p}(0)(w - \bar{w})(0)\, dx$$
$$= \Big\langle \bar{\zeta}, g'(\bar{y})(\bar{z}_w - \bar{z}_{\bar{w}}) \Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} - \int_{\overline{\Omega}_0} \Lambda_{\bar{\zeta}}\Big(g'(\bar{y})(\bar{z}_w - \bar{z}_{\bar{w}}) \Big) d\pi_{|\bar{\zeta}|},$$

where $\bar{z}_{\tilde{w}}$ (for $\tilde{w} = \bar{w}$ or $\tilde{w} = w$) is the solution of

$$(33)\, \frac{\partial z}{\partial t} + Az + \Phi'_y(\cdot, \bar{y}) z = 0 \quad \text{in } Q, \quad \frac{\partial z}{\partial n_A} + \Psi'_y(\cdot, \bar{y}) z = 0 \;\; \text{on } \Sigma, \quad z(0) = \tilde{w} \;\; \text{in } \Omega.$$

Let us set

$$I_\epsilon = \int_\Omega \Big(p_\epsilon(0)(w - w_\epsilon) - \bar{p}(0)(w - \bar{w}) \Big)\, dx$$
$$- \int_{\overline{\Omega}_0} \Lambda_\epsilon\Big(g'(y_\epsilon)(z_w - z_{w_\epsilon}) \Big) d\pi_{|\zeta_\epsilon|} + \int_{\overline{\Omega}_0} \Lambda_{\bar{\zeta}}\Big(g'(\bar{y})(\bar{z}_w - \bar{z}_{\bar{w}}) \Big)\, d\pi_{|\bar{\zeta}|}.$$

From (31) and (32), it follows that

$$
\begin{aligned}
\Big|I_\epsilon\Big| \leq & \Big| \int_\Omega \bar\alpha L'_y(x,\bar y(T),\bar w)(\bar z_w - \bar z_{\bar w})(T)\,dx \\
& - \int_\Omega \alpha_\epsilon L'_y(x,y_\epsilon(T),w_\epsilon))(z_{\epsilon,w} - z_{\epsilon,w_\epsilon})(T)\,dx\Big| \\
& + \Big|\Big\langle \bar\zeta, g'(\bar y)(\bar z_w - \bar z_{\bar w})\Big\rangle_{*,\overline{Q}\backslash\overline{\Omega}_0} - \Big\langle \zeta_\epsilon, g'(y_\epsilon)(z_{\epsilon,w} - z_{\epsilon,w_\epsilon})\Big\rangle_{*,\overline{Q}\backslash\overline{\Omega}_0}\Big| \\
\leq & \Big|\Big|\bar\alpha L'_y(\cdot,\bar y(T),\bar w) - \alpha_\epsilon L'_y(\cdot,y_\epsilon(T),w_\epsilon)\Big|\Big|_{1,\Omega}\Big|\Big|z_{\epsilon,w} - z_{\epsilon,w_\epsilon}\Big|\Big|_{\infty,Q} \\
& + \Big|\Big|\bar\alpha L'_y(\cdot,\bar y(T),\bar w)\Big|\Big|_{1,\Omega}\Big|\Big|(z_{\epsilon,w} - z_{\epsilon,w_\epsilon}) - (\bar z_w - \bar z_{\bar w})\Big|\Big|_{\infty,Q} \\
& + \Big|\Big\langle \bar\zeta - \zeta_\epsilon, g'(\bar y)(\bar z_w - \bar z_{\bar w})\Big\rangle_{*,\overline{Q}\backslash\overline{\Omega}_0}\Big| \\
& + \Big|\Big|g'(\bar y)(z_w - z_{\bar w}) - g'(y_\epsilon)(z_{\epsilon,w} - z_{\epsilon,w_\epsilon})\Big|\Big|_{\infty,Q}.
\end{aligned}
$$

Since the generalized sequence $(\alpha_\epsilon, w_\epsilon, y_\epsilon, z_{\epsilon,w_\epsilon}, z_{\epsilon,w}, \zeta_\epsilon)_\epsilon$ converges to $(\bar\alpha, \bar w, \bar y, \bar z_{\bar w}, \bar z_w, \bar\zeta)$, we obtain $\lim_\epsilon I_\epsilon = 0$. Therefore, by passing to the limit in the approximate optimality conditions satisfied by w_ϵ, we obtain

$$
(34)\qquad \int_\Omega H'_w\Big(x,\bar y(T),\bar w,\bar p(0),\bar\alpha\Big)(\bar w - w)\,dx + \int_{\overline{\Omega}_0} \Lambda_{\bar\zeta}\Big(g'(\bar y)(\bar z_{\bar w} - \bar z_w)\Big)d\pi_{|\bar\zeta|} \leq 0
$$

for all $w \in W_{ad}$. Let $z_{\tilde w}$ (for $\tilde w = \bar w$ or $\tilde w = w$) be the solution of (7) corresponding to $\tilde w$. Due to Proposition 4.10, we have

$$
(35)\qquad \int_{\overline{\Omega}_0} \Lambda_{\bar\zeta}\Big(g'(\bar y)\bar z_{\tilde w}\Big)d\pi_{|\bar\zeta|} = \int_{\overline{\Omega}_0} \Lambda_{\bar\zeta}\Big(g'(\bar y)z_{\tilde w}\Big)d\pi_{|\bar\zeta|}.
$$

By taking (34) and (35) into account, we obtain (8). On the other hand, by passing to the limit in

$$
\Big\langle \zeta_\epsilon, z\Big\rangle_{*,\overline{Q}\backslash\overline{\Omega}_0} \leq \Big\langle \zeta_\epsilon, g(y_\epsilon)\Big\rangle_{*,\overline{Q}\backslash\overline{\Omega}_0} \qquad \text{for all } z \in \mathcal{C},
$$

we obtain condition (4). Finally, by passing to the limit in ${\alpha_\epsilon}^2 + (||\zeta_\epsilon||_{(C_b(\overline{Q}\backslash\overline{\Omega}_0))'})^2 = 1$, we obtain $\bar\alpha^2 + (\lim_\epsilon ||\zeta_\epsilon||_{(C_b(\overline{Q}\backslash\overline{\Omega}_0))'})^2 = 1$. If $\bar\alpha > 0$, the proof is complete. If $\bar\alpha = 0$, we must prove that $||\bar\zeta||_{(C_b(\overline{Q}\backslash\overline{\Omega}_0))'} > 0$. Since $\mathrm{int}_{C_b(\overline{Q}\backslash\overline{\Omega}_0)}\mathcal{C} \neq \emptyset$, there exists a ball $B(z,2\rho) \subset \mathcal{C}$ in $C_b(\overline{Q} \setminus \overline{\Omega}_0)$, with center z and radius $2\rho > 0$. We can choose $\tilde z_\epsilon \in B(0,2\rho)$ such that $\langle \zeta_\epsilon, \tilde z_\epsilon\rangle_{*,\overline{Q}\backslash\overline{\Omega}_0} = \rho||\zeta_\epsilon||_{(C_b(\overline{Q}\backslash\overline{\Omega}_0))'}$. Since $z + \tilde z_\epsilon \in \mathcal{C}$, we have

$$
\Big\langle \zeta_\epsilon, z + \tilde z_\epsilon - g(y_\epsilon)\Big\rangle_{*,\overline{Q}\backslash\overline{\Omega}_0} \leq 0.
$$

By passing to the limit, we obtain $\rho + \langle \bar\zeta, z - g(\bar y)\rangle_{*,\overline{Q}\backslash\overline{\Omega}_0} \leq 0$, and thus $\bar\zeta \neq 0$. □

6. Application of Theorem 2.1. Consider state constraints of the form

$$
(36)\qquad a(x,t) \leq y(x,t) \leq b(x,t) \quad \text{ for all } (x,t) \in \overline{Q} \setminus \overline{\Omega}_0,
$$

where a and b are two functions in $C(\overline{Q})$ satisfying $a(x,t) < b(x,t)$ on $\overline{Q}$. The state constraints (36) may be written in the form (2) by setting

$$y \in \mathcal{C} = \{z \in C_b(\overline{Q} \setminus \overline{\Omega}_0) \mid a \le z \le b\}. \tag{37}$$

Theorem 6.1. *Suppose that* A1–A4 *are fulfilled and that the state constraint* (2) *is defined by* (37). *Then there exist* $\bar\alpha \ge 0$, $\bar p \in L^1(0,T;W^{1,1}(\Omega))$, $\bar\zeta \in (C_b(\overline{Q} \setminus \overline{\Omega}_0))'$, *two bounded linear transformations* $\Lambda_+ : C_b(\overline{Q} \setminus \overline{\Omega}_0) \longrightarrow L^\infty_{\pi_{\bar\zeta^+}}(\overline{Q})$, *and* $\Lambda_- : C_b(\overline{Q} \setminus \overline{\Omega}_0) \longrightarrow L^\infty_{\pi_{\bar\zeta^-}}(\overline{Q})$ *such that*

$$(\pi_{\bar\zeta^-}, \pi_{\bar\zeta^+}, \bar\alpha) \neq 0, \tag{38}$$

$$\Big\langle \pi_{\bar\zeta^+}, \bar y \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} = \Big\langle \pi_{\bar\zeta^+}, b \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0}, \quad \int_{\overline{\Omega}_0} \Lambda_+(\bar y)\, d\pi_{\bar\zeta^+} = \Big\langle \pi_{\bar\zeta^+}, b \Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)}, \tag{39}$$

$$\Big\langle \pi_{\bar\zeta^-}, \bar y \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} = \Big\langle \pi_{\bar\zeta^-}, a \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0}, \quad \int_{\overline{\Omega}_0} \Lambda_-(\bar y)\, d\pi_{\bar\zeta^-} = \Big\langle \pi_{\bar\zeta^-}, a \Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)}, \tag{40}$$

$$\begin{cases} -\frac{\partial \bar p}{\partial t} + A\bar p + \Phi'_y(x,t,\bar y)\bar p + [\pi_{\bar\zeta^+} - \pi_{\bar\zeta^-}]_{|Q} = 0 & \text{in } Q, \\ \frac{\partial \bar p}{\partial n_A} + \Psi'_y(s,t,\bar y)\bar p + [\pi_{\bar\zeta^+} - \pi_{\bar\zeta^-}]_{|\Sigma} = 0 & \text{on } \Sigma, \\ \bar p(x,T) + \bar\alpha L'_y(x,\bar y(T),\bar w) + [\pi_{\bar\zeta^+} - \pi_{\bar\zeta^-}]_{|\overline{\Omega}_T} = 0 & \text{on } \overline{\Omega}, \end{cases} \tag{41}$$

$$\int_\Omega H'_w\Big(x, \bar y(T), \bar w, \bar p(0), \bar\alpha\Big)(\bar w - w)\, dx \tag{42}$$

$$+\Big\langle \pi_{\bar\zeta^+}, b \Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)} - \Big\langle \pi_{\bar\zeta^-}, a \Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)} \le \int_{\overline{\Omega}_0} \Lambda_+(z_w)\, d\pi_{\bar\zeta^+} - \int_{\overline{\Omega}_0} \Lambda_-(z_w)\, d\pi_{\bar\zeta^-}$$

for all $w \in W_{ad}$, *where* z_w *is the solution of* (7) *corresponding to* w.

Proof. Due to Theorem 2.1, there exist $\bar\alpha \ge 0$, $\bar p \in L^1(0,T;W^{1,1}(\Omega))$, $\bar\zeta \in (C_b(\overline{Q} \setminus \overline{\Omega}_0))'$, and a bounded linear transformation $\Lambda_{\bar\zeta} : C_b(\overline{Q} \setminus \overline{\Omega}_0) \longrightarrow L^\infty_{\pi_{|\bar\zeta|}}(\overline{Q})$, such that (3)–(8) are satisfied. We consider $\bar\zeta$ as an element of $\mathcal{M}(\overline{Q} \times (\overline{Q} \setminus \overline{\Omega}_0)^\#)$. The condition (4) may be rewritten as

$$\Big\langle \bar\zeta, \tau(z) - \tau(\bar y) \Big\rangle_\# \le 0 \qquad \text{for all } z \in \mathcal{C}$$

(that is, $\langle \bar\zeta, \tau(z) - \tau(\bar y)\rangle_\# \le 0$ for all $\tau(z) \in \hat{\mathcal{C}} = \{\hat z \in C((\overline{Q}\setminus\overline{\Omega}_0)^\#) \mid \tau(a) \le \hat z \le \tau(b)\}$), where $\langle\cdot,\cdot\rangle_\#$ denotes the duality pairing $\langle\cdot,\cdot\rangle_{\mathcal{M}(\overline{Q}\times(\overline{Q}\setminus\overline{\Omega}_0)^\#)\times C(\overline{Q}\times(\overline{Q}\setminus\overline{\Omega}_0)^\#)}$ and τ is the operator defined in Proposition 4.2. Therefore,

$$\Big\langle \bar\zeta, \tau(\tilde z) - \tau(\tilde y) \Big\rangle_\# \le 0 \quad \text{for all } -\frac{b-a}{2} \le \tilde z \le \frac{b-a}{2} \quad \left(\text{with } \tilde y = \bar y - \frac{b+a}{2}\right)$$

and

$$\Big\langle \bar\zeta, \tau(\tilde y) \Big\rangle_\# = \sup_{|\tilde z| \le \frac{b-a}{2}} \Big\langle \bar\zeta, \tau(\tilde z) \Big\rangle_\# = \sup_{|\phi|\le 1} \Big\langle \bar\zeta, \tau\Big(\frac{b-a}{2}\phi\Big) \Big\rangle_\#$$

$$= \sup_{\tau(|\phi|)\le\tau(1)} \Big\langle \bar\zeta, \tau\Big(\frac{b-a}{2}\Big)\tau(\phi)\Big\rangle_{\#} = \sup_{|\tau(\phi)|\le 1} \Big\langle \bar\zeta, \tau\Big(\frac{b-a}{2}\Big)\tau(\phi)\Big\rangle_{\#} = \Big\langle |\bar\zeta|, \tau\Big(\frac{b-a}{2}\Big)\Big\rangle_{\#}.$$

From the definition of $\tilde y$ and from the above equality, by a straightforward calculation, we obtain

$$\Big\langle \bar\zeta^+, \tau(\bar y - b)\Big\rangle_{\#} + \Big\langle \bar\zeta^-, \tau(a - \bar y)\Big\rangle_{\#} = 0.$$

From $a \le \bar y \le b$ (and thus $\tau(a) \le \tau(\bar y) \le \tau(b)$), we see that

$$\Big\langle \bar\zeta^+, \tau(\bar y - b)\Big\rangle_{\#} = 0, \qquad \Big\langle \bar\zeta^-, \tau(a - \bar y)\Big\rangle_{\#} = 0. \tag{43}$$

Due to Theorem 4.7, there exist two bounded linear transformations $\Lambda_{\bar\zeta^+} : C((\overline{Q} \setminus \overline{\Omega}_0)^{\#}) \longrightarrow L^\infty_{\pi_{\bar\zeta^+}}(\overline{Q})$ and $\Lambda_{\bar\zeta^-} : C((\overline{Q} \setminus \overline{\Omega}_0)^{\#}) \longrightarrow L^\infty_{\pi_{\bar\zeta^-}}(\overline{Q})$, such that

$$\Big\langle \bar\zeta^+, \tau(h)\Big\rangle_{\#} = \Big\langle \pi_{\bar\zeta^+}, h\Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} + \int_{\overline{\Omega}_0} \Lambda_{\zeta^+}(\tau(h))\, d\pi_{\bar\zeta^+}, \tag{44}$$

$$\Big\langle \bar\zeta^-, \tau(h)\Big\rangle_{\#} = \Big\langle \pi_{\bar\zeta^-}, h\Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} + \int_{\overline{\Omega}_0} \Lambda_{\zeta^-}(\tau(h))\, d\pi_{\bar\zeta^-} \tag{45}$$

for all $h \in C_b(\overline{Q} \setminus \overline{\Omega}_0)$. Let us set $\Lambda_+ = \Lambda_{\zeta^+} \circ \tau$ and $\Lambda_- = \Lambda_{\zeta^-} \circ \tau$. From (43), the results are

$$\Big\langle \pi_{\bar\zeta^+}, \bar y\Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} + \int_{\overline{\Omega}_0} \Lambda_+(\bar y)\, d\pi_{\bar\zeta^+} = \Big\langle \pi_{\bar\zeta^+}, b\Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} + \int_{\overline{\Omega}_0} \Lambda_+(b)\, d\pi_{\bar\zeta^+},$$

$$\Big\langle \pi_{\bar\zeta^-}, \bar y\Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} + \int_{\overline{\Omega}_0} \Lambda_-(\bar y)\, d\pi_{\bar\zeta^-} = \Big\langle \pi_{\bar\zeta^-}, b\Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} + \int_{\overline{\Omega}_0} \Lambda_-(a)\, d\pi_{\bar\zeta^-}.$$

Since $a \le \bar y \le b$, due to Corollary 4.9 with the above equalities, we have

$$0 \le \int_{\overline{\Omega}_0} \Lambda_+(b - \bar y)\, d\pi_{\bar\zeta^+} = \Big\langle \pi_{\bar\zeta^+}, \bar y - b\Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} \le 0,$$

$$0 \le \int_{\overline{\Omega}_0} \Lambda_-(\bar y - a)\, d\pi_{\bar\zeta^-} = \Big\langle \pi_{\bar\zeta^-}, a - \bar y\Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} \le 0.$$

Consequently, we have proved (39) and (40). We still must prove (42). With (44) and (45), it follows that

$$\begin{aligned}
&\int_{\overline{\Omega}_0} \Lambda_{\bar\zeta}(z)\, d\pi_{|\bar\zeta|} = \Big\langle \bar\zeta, \tau(z)\Big\rangle_{\#} - \Big\langle \bar\zeta, z\Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} = \Big\langle \bar\zeta, z\Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} - \Big\langle \bar\zeta, z\Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} \\
&= \Big\langle \bar\zeta^+, z\Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} - \Big\langle \bar\zeta^+, z\Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} - \Big\langle \bar\zeta^-, z\Big\rangle_{*,\overline{Q}\setminus\overline{\Omega}_0} + \Big\langle \bar\zeta^-, z\Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} \\
&= \int_{\overline{\Omega}_0} \Lambda_+(z)\, d\pi_{\bar\zeta^+} - \int_{\overline{\Omega}_0} \Lambda_-(z)\, d\pi_{\bar\zeta^-} \qquad \text{for all } z \in C_b(\overline{Q} \setminus \overline{\Omega}_0).
\end{aligned}$$

The optimality condition (42) follows from (6), from the above equality, and from

$$\int_{\overline{\Omega}_0} \Lambda_+(z_{\bar w} - \bar y)\, d\pi_{\bar\zeta^+} = \int_{\overline{\Omega}_0} \Lambda_-(z_{\bar w} - \bar y)\, d\pi_{\bar\zeta^-} = 0.$$

The proof is complete. □

Corollary 6.2. *Suppose that assumptions of Theorem* 6.1 *are satisfied. Suppose in addition that there exists* $\tilde w \in W_{ad}$ *satisfying* $a(x,0) + \tilde\epsilon \le \tilde w(x) \le b(x,0) - \tilde\epsilon$ *on* Ω (*for some* $\tilde\epsilon > 0$). *Then there exist* $\bar\alpha \ge 0$, $\bar p \in L^1(0,T;W^{1,1}(\Omega))$, $\bar\mu_a \in \mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)$, *and* $\bar\mu_b \in \mathcal{M}_b(\overline{Q}\setminus\overline{\Omega}_0)$ *such that*

$$\bar\mu_a \ge 0, \qquad \bar\mu_b \ge 0, \qquad (\bar\mu_a, \bar\mu_b, \bar\alpha) \ne 0, \tag{46}$$

$$\Big\langle \bar\mu_b, \bar y \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} = \Big\langle \bar\mu_b, b \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0}, \qquad \Big\langle \bar\mu_a, \bar y \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0} = \Big\langle \bar\mu_a, a \Big\rangle_{b,\overline{Q}\setminus\overline{\Omega}_0}, \tag{47}$$

$$\bar p \textit{ satisfies } (41) \textit{ with } \pi_{\bar\zeta}{}_{|\overline{Q}\setminus\overline{\Omega}_0} \equiv \bar\mu_b - \bar\mu_a, \tag{48}$$

$$\int_\Omega H'_w\Big(x, \bar y(T), \bar w, \bar p(0), \bar\alpha\Big)(\bar w - w)\, dx \le 0 \tag{49}$$

for all $w \in W_{ad}$ *with* $a(0) \le w \le b(0)$.

Proof. Due to Theorem 6.1, there exist $\bar\alpha \ge 0$, $\bar p \in L^1(0,T;W^{1,1}(\Omega))$, $\bar\zeta \in (C_b(\overline{Q}\setminus\overline{\Omega}_0))'$, and a bounded linear transformation $\Lambda_{\bar\zeta} : C_b(\overline{Q}\setminus\overline{\Omega}_0) \longrightarrow L^\infty_{\pi_{\bar\zeta}}(\overline{Q})$ such that (38)–(42) are satisfied.

1. We claim that $(\pi_{\bar\zeta^+}{}_{|\overline{Q}\setminus\overline{\Omega}_0}, \pi_{\bar\zeta^-}{}_{|\overline{Q}\setminus\overline{\Omega}_0}, \bar\alpha) \ne 0$. Arguing by contradiction, we suppose that $(\pi_{\bar\zeta^+}{}_{|\overline{Q}\setminus\overline{\Omega}_0}, \pi_{\bar\zeta^-}{}_{|\overline{Q}\setminus\overline{\Omega}_0}, \bar\alpha) = 0$. It follows that $\bar p \equiv 0$. With (38), we have $(\pi_{\bar\zeta^-}{}_{|\overline{\Omega}_0}, \pi_{\bar\zeta^+}{}_{|\overline{\Omega}_0}) \ne 0$, and with (42), we deduce that

$$\int_{\overline{\Omega}_0} \Lambda_-(z_w)\, d\pi_{\bar\zeta^-} - \int_{\overline{\Omega}_0} \Lambda_+(z_w)\, d\pi_{\bar\zeta^+} \le \Big\langle \pi_{\bar\zeta^-}, a \Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)} - \Big\langle \pi_{\bar\zeta^+}, b \Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)}$$

for all $w \in W_{ad}$, where z_w is the solution of (7) corresponding to w. In particular,

$$\int_{\overline{\Omega}_0} \Lambda_-(z_{\tilde w})\, d\pi_{\bar\zeta^-} - \int_{\overline{\Omega}_0} \Lambda_+(z_{\tilde w})\, d\pi_{\bar\zeta^+} \le \langle \pi_{\bar\zeta^-}, a\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)} - \langle \pi_{\bar\zeta^+}, b\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)}. \tag{50}$$

With a comparison principle, we prove that

$$z_{a(0)+\tilde\epsilon}(x,t) \le z_{\tilde w}(x,t) \le z_{b(0)-\tilde\epsilon}(x,t) \quad \text{for all } (x,t) \in \overline{Q}\setminus\overline{\Omega}_0.$$

From Corollary 4.9, it follows that

$$\Big\langle \pi_{\bar\zeta^+}, a \Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)} + \tilde\epsilon\pi_{\bar\zeta^+}(\overline{\Omega}_0) \le \int_{\overline{\Omega}_0} \Lambda_+(z_{\tilde w})\, d\pi_{\bar\zeta^+} \le \Big\langle \pi_{\bar\zeta^+}, b \Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)} - \tilde\epsilon\pi_{\bar\zeta^+}(\overline{\Omega}_0),$$

$$\Big\langle \pi_{\bar\zeta^-}, a \Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)} + \tilde\epsilon\pi_{\bar\zeta^-}(\overline{\Omega}_0) \le \int_{\overline{\Omega}_0} \Lambda_-(z_{\tilde w})\, d\pi_{\bar\zeta^-} \le \Big\langle \pi_{\bar\zeta^-}, b \Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)} - \tilde\epsilon\pi_{\bar\zeta^-}(\overline{\Omega}_0),$$

and thus

$$\begin{aligned}&\Big\langle \pi_{\bar\zeta^-}, a\Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)} - \Big\langle \pi_{\bar\zeta^+}, b\Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)}\\ &\le \int_{\overline{\Omega}_0} \Lambda_-(z_{\tilde w})\, d\pi_{\bar\zeta^-} - \int_{\overline{\Omega}_0} \Lambda_+(z_{\tilde w})\, d\pi_{\bar\zeta^+} - \tilde\epsilon[\pi_{\bar\zeta^-}(\overline{\Omega}_0) + \pi_{\bar\zeta^+}(\overline{\Omega}_0)],\end{aligned}$$

which is in contradiction with (50). By setting $\bar\mu_a \equiv \pi_{\bar\zeta^-}{}_{|\overline{Q}\backslash\overline{\Omega}_0}$ and $\bar\mu_b \equiv \pi_{\bar\zeta^+}{}_{|\overline{Q}\backslash\overline{\Omega}_0}$, we obtain (46), (47), and (48).

2. Let $a(0) \le w \le b(0)$. With a comparison principle and Corollary 4.9, it follows that

$$\begin{aligned}&\int_{\overline{\Omega}_0} \Lambda_+(z_w)\, d\pi_{\bar\zeta^+} \le \int_{\overline{\Omega}_0} \Lambda_+(z_{b(0)})\, d\pi_{\bar\zeta^+} = \Big\langle \pi_{\bar\zeta^+}, b\Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)},\\ &\int_{\overline{\Omega}_0} \Lambda_-(z_{a(0)})\, d\pi_{\bar\zeta^-} = \Big\langle \pi_{\bar\zeta^-}, a\Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)} \le \int_{\overline{\Omega}_0} \Lambda_-(z_w)\, d\pi_{\bar\zeta^-}.\end{aligned}$$

Taking (42) into account, we obtain

$$\begin{aligned}&\int_\Omega H'_w(x, \bar y(T), \bar w, \bar p(0), \bar\alpha)(\bar w - w)\, dx\\ &\le \Big\langle \pi_{\bar\zeta^-}, a\Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)} - \Big\langle \pi_{\bar\zeta^+}, b\Big\rangle_{\mathcal{M}(\overline{\Omega}_0)\times C(\overline{\Omega}_0)} + \int_{\overline{\Omega}_0} \Lambda_+(z_w)\, d\pi_{\bar\zeta^+} - \int_{\overline{\Omega}_0} \Lambda_-(z_w)\, d\pi_{\bar\zeta^-}\\ &\le 0 \qquad \text{for all } w \in W_{ad} \text{ with } \ a(0) \le w \le b(0).\end{aligned}$$

The proof is complete. □

Acknowledgment. The authors would like to thank E. Casas for helpful remarks during the preparation of this paper.

REFERENCES

[1] N. Arada and J. P. Raymond, *Dirichlet boundary control of semilinear elliptic equations*, in 18th IFIP TC7 Conference on Modelling and Optimization, Detroit, MI, 1997.

[2] N. Arada and J. P. Raymond, *Dirichlet Boundary Control of Semilinear Parabolic Equations, Part* 2: *Problems with Pointwise State Constraints*, Technical report 98.17, MIP, Toulouse, France, submitted.

[3] V. Barbu, *Analysis and Control of Nonlinear Infinite Dimentional Systems*, Academic Press, New York, 1993.

[4] J. F. Bonnans and E. Casas, *An extension of Pontryagin's principle for state-constrained optimal control of semilinear elliptic equations and variational inequalities*, SIAM J. Control Optim., 33 (1995), pp. 274–298.

[5] E. Casas, *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations*, SIAM J. Control Optim., 35 (1997), pp. 1297–1327.

[6] J. B. Conway, *A Course in Functional Analysis*, Springer-Verlag, Berlin, Heidelberg, New York, 1990.

[7] R. J. DiPerna and A. J. Majda, *Oscillation and concentrations in the weak solutions of the incompressible fluid equations*, Comm. Math. Phys., 108 (1987), pp. 667–689.

[8] N. Dunford and J. T. Schwartz, *Linear Operators*, Part 1, Interscience Publishers, New York, London, 1958.

[9] H. O. Fattorini, *Optimal control problems with state constraints for semilinear distributed-parameter systems*, J. Optim. Theory Appl., 88 (1996), pp. 25–59.

[10] H. O. Fattorini and T. Murphy, *Optimal control problems for nonlinear parabolic boundary control systems: The Dirichlet boundary condition*, Differential Integral Equations, 7 (1994), pp. 1367–1388.

[11] X. J. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser, Boston, Basel, Berlin, 1995.

[12] B. Mordukhovich and K. Zhang, *Dirichlet boundary control of parabolic systems with pointwise state constraints*, in International Conference on Control and Estimations of Distributed Parameter Systems, Internat. Ser. Numer. Math., W. Desch, F. Kappel, and K. Kunisch, eds., Birkhäuser-Verlag, Basel, Switzerland, 1997, pp. 231–246.

[13] B. Mordukhovich and K. Zhang, *Minimax control of parabolic systems with Dirichlet boundary condition and state constraints*, Appl. Math. Optim., 36 (1997), pp. 323–360.

[14] J. P. Raymond, *Nonlinear boundary control of semilinear parabolic equations with pointwise state constraints*, Discrete Contin. Dynam. Systems, 3 (1997), pp. 341–370.

[15] J. P. Raymond and H. Zidani, *Hamiltonian Pontryaguin's principles for control problems governed by semilinear parabolic equations*, Appl. Math. Optim., 39 (1999), pp. 143–177.

[16] J. P. Raymond and H. Zidani, *Pontryagin's principles for state-constrained control problems governed by parabolic equations with unbounded controls*, SIAM J. Control Optim., 36 (1998), pp. 1853–1879.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 4, pp. 1033–1047

ON THE WEAK CLOSURE OF SETS OF FEASIBLE STATES FOR LINEAR ELLIPTIC EQUATIONS IN THE SCALAR CASE*

ULDIS RAITUMS†

Abstract. A suitable description of the weak closure of feasible states is given for the family of equations

$$\mathrm{div} A(x)\nabla u = f(x) \text{ in } \Omega\ ,\ u \in H_0^1(\Omega),\ \Omega \subset \mathbf{R}^n,$$

with $A \in \mathcal{M}$, where the set $\mathcal{M}$ consists of all measurable symmetric matrices whose eigenvalues at almost every $x \in \Omega$ belong to a given finite set $\{(\lambda_1^1, \dots, \lambda_n^1); \dots; (\lambda_1^N, \dots, \lambda_n^N)\} \subset \mathbf{R}^n$ and which satisfy additional constraints on the measure of sets where the eigenvalues of A are equal to some $(\lambda_1^i, \dots, \lambda_n^i)$, $i = 1, \dots, N$. Applications to optimal control problems are also considered.

Key words. elliptic equation, optimal control, feasible states, weak closure

AMS subject classifications. 49J20, 49M20

PII. S0363012997324326

1. Introduction.

For optimal control problems of the type

$$I(u) \;\to\; \min,\ \ u \in H_0^1(\Omega),\ \ A \in \mathcal{M}, \tag{1.1}$$

$$\mathcal{A}u \;:=\; \mathrm{div} A\nabla u = f \ \text{ in } \Omega, \tag{1.2}$$

where $\mathcal{M}$ is a given set of measurable symmetric positive definite matrices and I is a weakly continuous functional, it is very natural to extend the set $\mathcal{M}$ to the set $G\mathcal{M}$ which corresponds to the G-closure of the set of admissible operators of the type (1.2) defined by $A \in \mathcal{M}$. Since analytical descriptions of $G\mathcal{M}$ are known only for a few cases, the problem of the G-closure can be substituted by a problem of finding another larger set $\mathcal{M}'$ of matrices, $\mathcal{M} \subset \mathcal{M}'$, such that the set $Z(\mathcal{M}', f)$ of all solutions of (1.2) with $A \in \mathcal{M}'$ is equal to the closure in the weak topology of $H_0^1(\Omega)$ of the set $Z(\mathcal{M}, f)$ of all solutions of (1.2) with $A \in \mathcal{M}$.

If such a set $\mathcal{M}'$ is found, then from the point of view of solving the problem (1.1)–(1.2) there is not a great difference if one uses the set $G\mathcal{M}$ or the set $\mathcal{M}'$. Only two questions remain: how to interpret an optimal solution from $\mathcal{M}'$ and how to construct a minimizing sequence for the original problem by using the knowledge of the optimal solution from $\mathcal{M}'$.

In this paper, we consider the problem (1.1)–(1.2) with $\mathcal{M} = \mathcal{M}(\mathbf{d})$, where $\mathcal{M}(\mathbf{d})$ is defined by means of a given finite set of matrices $\{D_1; \dots; D_N\} \subset \mathrm{R}^{n\times n}$, a vector $\mathbf{d} = (d_1, \dots, d_N) \in \mathrm{R}^N$, and

*Received by the editors July 11, 1997; accepted for publication (in revised form) July 30, 1998; published electronically April 7, 1999. This research was supported in part by Latvian Council of Sciences grant 96.0248.

http://www.siam.org/journals/sicon/37-4/32432.html

†Institute of Mathematics and Computer Science, University of Latvia, LV-1459 Riga, Latvia (raitums@cclu.lv).

$$\begin{aligned}\mathcal{M}(\mathbf{d}) \;=\; \Big\{ & A \in [L_2(\Omega)]^{n\times n} \mid \\ & A(x) = \sum_{i=1}^{N} \theta_i(x) Q_i(x) D_i Q_i^{-1}(x) \text{ almost everywhere (a.e.) } x \in \Omega, \\ & Q_i(x) \in \mathcal{O} \;\text{ a.e. } x \in \Omega, \\ & \theta_i(x) = 0 \text{ or } \;1,\; i = 1, \dots, N,\;\; \theta_1(x) + \cdots + \theta_N(x) \\ & = 1 \;\text{ a.e. } x \in \Omega, \\ & \int_\Omega \theta_i(x)dx \le d_i,\; i = 1, \dots, N \Big\}.\end{aligned}$$

Here $\mathcal{O}$ is the set of all orthogonal constant $n \times n$ matrices.

From the point of view of practical problems the matrices D_i play the role of properties of given N materials (not necessary isotropic) and the additional constraints on θ_i play the role of constraints on the volumes occupied by each material.

The above-formulated problem and its analogues have been studied by many authors. Most of these investigations were based on the evaluation of the G-closure of the set of admissible matrices. For problems without restrictions on volumes occupied by each material the corresponding G-closures were obtained by Frankfort and Murat [2] ($n = N = 2$, anisotropic materials) and Lurie and Cherkaev [4] (isotropic materials, arbitrary n and m). The characterization of mixtures using two isotropic materials in given proportions were given in Lurie and Cherkaev [4], [5], Murat and Tartar [8], and Tartar [12]. Many investigations have been devoted to bounds on effective properties of composites. We mention here Milton and Kohn [7], Milton [6], Nesi [9], Zhikov, Kozlov, and Oleinik [15], and references therein. Nevertheless, as far as we know, an analytical description of the G-closure for the case of N, $N \geq 3$, does not exist, given isotropic materials taken in prescribed proportion.

Various aspects of related optimal control problems are discussed in Lurie [3]. In Cherkaev [1] the idea was developed that from the point of view of optimization it is necessary to find out only a part of the G-closure which really defines the only candidates to optimality.

The first results different from pure extensions via the G-closure were obtained by Raitums [10] for the case without restrictions on the amounts of materials. Almost at the same time similar ideas were developed in Murat and Tartar [8] for the case of two isotropic materials, with restrictions on amounts of materials. In Tartar [13] it was indicated that these ideas in a natural way can be extended to the case of an arbitrary number of anisotropic materials.

We must especially mention the recently published paper by Tartar [14], who via a different approach obtained a successful extension of the set $\mathcal{M}(\mathbf{d})$ which coincides with the set $W\mathcal{M}(\mathbf{d})$ (see section 2 for the precise definition) described in our paper. The main difference between our approach and Tartar's [14] is that we give an explicit analytical procedure (proofs of Lemma 3.6 and Theorem 2.1 below) which describes how for a given pair $(A_*, u_*) \in W\mathcal{M}(\mathbf{d}) \times H_0^1(\Omega)$,

$$\operatorname{div} A_* \nabla u_* = f \text{ in } \Omega,$$

a sequence $\{A_k\} \subset \mathcal{M}(\mathbf{d})$ (rank 2 laminated structure) is constructed such that the corresponding sequence $\{u_k\}$ of solutions of

$$\operatorname{div} A_k \nabla u_k = f \text{ in } \Omega, \qquad k = 1, 2, \ldots,$$

converges to u_* weakly in $H_0^1(\Omega)$.

In the following sections we give an explicit description of a set of matrices $W\mathcal{M}(\mathbf{d})$ such that for every fixed $f \in H^{-1}(\Omega)$ the set $Z(W\mathcal{M}(\mathbf{d}), f)$ of all solutions of (1.2) with $A \in W\mathcal{M}(\mathbf{d})$ is equal to the closure in the weak topology of $H_0^1(\Omega)$ of the set of all solutions of (1.2) with $A \in \mathcal{M}(\mathbf{d})$. Moreover, we show that the set $W\mathcal{M}(\mathbf{d})$ defines a G-closed family of operators and that for every solution u_0 of (1.2) with some $A_0 \in W\mathcal{M}(\mathbf{d})$ there exists a sequence of matrices $\{A_k\} \subset \mathcal{M}(\mathbf{d})$ which defines a rank 2 laminated structure and G-converges to a matrix A_* such that

$$A_*(x)\nabla u_0(x) = A_0(x)\nabla u_0(x) \text{ a.e. } x \in \Omega.$$

2. Notation and statements of results. Let $n \geq 2$ be an integer, let R^n be the n-dimensional Euclidean space with elements $x = (x_1, \ldots, x_n)$, $\xi = (\xi_1, \ldots, \xi_n)$, and let $\Omega \subset \mathrm{R}^n$ be a bounded domain with a uniformly Lipschitz boundary $\partial\Omega$.

Throughout this paper we shall use the following notations:

$\lvert . \rvert, \langle . \rangle$	the norm and the scalar product in R^n, respectively.
$\lvert E\rvert$ or meas$E$	the Lebesgue measure of a measurable subset $E \subset \Omega$.
$\mathcal{O}$	the set of all constant $n \times n$ orthogonal matrices.
Q	a constant orthogonal matrix or a matrix function $Q = Q(x),\ x \in \Omega$, with values $Q(x) \in \mathcal{O}$ a.e. $x \in \Omega$.
ν, μ	fixed positive constants with $0 < \nu < \mu$.
$M(\nu, \mu)$	the set of all symmetric constant $n \times n$ matrices D such that $\langle D\xi, \xi\rangle \geq \nu\lvert\xi\rvert^2 \ \forall \xi \in \mathrm{R}^n, \quad \lvert D\xi\rvert \leq \mu\lvert\xi\rvert \ \forall \xi \in \mathrm{R}^n$.
$\mathcal{M}(\nu, \mu)$	the set $\{A \in [L_2(\Omega)]^{n\times n} \mid A(x) \in M(\nu,\mu) \text{ a.e. } x \in \Omega\}$.
$\lambda_j(A)(x)$	eigenvalues of a matrix $A \in \mathcal{M}(\nu,\mu)$ at a point x arranged in the increasing order, i.e., $\nu \leq \lambda_1(A)(x) \leq \cdots \leq \lambda_n(A)(x) \leq \mu$.
$\lambda_j(A)$	the corresponding functions with values at a point $x \in \Omega$ equal to $\lambda_j(A)(x),\ j = 1, \ldots, n$, respectively.
$\mathbf{l}_j$	vector functions with values in R^n.
$w$lim	the limit in the weak topology of $L_2(\Omega)$ or its Cartesian products.
G-limit	the limit in the sense of the G-convergence.
diag$(a_1, \ldots, a_n)$	a diagonal matrix $\begin{pmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_n \end{pmatrix}$.
$\mathbf{a} \leq \mathbf{b}$	for vectors $\mathbf{a} = (a_1, \ldots, a_n)$, $\mathbf{b} = (b_1, \ldots, b_n)$ the inequality $\mathbf{a} \leq \mathbf{b}$ means that $a_j \leq b_j,\ j = 1, \ldots, n$.
$A \leq B$	for matrices $A, B \in \mathcal{M}(\nu,\mu)$ the inequality $A \leq B$. means that $\langle (B(x) - A(x))\,\xi, \xi\rangle \geq 0 \ \forall \xi \in \mathrm{R}^n$ a.e. $x \in \Omega$.

Consider elliptic operators

$$\begin{aligned} \mathcal{A} &: H_0^1(\Omega) \to H^{-1}(\Omega), \\ \mathcal{A}u &:= \operatorname{div} A\nabla u \end{aligned} \tag{2.1}$$

defined by means of matrices $A \in \mathcal{M}(\nu, \mu)$. To emphasize that an operator $\mathcal{A}$ of the kind (2.1) is defined by a matrix A we shall use the notation $\mathcal{A}(A)$.

We shall say that a sequence of matrices $\{A_k\} \subset \mathcal{M}(\nu, \mu)$ G-converges to a matrix A_0 if and only if the corresponding sequence of operators $\{\mathcal{A}(A_k)\}$ G-converges to the operator $\mathcal{A}(A_0)$. For more details of the G-convergence see Zhikov, Kozlov, and Oleinik [15].

It is well known (see, for instance, Zhikov, Kozlov, and Oleinik [15] or Raitums [11]) that every subset $\mathcal{M} \subset \mathcal{M}(\nu, \mu)$ possesses a G-closure which belongs to $\mathcal{M}(\nu, \mu)$. We shall denote by $G\mathcal{M}$ the G-closure of a set $\mathcal{M} \subset \mathcal{M}(\nu, \mu)$. We remind the reader that $G\mathcal{M}$ consists of all matrices B such that there exists a sequence $\{A_k\} \subset \mathcal{M}$ which G-converges to B.

The G-convergence of a sequence of matrices $\{A_k\}$ to a matrix A_0 shall be denoted by

$$A_k \xrightarrow{G} A_0 \ \text{ or } \ A_0 = G - \text{limit} A_k.$$

We shall say that a matrix A_0 is a rank 1 composite (or that A_0 is obtained by means of a rank 1 laminated structure) if there exists a sequence of matrices A^s, $s = 1, 2, \dots$, such that

(i) $A^s \to A_0$ strongly in $[L_2(\Omega)]^{n \times n}$ as $s \to \infty$;

(ii) for every matrix A^s there exists a sequence $\{A_k^s\} \subset \mathcal{M}(\nu, \mu)$ of admissible matrices which G-converges to A^s and for almost every $x_0 \in \Omega$ there exists a neighborhood $E^s(x_0)$ of x_0 and a constant vector $\mathbf{l}^s \in \mathrm{R}^n$ such that

$$A_k^s(x) = A_k^s(\langle x, \mathbf{l}^s \rangle) \ \text{ in } \ E^s(x_0), \qquad k = 1, 2, \dots.$$

We shall say that a matrix A_0 is a rank 2 composite (or that A_0 is obtained by means of a rank 2 laminated structure) if A_0 is constructed from rank 1 composites in the same way as rank 1 composites are constructed from admissible matrices.

Let $N \geq 2$ be an integer, let $D_i \in M(\nu, \mu)$, $i = 1, \dots, N$, be diagonal matrices,

$$D_i = \text{diag}(\lambda_1(D_i), \dots, \lambda_n(D_i)), \ \ i = 1, \dots, N,$$

and let $\mathbf{d} = (d_1, \dots, d_N)$ be a vector with nonnegative components such that

$$d_1 + \dots + d_N \geq |\Omega|.$$

We introduce the following sets:

$$\begin{aligned} S = \{\theta \in [L_2(\Omega)]^N \ \ | \ \ & \theta = (\theta_1, \dots, \theta_N), \ \ \theta_i(x) = 0 \ \text{ or } 1, \ \ i = 1, \dots, N, \\ & \theta_1(x) + \dots + \theta_N(x) = 1 \ \text{ a.e. } x \in \Omega\}, \end{aligned}$$

$$\begin{aligned} \mathcal{K}(\mathbf{d}) = \Big\{ & \mathbf{p} \in [L_2(\Omega)]^N \ \ | \ \ \mathbf{p} = (p_1, \dots, p_N), \ \ p_i(x) \geq 0 \ \text{ a.e. } x \in \Omega, \ \ i = 1, \dots, N, \\ & p_1(x) + \dots + p_N(x) = 1 \ \text{ a.e. } \ x \in \Omega, \\ & \int_\Omega p_i(x) dx \leq d_i, \ \ i = 1, \dots, N \Big\}, \end{aligned}$$

$$\begin{aligned} \mathcal{M}(\mathbf{d}) = \Big\{ & A \in \mathcal{M}(\nu, \mu) \ \ | \ \ A(x) = \sum_{i=1}^{N} \theta_i(x) Q_i(x) D_i Q_i^{-1}(x) \ \text{ a.e. } x \in \Omega, \\ & \theta = (\theta_1, \dots, \theta_N) \in S, \ \ Q_i(x) \in \mathcal{O} \ \text{ a.e. } x \in \Omega, \ \ i = 1, \dots, N, \\ & \int_\Omega \theta_i(x) dx \leq d_i, \ \ i = 1, \dots, N \Big\}, \end{aligned}$$

$$\Lambda(\mathbf{p}) = \Bigg\{ (\lambda_1, \lambda_n) \in [L_2(\Omega)]^2 \quad | \quad \lambda_1(x) \le \lambda_n(x) \ \text{ a.e. } x \in \Omega,$$
$$\left(\frac{1}{\lambda_1(x)}, \lambda_n(x)\right) \le \left(\sum_{i=1}^{N} p_i(x) \frac{1}{\lambda_1(D_i)}, \ \sum_{i=1}^{N} p_i(x) \lambda_n(D_i)\right) \ \text{ a.e. } x \in \Omega \Bigg\},$$

$$\Lambda_0(\mathbf{p}) = \Bigg\{ (\lambda_1, \lambda_n) \in \Lambda(\mathbf{p}) \quad |$$
$$\left(\frac{1}{\lambda_1(x)}, \lambda_n(x)\right) = \left(\sum_{i=1}^{N} p_i(x) \frac{1}{\lambda_1(D_i)}, \ \sum_{i=1}^{N} p_i(x) \lambda_n(D_i)\right) \ \text{ a.e. } x \in \Omega \Bigg\},$$

$$\Lambda(\mathbf{d}) = \bigcup_{\mathbf{p} \in \mathcal{K}(\mathbf{d})} \Lambda(\mathbf{p}),$$
$$\Lambda_0(\mathbf{d}) = \bigcup_{\mathbf{p} \in \mathcal{K}(\mathbf{d})} \Lambda_0(\mathbf{p}).$$

By $\Lambda(\mathbf{p})(x)$ and $\Lambda_0(\mathbf{p})(x)$ we shall denote the sets of values defined by functions from $\Lambda(\mathbf{p})$ and $\Lambda_0(\mathbf{p})$ at a point x, respectively.

Consider the optimal control problem

$$\begin{aligned} &I(u) \to \min, \\ &\operatorname{div} A \nabla u = f \ \text{ in } \Omega, \ A \in \mathcal{M}(\mathbf{d}), \ \ u \in H_0^1(\Omega), \end{aligned} \tag{2.2}$$

where the element $f \in H^{-1}(\Omega)$ and the functional I are fixed. We always suppose that the functional I is weakly continuous on $H_0^1(\Omega)$.

Such problems correspond to the optimal layout of N given anisotropic materials described by means of matrices $D_1, \ldots, D_N$, respectively, with additional constraints on the volumes occupied by each material.

Together with the problem (2.2) we will consider the problem

$$\begin{aligned} &I(u) \to \min, \\ &\operatorname{div} A \nabla u = f \ \text{ in } \Omega, \ A \in G\mathcal{M}(\mathbf{d}), \ u \in H_0^1(\Omega), \end{aligned} \tag{2.3}$$

which is the extension of the problem (2.2) via the G-closure, and the problem

$$\begin{aligned} &I(u) \to \min, \\ &\operatorname{div} A \nabla u = f \ \text{ in } \Omega, \ \ A \in W\mathcal{M}(\mathbf{d}), \ u \in H_0^1(\Omega), \end{aligned} \tag{2.4}$$

where

$$W\mathcal{M}(\mathbf{d}) = \{A \in \mathcal{M}(\nu, \mu) \quad | \quad (\lambda_1(A), \lambda_n(A)) \in \Lambda(\mathbf{d})\}. \tag{2.5}$$

It is clear that problem (2.3) possesses an optimal solution.

For a given set of matrices $\mathcal{M} \subset \mathcal{M}(\nu, \mu)$ denote by $Z(\mathcal{M}, f)$ the set of all feasible states of the state equation

$$\operatorname{div} A \nabla u = f \ \text{ in } \Omega$$

with $A \in \mathcal{M}$, i.e.,

$$Z(\mathcal{M}, f) = \{u \in H_0^1(\Omega \mid \text{there exists a matrix } A \in \mathcal{M} \text{ such that } \operatorname{div} A\nabla u = f \text{ in } \Omega\}.$$

The main results of the paper are the following.

Theorem 2.1. *For every fixed $f \in H^{-1}(\Omega)$ the problems (2.3) and (2.4) have one and the same set of feasible states, i.e.,*

$$Z\left(G\mathcal{M}(\mathbf{d}), f\right) = Z\left(W\mathcal{M}(\mathbf{d}), f\right),$$

and the set $Z\left(W\mathcal{M}(\mathbf{d}), f\right)$ is closed in the weak topology of $H_0^1(\Omega)$.

Theorem 2.2. *The set $W\mathcal{M}(\mathbf{d})$ is G-closed.*

Theorem 2.3. *Problem (2.4) has at least one optimal solution $(A_0, u_0) \in W\mathcal{M}(\mathbf{d}) \times H_0^1(\Omega)$ and there exists a matrix $A_* \in G\mathcal{M}(\mathbf{d})$ such that*

$$A_*(x)\nabla u_0(x) = A_0(x)\nabla u_0(x) \text{ a.e. } x \in \Omega.$$

The pair (A_*, u_0) is an optimal solution of the problem (2.3) and the matrix A_* is a rank 2 composite constructed by means of matrices from $\mathcal{M}(\mathbf{d})$.

3. Properties of the set $W\mathcal{M}(\mathbf{d})$.

Lemma 3.1. *The set*

$$\Lambda = \left\{ \left(\frac{1}{\lambda_1(A)}, \lambda_n(A) \right) \in [L_2(\Omega)]^2 \mid A \in W\mathcal{M}(\mathbf{d}) \right\}$$

is convex and closed.

Proof. From the definitions of the sets $\Lambda(\mathbf{p})$, $\Lambda(\mathbf{d})$, and $W\mathcal{M}(\mathbf{d})$ it immediately follows that

$$\begin{aligned} \Lambda = \Bigg\{ &\left(\frac{1}{\lambda_1}, \lambda_n \right) \in [L_2(\Omega)]^2 \mid \lambda_1(x) \le \lambda_n(x) \text{ a.e. } x \in \Omega, \\ &\left(\frac{1}{\lambda_1(x)}, \lambda_n(x) \right) \le \left(\sum_{i=1}^N p_i(x) \frac{1}{\lambda_1(D_i)}, \sum_{i=1}^N p_i(x)\lambda_n(D_i) \right) \text{ a.e. } x \in \Omega, \\ &\mathbf{p} \in \mathcal{K}(\mathbf{d}) \Bigg\}. \end{aligned}$$

The set $\mathcal{K}(\mathbf{d})$ is convex and closed; therefore the set Λ is convex and closed too. □

Lemma 3.2. *Let $\mathbf{p} \in \mathcal{K}(\mathbf{d})$ and let the sequence $\{A_k\} \subset \mathcal{M}(\mathbf{d})$ be defined as*

$$\begin{aligned} &A_k(x) = \sum_{i=1}^N \theta_i^k(x) Q_i^k(x) D_i \left(Q_i^k(x)\right)^{-1} \text{ a.e. } x \in \Omega,\ k = 1, 2, \dots, \\ &Q_i^k(x) \in \mathcal{O} \text{ a.e. } x \in \Omega,\ i = 1, \dots, N,\ k = 1, 2, \dots, \qquad (3.1) \\ &\theta^k = (\theta_1^k, \dots, \theta_N^k) \in S,\ k = 1, 2, \dots, \\ &\theta^k \rightharpoonup \mathbf{p} \text{ weakly as } k \to \infty. \end{aligned}$$

If the sequence $\{A_k\}$ G-converges to a matrix A_0 then

$$(\lambda_1(A_0), \lambda_n(A_0)) \in \Lambda(\mathbf{p}).$$

Proof. Without loss of generality we can assume that the sequences $\{A_k\}$ and $\{A_k^{-1}\}$ weakly converge to A_+ and A_-^{-1}, respectively. It is well known (see, for instance, Zhikov, Kozlov, and Oleinik [15]) that

$$A_- \le A_0 \le A_+,$$

which gives that

$$\lambda_1(A_-)(x) \le \lambda_1(A_0)(x) \le \lambda_n(A_0)(x) \le \lambda_n(A_+)(x) \text{ a.e. } x \in \Omega. \tag{3.2}$$

On the other hand, for every fixed nonnegative $\varphi \in L_2(\Omega)$

$$\begin{aligned}&\int_\Omega \frac{1}{\lambda_1(A_-)}\varphi dx \le \lim_{k\to\infty}\int_\Omega \max_{|\xi(x)|\le 1}\langle A_k^{-1}(x)\xi(x),\xi(x)\rangle\varphi dx\\ &= \lim_{k\to\infty}\int_\Omega \max_{|\xi(x)|\le 1}\left\langle \sum_{i=1}^N \theta_i^k(x)Q_i^k(x)D_i^{-1}\left(Q_i^k(x)\right)^{-1}\xi(x),\xi(x)\right\rangle\varphi dx\\ &\le \lim_{k\to\infty}\int_\Omega \sum_{i=1}^N \theta_i^k\frac{1}{\lambda_1(D_i)}\varphi dx = \int_\Omega \sum_{i=1}^N p_i\frac{1}{\lambda_1(D_i)}\varphi dx.\end{aligned}$$

Analogously, for every nonnegative $\varphi \in L_2(\Omega)$,

$$\int_\Omega \lambda_n(A_+)\varphi dx \le \int_\Omega \sum_{i=1}^N p_i\lambda_n(D_i)\varphi dx.$$

Both these estimates together with (3.2) give that

$$\begin{gathered}\frac{1}{\lambda_1(A_0)(x)} \le \frac{1}{\lambda_1(A_-)(x)} \le \sum_{i=1}^N p_i(x)\frac{1}{\lambda_1(D_i)} \text{ a.e. } x\in\Omega,\\ \lambda_n(A_0)(x) \le \lambda_n(A_+)(x) \le \sum_{i=1}^N p_i(x)\lambda_n(D_i) \text{ a.e. } x\in\Omega,\end{gathered}$$

which proves the statement of Lemma 3.2. □

LEMMA 3.3. *Let $\mathbf{p} \in \mathcal{K}(\mathbf{d})$, let $(\lambda_1^0, \lambda_n^0) \in \Lambda_0(\mathbf{p})$, and let $(\mathbf{l}_1, \dots, \mathbf{l}_n)$ be an n-tuple of measurable vector functions orthonormed at a.e. $x \in \Omega$. Let the functions $\mathbf{p}, \mathbf{l}_1, \dots, \mathbf{l}_n$ be piecewise constant with one and the same partition $\Omega = \Omega_1 \cup \dots \cup \Omega_m$.*

Then there exists a sequence of matrices $A_k \in \mathcal{M}(\mathbf{d})$, $k = 1, 2, \dots$, of the type (3.1) such that

- (i) *$\{A_k\}$ G-converges to a matrix $A_0 \in G\mathcal{M}(\mathbf{d})$ and $(\lambda_1(A_0), \lambda_n(A_0)) = (\lambda_1^0, \lambda_n^0)$ in Ω, $A_0\mathbf{l}_j = \lambda_j(A_0)\mathbf{l}_j$ in Ω, $j = 1, \dots, n$;*
- (ii) *$A_k \rightharpoonup A_+$ weakly as $k \to \infty$ and $\lambda_n(A_+) = \lambda_n^0$ in Ω;*
- (iii) *$A_k^{-1} \rightharpoonup A_-^{-1}$ weakly as $k \to \infty$ and $\lambda_1(A_-) = \lambda_1^0$ in Ω;*
- (iv) *the functions θ^k in the representation (3.1) of matrices A_k satisfy $\int_{\Omega_r}\theta_i^k dx = \int_{\Omega_r} p_i dx$, $i = 1, \dots, N$, $r = 1, \dots, m$, $k = 1, 2, \dots$, $\theta^k \rightharpoonup \mathbf{p}$ weakly as $k \to \infty$;*
- (v) *in every Ω_r, $r = 1, \dots, m$, $A_k(x) = A_k(\langle x, \mathbf{l}_1(x)\rangle)$, $k = 1, 2, \dots$.*

Proof. Let $Q = Q(x)$, $x \in \Omega$, be a matrix such that its jth column is equal to $\mathbf{l}_j$, $j = 1, \dots, n$, respectively. It is obvious that $Q(x) \in \mathcal{O}$ a.e. $x \in \Omega$. Define the matrices A_k as

$$
\begin{aligned}
&A_k(x) = \sum_{i=1}^{N} \theta_i^k(x) Q(x) D_i Q^{-1}(x) \ \text{ a.e. } x \in \Omega,\\
&(\theta_1^k, \dots, \theta_N^k) = \theta^k \in S,\\
&\int_{\Omega_r} (\theta_i^k - p_i) dx = 0, \ \ i = 1, \dots, N, \ \ r = 1, \dots, m, \ \ k = 1, 2, \dots,\\
&\theta^k(x) = \theta^k(\langle x, \mathbf{l}_1(x)\rangle) \ \text{ a.e. } x \in \Omega,\\
&\theta^k \rightharpoonup \mathbf{p} \text{ weakly as } k \to \infty.
\end{aligned}
\tag{3.3}
$$

By construction $A_k \in \mathcal{M}(\mathbf{d})$, $k = 1, 2, \dots,$ and in every Ω_r, $r = 1, \dots, m$, $A_k(x) = A_k(\langle x, \mathbf{l}_1\rangle)$, $k = 1, 2, \dots$.

The function $\mathbf{l}_1$ is constant in every Ω_r, $r = 1, \dots, m$; therefore, by the well-known formulae for layered structures and by the locality properties of the G-convergence (see, for instance, Zhikov, Kozlov, and Oleinik [15]) the sequence $\{A_k\}$ defined by (3.3) has properties (i)–(v). □

LEMMA 3.4. *Let $\mathbf{p} \in \mathcal{K}(\mathbf{d})$, let $(\lambda_1^0, \lambda_n^0) \in \Lambda_0(\mathbf{p})$ and let $(\mathbf{l}_1, \dots, \mathbf{l}_n)$ be an n-tuple of measurable vector functions orthonormed at a.e. $x \in \Omega$. Then there exists a matrix $A_0 \in G\mathcal{M}(\mathbf{d})$ such that*

(i) *A_0 is the rank 1 composite defined by means of matrices from $\mathcal{M}(\mathbf{d})$;*

(ii) *$(\lambda_1(A_0), \lambda_n(A_0)) = (\lambda_1^0, \lambda_n^0)$ in Ω;*

(iii) *the eigenvectors of A_0 are equal to $\mathbf{l}_1, \dots, \mathbf{l}_n$, respectively, i.e.,*

$$A_0 \mathbf{l}_j = \lambda_j(A_0) \mathbf{l}_j \ \ \textit{in } \Omega, \ \ j = 1, \dots, n.$$

Proof. For every integer $s = 1, 2, \dots,$ there exist piecewise approximations $\mathbf{p}^s, \mathbf{l}_1^s, \dots, \mathbf{l}_n^s$ of $\mathbf{p}, \mathbf{l}_1, \dots, \mathbf{l}_n$, respectively, with one and the same partition

$$\Omega = \Omega_1^s \cup \dots \cup \Omega_{m_s}^s$$

(the corresponding approximations $(\lambda_1^s, \lambda_n^s)$ for $(\lambda_1^0, \lambda_n^0)$ are uniquely defined by $\mathbf{p}^s$) such that

$$
\begin{aligned}
&\mathbf{p}^s \to \mathbf{p} \ \text{ strongly as } \ s \to \infty,\\
&\mathbf{l}_j^s \to \mathbf{l}_j \ \text{ strongly as } s \to \infty, \ \ j = 1, \dots, n,\\
&\int_\Omega p_i^s dx = \int_\Omega p_i dx, \ \ i = 1, \dots, N, \ \ s = 1, 2, \dots,
\end{aligned}
\tag{3.4}
$$

such that $\mathbf{p}^s \in \mathcal{K}(\mathbf{d})$, $s = 1, 2, \dots,$ and such that every n-tuple $(\mathbf{l}_1^s, \dots, \mathbf{l}_n^s)$ is orthonormed at a.e. $x \in \Omega$. Our approximations satisfy the assumptions of Lemma 3.3; hence, there exist piecewise constant matrices $A^s \in G\mathcal{M}(\mathbf{d})$, $s = 1, 2, \dots,$ such that every matrix A_s is obtained by means of a rank 1 laminated structure from matrices from $\mathcal{M}(\mathbf{d})$ and

$$
\begin{aligned}
&\lambda_1(A_s) = \left(\sum_{i=1}^{N} p_i^s \frac{1}{\lambda_1(D_i)} \right)^{-1} \ \ \text{ in } \Omega,\\
&\lambda_j(A_s) = \sum_{i=1}^{N} p_i^s \lambda_j(D_i) \ \text{ in } \Omega, \ \ j = 2, \dots, n,\\
&A_s \mathbf{l}_j^s = \lambda_j(A_s) \mathbf{l}_j^s \ \text{ in } \Omega, \ \ j = 1, \dots, n,\\
&s = 1, 2, \dots.
\end{aligned}
$$

Here we have used the fact that the eigenvalues $\lambda_j(A_0),\ j=2,\dots,n$, in the construction of A_0 in the proof of Lemma 3.3 are the limits of the sequences $\{\lambda_j(A_k)\},\ j=2,\dots,n$, respectively.

Now, from the strong convergences (3.4) it follows immediately that the sequence $\{A_s\}$ converges strongly to a matrix A_0 which has the properties (i)–(iii). In turn, the standard diagonal process (matrices A_s are G-limit matrices for sequences from $\mathcal{M}(\mathbf{d})$) gives that $A_0 \in G\mathcal{M}(\mathbf{d})$. □

LEMMA 3.5. *The set $W\mathcal{M}(\mathbf{d})$ is G-closed.*

Proof. Let the sequence $\{A_k\} \subset W\mathcal{M}(\mathbf{d})$ G-converges to a matrix A_0. Without losing generality we can assume that

$$\begin{aligned}&A_k \rightharpoonup A_+ \text{ weakly as } k\to\infty,\\ &A_k^{-1} \rightharpoonup A_-^{-1} \text{ weakly as } k\to\infty.\end{aligned}$$

Since $(\lambda_1(A_k),\lambda_n(A_k)) \in \Lambda(\mathbf{p}^k)$ for some $\mathbf{p}^k \in \mathcal{K}(\mathbf{d}),\ k=1,2,\dots$, then

$$\left(\frac{1}{\lambda_1(A_k)},\lambda_n(A_k)\right) \le \left(\sum_{i=1}^N p_i^k \frac{1}{\lambda_1(D_i)},\sum_{i=1}^N p_i^k\lambda_n(D_i)\right) \text{ in } \Omega,\quad k=1,2,\dots. \tag{3.5}$$

Let $\mathbf{p}^k \rightharpoonup \mathbf{p}^0$ weakly as $k\to\infty$. Analogously to the proof of Lemma 3.2, from estimate (3.5) we obtain that

$$\left(\frac{1}{\lambda_1(A_0)},\lambda_n(A_0)\right) \le \left(\sum_{i=1}^N p_i^0 \frac{1}{\lambda_1(D_i)},\sum_{i=1}^N p_i^0\lambda_n(D_i)\right) \text{ in } \Omega.$$

The set $\mathcal{K}(\mathbf{d})$ is weakly closed; hence, $\mathbf{p}^0 \in \mathcal{K}(\mathbf{d})$ and $(\lambda_1(A_0),\lambda_n(A_0)) \in \Lambda(\mathbf{p}^0)$, which gives the inclusion $A_0 \in W\mathcal{M}(\mathbf{d})$. □

LEMMA 3.6. *Let $\mathbf{p} \in \mathcal{K}(\mathbf{d})$ and let $(\lambda_1^0,\lambda_n^0) \in \Lambda_0(\mathbf{p})$. Then there exists a pair of functions $(\alpha,\beta) = (\alpha(t,x),\beta(t,x)),(t,x) \in [0,1]\times\Omega$, such that these functions are continuous in t for a.e. $x\in\Omega$, measurable in x and that*

$$\begin{aligned}&(\alpha(t,x),\beta(t,x)) \in \Lambda(\mathbf{p})(x) \quad a.e.\ x\in\Omega,\quad 0\le t\le 1,\\ &(\alpha(0,x),\beta(0,x)) = \left(\lambda_1^0(x),\lambda_n^0(x)\right) \quad a.e.\ x\in\Omega,\\ &\alpha(1,x)=\beta(1,x) \quad a.e.\ x\in\Omega.\end{aligned}$$

In addition, for every fixed measurable-in-x function $\theta=\theta(x),x\in\Omega$, such that $0\le\theta(x)\le 1$ a.e. $x\in\Omega$, and every fixed measurable n-tuple $(\mathbf{l}_1,\dots,\mathbf{l}_n)$ of orthonormed at a.e. $x\in\Omega$ vector functions there exists a rank 2 composite $A_\theta \in G\mathcal{M}(\mathbf{d})$ such that

$$\begin{aligned}&(\lambda_1(A_\theta)(x),\lambda_n(A_\theta)(x)) = (\alpha(\theta(x),x),\beta(\theta(x),x)) \quad a.e.\ x\in\Omega,\\ &A_\theta\mathbf{l}_j = \lambda_j(A_\theta)\mathbf{l}_j \ in\ \Omega,\quad j=1,\dots,n.\end{aligned}$$

Proof. Let Q be a matrix such that its jth column is equal to $\mathbf{l}_j,\ j=1,\dots,n$. By virtue of Lemma 3.4 and Corollary 3.1 there exists a rank 1 composite $B_0 \in G\mathcal{M}(\mathbf{d})$

such that

$$\begin{aligned}
&B_0 = Q\mathrm{diag}\,(\lambda_1(B_0),\dots,\lambda_n(B_0))\,Q^{-1} \ \text{ in } \Omega,\\
&(\lambda_1(B_0),\lambda_n(B_0)) = (\lambda_1^0,\lambda_n^0) \ \text{ in } \Omega,\\
&B_0 = \text{G-limit} \sum_{i=1}^{N} \theta_i^k Q D_i Q^{-1},\\
&\int_\Omega \theta_i^k dx = \int_\Omega p_i dx, \ i=1,\dots,N, \ \ k=1,2,\dots,\\
&\theta^k \rightharpoonup \mathbf{p} \ \text{ weakly as } k\to\infty.
\end{aligned}$$

Analogously, if we use, instead of the matrices D_i, the clusters of matrices $\{\mathrm{diag}\,(\lambda_{j_1}(D_i),\dots,\lambda_{j_n}(D_i))\}$ with different order of eigenvalues on the diagonal and using layers orthogonal to $\mathbf{l}_1$ or $\mathbf{l}_n$ we obtain rank 1 composites $B_1, B_2 \in G\mathcal{M}(\mathbf{d})$ such that

$$\begin{aligned}
&B_1 = Q\mathrm{diag}(\lambda_1^0, a,\dots,a,\lambda_n^0)Q^{-1} \ \text{ in } \Omega,\\
&B_2 = Q\mathrm{diag}(\lambda_n^0, a,\dots,a,\lambda_1^0)Q^{-1} \ \text{ in } \Omega,\\
&a = \frac{1}{n-2}\,(\lambda_2(B_0)+\dots+\lambda_{n-1}(B_0)) \ \text{ in } \Omega.
\end{aligned}$$

Both matrices B_1 and B_2 locally at a.e. $x\in\Omega$ are the rank 1 composites constructed by means of the matrices D_i (and their rotations) taken in the proportions defined by $\mathbf{p}(x)$. The only difference is that the matrix B_1 is constructed by means of layers orthogonal to $\mathbf{l}_1$ (more precisely to approximations of $\mathbf{l}_1$) but the matrix B_2 is constructed by means of layers orthogonal to $\mathbf{l}_n$.

Hence, every matrix of the type

$$\begin{aligned}
&B = \theta B_2 + (1-\theta)B_1,\\
&\theta\in L_2(\Omega), \ \ \theta(x) = 0 \text{ or } 1 \ \text{ a.e. } x\in\Omega
\end{aligned}$$

has the same local distribution of involved matrices and B is a rank 1 composite constructed by means of matrices from $\mathcal{M}(\mathbf{d})$.

Let $Q_s, \lambda_1^s, a^s, \lambda_n^s, s = 1,2,\dots,$ be piecewise approximations in the norm of $[L_2(\Omega)]^{n\times n}$ or $L_2(\Omega)$ of the matrix Q and the functions $\lambda_1^0, a, \lambda_n^0$, respectively, with one and the same partition of Ω for a chosen s. Let $\mathbf{l}_1^s$ be the first column of Q_s and let

$$\begin{aligned}
&B_1^s = Q_s\mathrm{diag}(\lambda_1^s, a^s,\dots,a^s,\lambda_n^s)Q_s^{-1} \ \text{ in } \Omega,\\
&B_2^s = Q_s\mathrm{diag}(\lambda_n^s, a^s,\dots,a^s,\lambda_1^s)Q_s^{-1} \ \text{ in } \Omega.
\end{aligned}$$

For a given $\theta\in L_2(\Omega)$, $\ 0\le\theta(x)\le 1$ a.e. $x\in\Omega$, define the sequence

$$C_k^s = \theta_{ks}B_2^s + (1-\theta_{ks})B_1^s, \ \ k=1,2,\dots,$$

where

$$\begin{aligned}
&\theta_{ks}(x) = 0 \ \text{ or } 1 \ \text{ a.e. } x\in\Omega, \ \ k=1,2,\dots,\\
&\theta_{ks} \rightharpoonup \theta \text{ weakly as } k\to\infty, \ \ s=1,2,\dots,\\
&\theta_{ks}(x) = \theta_{ks}\left(\langle x, \mathbf{l}_1^s\rangle\right), \ \ k,s = 1,2,\dots.
\end{aligned}$$

For every fixed $s = 1, 2, \dots,$ the sequence $\{C^s_k\}$ G-converges as $k \to \infty$ to the matrix

$$C^s = Q_s \mathrm{diag}\left(\left(\frac{1-\theta}{\lambda_1^s} + \frac{\theta}{\lambda_n^s}\right)^{-1}, a^s, \dots, a^s, (1-\theta)\lambda_n^s + \theta\lambda_1^s\right) Q_s^{-1} \text{ in } \Omega.$$

In turn, the sequence $\{C^s\}$ converges strongly as $s \to \infty$ to the matrix

$$C_\theta = Q \mathrm{diag}\left(\left(\frac{1-\theta}{\lambda_1^0} + \frac{\theta}{\lambda_n^0}\right)^{-1}, a, \dots, a, (1-\theta)\lambda_n^0 + \theta\lambda_1^0\right) Q^{-1} \text{ in } \Omega. \tag{3.6}$$

The matrix C_θ belongs to $G\mathcal{M}(\mathbf{d})$ and C_θ is a rank 2 composite.

Let us represent the domain Ω as $\Omega = E_1 \cup E_2$, where

$$\begin{aligned} E_1 &= \{x \in \Omega \quad | \quad a(x) \le \left(\lambda_1^0(x)\, \lambda_n^0(x)\right)^{1/2}\}, \\ E_2 &= \{x \in \Omega \quad | \quad a(x) > \left(\lambda_1^0(x)\, \lambda_n^0(x)\right)^{1/2}\}. \end{aligned}$$

Define the function $h_1 = h_1(x), x \in \Omega$, as the solution of the equation

$$\frac{1-h_1}{\lambda_1^0(x)} + \frac{h_1}{\lambda_n^0(x)} = \frac{1}{a(x)}, \quad x \in \Omega;$$

the function $h_2 = h_2(x),\ x \in \Omega$, as the solution of the equation

$$(1-h_2)\lambda_n^0(x) + h_2\lambda_1^0(x) = a(x), \quad x \in \Omega;$$

and the functions h_-, h_+ as

$$\begin{aligned} h_+(x) &= \max\{h_1(x); h_2(x)\}, \quad x \in \Omega, \\ h_-(x) &= \min\{h_1(x); h_2(x)\}, \quad x \in \Omega. \end{aligned}$$

By construction

$$\begin{aligned} 0 \le h_-(x) = h_1(x) \le h_2(x) = h_+(x) \le 1 \ \text{ in } E_1, \\ 0 \le h_-(x) = h_2(x) \le h_1(x) = h_+(x) \le 1 \ \text{ in } E_2. \end{aligned}$$

Now we are able to define the functions α and β:

$$\begin{aligned} &(\alpha(t,x), \beta(t,x)) = \left(\left(\frac{1-t}{\lambda_1^0(x)} + \frac{t}{\lambda_n^0(x)}\right)^{-1}, (1-t)\lambda_n^0(x) + t\lambda_1^0\right), \\ &\qquad x \in \Omega, \quad 0 \le t \le h_-(x), \\ &(\alpha(t,x), \beta(t,x)) = \left(a(x), (1-t)\lambda_n^0(x) + t\lambda_1^0(x)\right), \\ &\qquad x \in E_1, \quad h_-(x) < t < h_+(x), \\ &(\alpha(t,x), \beta(t,x)) = \left(\left(\frac{1-t}{\lambda_1^0(x)} + \frac{t}{\lambda_n^0(x)}\right)^{-1}, a(x)\right), \\ &\qquad x \in E_2, \quad h_-(x) < t < h_+(x), \\ &(\alpha(t,x), \beta(t,x)) = (a(x), a(x)), \ x \in \Omega, \quad h_+(x) \le t \le 1. \end{aligned} \tag{3.7}$$

By construction the pair (α, β) has the desired properties and

$$(\lambda_1(C_\theta)(x), \lambda_n(C_\theta)(x)) = (\alpha(\theta(x), x), \beta(\theta(x), x)), \ x \in \Omega, \quad 0 \le \theta(x) \le h_+(x).$$

For $0 \le \theta(x) \le h_{\text{-}}(x)$ the eigenvectors $\mathbf{1}_1(x)$ and $\mathbf{1}_n(x)$ correspond to $\lambda_1(C_\theta)(x)$ and $\lambda_n(C_\theta)(x)$, respectively. However, for $h_{\text{-}}(x) \le \theta(x) \le h_t(x)$, in E_1 the eigenvectors $\mathbf{1}_j(x), j = 1, \dots, n-1$, and $\mathbf{1}_n(x)$ and in E_2 the eigenvectors $\mathbf{1}_1(x)$ and $\mathbf{1}_j(x)$, $j = 2, \dots, n$, respectively, correspond to $\lambda_1(C_\theta)(x)$ and $\lambda_n(C_\theta)(x)$.

Since the set $\mathcal{M}(\mathbf{d})$ is invariant with respect to rotations then together with the matrix $C_\theta \in G\mathcal{M}(\mathbf{d})$ the set $G\mathcal{M}(\mathbf{d})$ contains the matrix

$$A_\theta = Q\text{diag}\left(\lambda_1(C_\theta), \dots, \lambda_n(C_\theta)\right) Q^{-1} \ \text{ in } \Omega$$

which has the needed properties. It is obvious that the matrix A_θ is a rank 2 composite too. □

4. The extended problem. In this section we give the proofs of Theorems 2.1–2.3.

To begin with we will recall some properties of symmetric matrices.

Let $A \in \mathcal{M}(\nu, \mu)$ and let $\mathbf{a}, \mathbf{b} \in [L_2(\Omega)]^n$, $\mathbf{a}(x) \ne 0,\ \mathbf{b}(x) \ne 0$ a.e. $x \in \Omega$. Suppose that

$$A\mathbf{a} = \mathbf{b} \ \text{ in } \Omega.$$

Then there exists a matrix $A' \in \mathcal{M}(\nu, \mu)$ such that

$$\begin{gathered} \lambda_1(A') = \lambda_1(A), \ \ \lambda_2(A') = \dots = \lambda_n(A') \le \lambda_n(A) \ \text{ in } \Omega, \\ A'\mathbf{a} = \mathbf{b} \ \text{ in } \Omega. \end{gathered}$$

Furthermore, there exists a family of measurable symmetric $n \times n$ matrix functions $B_\tau,\ \tau < 0$, defined as

$$\begin{aligned} &\lambda_1(B_\tau)(x) = \tau F(x) + G(x) \ \text{ a.e. } x \in \Omega, \\ &\lambda_2(B_\tau)(x) = \dots = \lambda_n(B_\tau)(x) = -\frac{1}{\tau}F(x) + G(x) \ \text{ a.e. } x \in \Omega, \\ &\mathbf{l}_1(B_\tau) = \frac{\mathbf{b} - \lambda_n(B\tau)\mathbf{a}}{|\mathbf{b} - \lambda_n(B_\tau)\mathbf{a}|} \ \text{ in } \Omega, \\ &\mathbf{l}_n(B_\tau) = \frac{\mathbf{b} - \lambda_1(B_\tau)\mathbf{a}}{|\mathbf{b} - \lambda_1(B_\tau)\mathbf{a}|} \ \text{ in } \Omega, \end{aligned} \tag{4.1}$$

where

$$F(x) = \frac{\left(|\mathbf{a}(x)|^2|\mathbf{b}(x)|^2 - (\langle\mathbf{a}(x), \mathbf{b}(x)\rangle)^2\right)^{1/2}}{|\mathbf{a}(x)|^2} \ \text{ a.e. } x \in \Omega,$$

$$G(x) = \frac{\langle\mathbf{a}(x), \mathbf{b}(x)\rangle}{|\mathbf{a}(x)|^2} \ \text{ a.e. } x \in \Omega,$$

such that

$$B_\tau\mathbf{a} = \mathbf{b} \ \text{ in } \Omega.$$

In addition, for a.e. $x \in \Omega$ there exists $\tau = \tau(x)$ such that

$$B_{\tau(x)}(x) = A'(x).$$

Here by $\mathbf{l}_1(B_\tau)$ and $\mathbf{l}_n(B_\tau)$ we denote the eigenvectors of B_τ which correspond to $\lambda_1(B_\tau)$ and $\lambda_n(B_\tau)$, respectively. It is clear that for a.e. $x \in \Omega$ the vectors $\mathbf{a}(x)$ and $\mathbf{b}(x)$ belong to the linear hull of $(\mathbf{l}_1(B_\tau), \mathbf{l}_n(B_\tau))$.

All these results can be found in Raitums [10], [11].

Proof of Theorem 2.1. Because $\mathcal{M}(\mathbf{d}) \subset W\mathcal{M}(\mathbf{d})$ and the set $W\mathcal{M}(\mathbf{d})$ is G-closed then $G\mathcal{M}(\mathbf{d}) \subset W\mathcal{M}(\mathbf{d})$ and, as a consequence,

$$Z\left(G\mathcal{M}(\mathbf{d}), f\right) \subset Z\left(W\mathcal{M}(\mathbf{d}), f\right) \ \forall\, f \in H^{-1}(\Omega).$$

Let $u_0 \in Z\left(W\mathcal{M}(\mathbf{d}), f\right)$, i.e., there exists a matrix $A_0 \in W\mathcal{M}(\mathbf{d})$ such that

$$\mathrm{div} A_0 \nabla u_0 = f \ \text{ in } \Omega.$$

Since $A_0 \in W\mathcal{M}(\mathbf{d})$ then there exists a $\mathbf{p}^0 \in \mathcal{K}(\mathbf{d})$ such that $(\lambda_1(A_0), \lambda_n(A_0)) \in \Lambda(\mathbf{p}^0)$. Let $(\lambda_1^0, \lambda_n^0) \in \Lambda_0(\mathbf{p}^0)$. We point out that the pair $(\lambda_1^0, \lambda_n^0) \in \Lambda_0(\mathbf{p}^0)$ is uniquely defined by $\mathbf{p}^0$.

By virtue of Lemma 3.6 there exists a pair of functions $(\alpha, \beta) = (\alpha(t,x), \beta(t,x))$ such that these functions are continuous in t for a.e. $x \in \Omega$ and that for a.e. $x \in \Omega$ the curve $(\alpha(t,x), \beta(t,x))\,,\ 0 \le t \le 1$, connect the point $\big(\lambda_1^0(x), \lambda_n^0(x)\big)$ with the bisectrix $\lambda_1 = \lambda_n$. This curve belongs to $\Lambda(\mathbf{p}^0)(x)$.

On the other hand, there exists a family of matrices B_τ defined by formulae (4.1) with $\mathbf{a} = \nabla u_0,\ \mathbf{b} = A_0 \nabla u_0$ such that

$$B_\tau \nabla u_0 = A_0 \nabla u_0 \text{ in } \Omega,$$

and for some $\tau = \tau(x)$

$$\big(\lambda_1(B_{\tau(x)})(x), \lambda_n(B_{\tau(x)})(x)\big) \in \Lambda(\mathbf{p}^0)(x) \text{ a.e. } x \in \Omega.$$

The matrices B_τ continuously depend on τ for a.e. $x \in \Omega$.

Because the functions $\lambda_1(B_\tau)(x)$ and $\lambda_n(B_\tau)(x)$ are increasing then for a.e. $x \in \Omega$ the curves $(\alpha(t,x), \beta(t,x))$ and $(\lambda_1(B_\tau)(x), \lambda_n(B_\tau)(x))$ intersect, i.e., there exist measurable functions $\tau = \tau(x),\ t = t(x)$ such that

$$\big(\lambda_1(B_{\tau(x)})(x), \lambda_n(B_{\tau(x)})(x)\big) = \big(\lambda_1(C_{t(x)})(x), \lambda_n(C_{t(x)})(x)\big) \text{ a.e. } x \in \Omega$$

(the matrix $C_{t(.)}$ we take from the proof of Lemma 3.6) and that

$$B_{\tau(x)}(x) \nabla u_0(x) = A_0(x) \nabla u_0(x) \text{ a.e. } x \in \Omega.$$

We remind the reader that for a.e. $x \in \Omega$ the vectors $\nabla u_0(x)$ and $A_0(x) \nabla u_0(x)$ always belong to the linear hull of $\big(\mathbf{l}_1(B_{\tau(x)})(x), \mathbf{l}_n(B_{\tau(x)})(x)\big)$.

The set $GM(\mathbf{d})$ is invariant with respect to rotations; hence, $GM(\mathbf{d})$ contains a matrix C such that its eigenvalues coincide with the eigenvalues of $C_{t(.)}$, and its eigenvectors coincide with the eigenvectors of $B_{\tau(.)}$. Thus, we have obtained a matrix C such that

(1) $C \in G\mathcal{M}(\mathbf{d})$ and C is a rank 2 composite constructed by means of matrices from $\mathcal{M}(\mathbf{d})$;
(2) $(\lambda_1(C), \lambda_n(C)) \in \Lambda(\mathbf{p}^0)$ in Ω;
(3) $C \nabla u_0 = A_0 \nabla u_0$ in Ω.

The properties of the matrix C give that

$$\mathrm{div} C \nabla u_0 = f \ \text{ in } \Omega,$$

i.e., that $u_0 \in Z(G\mathcal{M}(\mathbf{d}), f)$. Hence, $Z(W\mathcal{M}(\mathbf{d}), f) = Z(G\mathcal{M}(\mathbf{d}), f)$ for every $f \in H^{-1}(\Omega)$. Since the set $Z(G\mathcal{M}(\mathbf{d}), f)$ is closed in the weak topology of $H_0^1(\Omega)$ then the set $Z(W\mathcal{M}(\mathbf{d}), f)$ is closed in the weak topology of $H_0^1(\Omega)$ too. □

Proof of Theorem 2.2. The statement of Theorem 2.2 is equal to the statement of Lemma 3.5. □

Proof of Theorem 2.3. Since the functional I in the problem (2.4) is weakly continuous on $H_0^1(\Omega)$ and the set $Z(W\mathcal{M}(\mathbf{d}), f)$ is closed in the weak topology of $H_0^1(\Omega)$ then the functional I attains its minimum on $Z(W\mathcal{M}(\mathbf{d}), f)$ at an element u_0. By the definition of $Z(W\mathcal{M}(\mathbf{d}), f)$ there exists a matrix $A_0 \in W\mathcal{M}(\mathbf{d})$ such that

$$\mathrm{div} A_0 \nabla u_0 = f \text{ in } \Omega.$$

Hence, the pair (A_0, u_0) is an optimal solution of the problem (2.4).

In turn, if we put $A_* = C$, where the matrix C is defined in the proof of Theorem 2.1, then A_* satisfies all demands of the statement of Theorem 2.3. □

In conclusion we point out only that Lemma 3.6 and the proof of Theorem 2.1 give a possibility of constructing a minimizing sequence for the initial problem (2.2) via the knowledge of an optimal solution (A_0, u_0) of the problem (2.4).

Indeed, the knowledge of the matrix A_0 gives a $\mathbf{p}^0 \in \mathcal{K}(\mathbf{d})$ with $(\lambda_1(A_0), \lambda_n(A_0)) \in \Lambda(\mathbf{p}^0)$ and the pair $(\lambda_1^0, \lambda_n^0) \in \Lambda_0(\mathbf{p}^0)$. Lemma 5.6 gives the construction for the pair (α, β). In reality the functions α and β do not depend on the eigenvectors of the matrix A_0. In turn, the knowledge of the pair $(\nabla u_0, A_0 \nabla u_0)$ gives the necessary data for the construction of the family $\{B_\tau\}$ in the proof of Theorem 2.1.

After computing intersection points $r(x)$ for the curves $(\alpha(t,x), \beta(t,x))$ and $(\lambda_1(B_\tau)(x), \lambda_n(B_\tau)(x))$ we will know the matrix C from the proof of Theorem 2.1. Finally, the construction of the matrices C_θ from the proof of Lemma 3.6 gives the matrices $A_k \in \mathcal{M}(\mathbf{d})$, $k = 1, 2, \ldots$, for the minimizing sequence $\{(A_k, u_k)\}$ in the problem (2.2).

From the point of view of applications it is more convenient from the very beginning to start with piecewise approximations of A_0 and ∇u_0. After that all steps in the construction of the minimizing sequence of controls $\{A_k\}$ become simple and clear, especially the construction of the rank 2 composites.

Acknowledgment. The author is indebted to unknown referees who considerably improved the exposition of this paper.

REFERENCES

[1] A.V. Cherkaev, *Reducing of optimal design problems to minimal variational problems*, in Composite Media and Homogenization Theory, G. Dal Masso and D. Dell'Antonio, eds., World Scientific, Singapore, 1995, pp. 139–166.

[2] G.A. Frankfort and F. Murat, *Optimal bounds for conduction in two-dimensional, two-phase, anisotropic media*, in Non-classical Continuum Mechanics, R. J. Knops and A. A. Lacey, eds., Cambridge University Press, Cambridge, UK, 1987, pp. 197–212.

[3] K.A. Lurie, *Applied Optimal Control*, Plenum, New York, 1993.

[4] K.A. Lurie and A.V. Cherkaev, *The effective characteristics of composite materials and optimal design of construction*, Advances in Mechanics (Poland), 9 (1986), pp. 3–81 (in Russian).

[5] K.A. Lurie and A.V. Cherkaev, *Exact estimates of conductivity of composites formed by two isotropically conducting media taken in prescribed proportions*, Proc. Roy. Soc. Edinburgh Sect. A, 99 (1984), pp. 71–87.

[6] G.W. Milton, *On characterizing the set of possible effective tensors of composites: The variational method and the translation method*, Comm. Pure Appl. Math., 43 (1990), pp. 63–125.

[7] G.W. Milton and R.V. Kohn, *Variational bounds on the effective moduli of anisotropic composites*, J. Mech. Phys. Solids, 36 (1988), pp. 597–629.

[8] F. Murat and L. Tartar, *Calcul des variations et homogenisation*, in Les Methodes de l'Homogenisation: Theorie et Applications en Physique, Coll. de la Dir. des Etudes et Recherches EDF, Eyrolles, 1985, pp. 319–369.

[9] V. Nesi, *Bounds on the effective conductivity of two-dimensional composites made of $n \geq 3$ isotropic phases in prescribed volume fraction: The weighted translation method*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 1219–1239.

[10] U.E. Raitums, *On the existence of solutions in optimal coefficients control problems for linear elliptic equations*, Differentsialnye Uravneniya, 19 (1983), pp. 1040–1047 (in Russian).

[11] U.E. Raitums, *Optimal Control Problems for Elliptic Equations*, Zinatne, Riga, 1989 (in Russian).

[12] L. Tartar, *Estimations fines des coefficients homogeneises*, in Ennio De Giorgi's Colloquium, P. Kree, ed., Pitman Res. Notes Math. Ser. 125, Longman, London, 1985, pp. 168–187.

[13] L. Tartar, *Remarks on optimal design problems*, in Calculus of Variations, Homogenization and Continuous Mechanics, G. Buttazzo, G. Bouchitte, and P. Suquet, eds., World Scientific, Singapore, 1994, pp. 279–296.

[14] L. Tartar, *Remarks on the homogenization method in optimal design method*, in Homogenization and Applications to Material Science, D. Cioranescu, et al., eds., GAKUTO Internat. Ser. Math. Sci. Appl. 9, Gakkotosho, Tokyo, 1997, pp. 393–412.

[15] V.V. Zhikov, S.M. Kozlov, and O.A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 4, pp. 1048–1069

RISK-SENSITIVE CONTROL OF FINITE STATE MACHINES ON AN INFINITE HORIZON II*

WENDELL H. FLEMING† AND DANIEL HERNÁNDEZ-HERNÁNDEZ‡

Abstract. In this paper we present a sequel of [Fleming and Hernandez-Hernandez, *SIAM J. Control Optim.*, 35 (1997), pp. 1790–1810], extending those results to discrete time, output feedback (partial state information) systems. Our aim is to study both robust and risk-sensitive control of such systems on an infinite horizon. Appropriate information states are defined for each case, and dynamic programming results are obtained.

Key words. risk-sensitive control, robust control, dynamic games, information states

AMS subject classifications. 93E20, 93B36, 90D15, 90D50

PII. S0363012997321498

1. Introduction.

Risk-sensitive control provides a link between stochastic and robust control. Extensive literature can be found on this subject, covering different kinds of models. Jacobson [J] established for the first time this connection for linear systems, while for a continuous variable, finite time horizon, this idea was introduced by Whittle [W]. Nonlinear continuous variable models were considered in [F-McE, F-McE1], where Whittle's idea was given as a rigorous mathematical basis using viscosity solution methods. In addition, discrete time, finite horizon, output feedback systems were studied in [B-J, J-B-E], and the infinite horizon, state feedback case for finite state machines was treated in [F-HH]. See also [FG-M], where risk-sensitive control problems on a finite time horizon for hidden Markov models were considered.

In this paper we are concerned with the infinite horizon, partially observed, risk-sensitive control problem with finite state space and long run average cost. This problem is solved by defining an appropriate information state that turns out to be a sufficient statistic and an alternate risk-sensitive control problem with full state information. The optimal growth rate is expressed as the upper value of a stochastic dynamic game with average cost criteria. Introducing a discounted cost stochastic dynamic game (cf. [F-McE, HH-M]), we prove that its value function satisfies an Isaacs equation. By using the vanishing discount approach, it is shown that the risk-sensitive dynamic programming equation holds. We also derive an optimal output feedback control that is separated through the information state. See section 4.

It is well known that a partial state information dynamic game arises naturally when an output feedback H^{∞}-robust control is considered. In [B-J] (see also [J-B-E, J-B]) this game was studied for finite state systems on a finite time horizon. By introducing an information state, they define an equivalent dynamic game with complete state information. In section 5 we analyze this deterministic dynamic game with average cost per unit time criterion. Redefining the information state, we prove

*Received by the editors May 14, 1997; accepted for publication (in revised form) December 1, 1997; published electronically April 14, 1999.

http://www.siam.org/journals/sicon/37-4/32149.html

†Division of Applied Mathematics, Brown University, Providence, RI 02912 (whf@cfm.brown.edu). The research of this author was partially supported by NSF grant DMS-9531276 and ARO grant DAAL03-92-G0115.

‡Departamento de Matemáticas, CINVESTAV-IPN, Apartado postal 14-740, México D.F. 07000, Mexico (dher@math.cinvestav.mx). The research of this author was partially supported by CONACYT grant 3115P-E9608 and NSF grant EEC-9402384.

the existence of a solution to the corresponding Isaacs equation without going through the small noise limit of the risk-sensitive dynamic programming equation, as was done in our previous work [F-HH]. When the upper value of this game is zero, it is shown that a robust output feedback control can be obtained from the Isaacs equation. We also analyze the case where the upper value is greater than zero, relating this fact to Dower and James's definition [D-J] of finite power gain.

The paper is organized as follows. To illustrate the ideas in a simple setting, in section 2 we consider uncontrolled finite state Markov chains. The risk-sensitive index is defined, and it is expressed in terms of a normalized information state. Also, it is characterized as the optimal value of a stochastic control problem with average cost criteria. Random perturbations of a finite state machine are considered in section 3, describing the strength of the perturbations through a parameter ε. The small noise limit is studied employing results from [F-HH]. In section 4 the partially observed risk-sensitive control problem is solved using estimates for products of random matrices. Finally, in section 5 we analyze the partial state information dynamic game described above.

2. The risk-sensitive index.

In this section we define the risk-sensitive index for a discrete time partially observed Markov process. Introducing an information state, we express the risk-sensitive index in terms of this completely observed state and study its related linear eigenvalue problem. Through this section no minimizing control is considered, and we postpone up to section 4 discussions involving risk-sensitive control problems.

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be an underlying probability space, and let X be a finite set with N elements. On X consider a Markov chain $\{x_t\}$ with stationary transition probability matrix $P = (P_{xx'})$ and initial distribution θ_0. Further, let $\{y_t,\, t = 1, 2, \ldots\}$ be a stochastic process with state space Y, with Y a finite set with M elements, and with associated matrix $Q = (Q_{xy})$, where Q_{xy} is the probability of receiving message y when the state is x. The realizations of the process $\{y_t\}$ are referred to as the measurements (or observations) of the system. We denote by $\mathcal{Y}^t$ the σ-algebra $\sigma(y_1, \ldots, y_t)$, with $\mathcal{Y}^0 := \{\emptyset, \Omega\}$.

Throughout this section we assume the following.

(A) The transition matrix P is irreducible, and the matrix Q is positive (i.e., $Q_{xy} > 0$ for all x, y).

Given a nonnegative function l defined on X and the risk averse factor $\mu > 0$, the risk-sensitive index is defined by

$$\lambda := \lim_{T\to\infty} \frac{1}{\mu} \cdot \frac{1}{T} \log E_{\theta_0} \exp\left\{ \mu \sum_{t=0}^{T-1} l(x_t) \right\}. \tag{2.1}$$

Remark 2.1. Note that λ has nothing to do with the observation process, and therefore it could be studied by the methods developed in [F-HH]. However, in order to introduce the main ideas and make the presentation easier in section 4, we rewrite the expectation in (2.1) in terms of a normalized "information state" process θ_t and study by stochastic control methods its behavior as T goes to infinity. If $\theta_0 = \delta_x$ is a Dirac distribution, then (2.1) is the same as [F-HH, Formula (2.15)] with $\varepsilon = 1$.

Remark 2.2. In [B-J] (see also [J-B-E]) an output feedback risk-sensitive control problem on a finite horizon is considered, where an information state is derived that turns out to be a sufficient statistic. However, in order to solve the infinite horizon case, we have to introduce another statistic, which is just the normalization of the one introduced by Baras and James [B-J].

Let $\mathcal{G}^T$ be the σ-algebra $\sigma(x_0, \ldots, x_T; y_1, \ldots, y_T)$. Then (cf. [R-S]), there exists a probability measure $\mathcal{P}^+$ on $\mathcal{G}^T$ defined by

$$\mathcal{P}^+(x_0, x_1, \ldots, x_T; y_1, \ldots, y_T) = \prod_{t=0}^{T-1} \left[P_{x_t x_{t+1}} \cdot \frac{1}{M} \right] \theta_0(x_0) \tag{2.2}$$

such that $\mathcal{P}$ is absolutely continuous with respect to $\mathcal{P}^+$, with

$$\frac{d\mathcal{P}}{d\mathcal{P}^+} |_{\mathcal{G}^T} = \overline{L}_T =: \prod_{t=0}^{T-1} M Q_{x_{t+1} y_{t+1}}.$$

Moreover, under $\mathcal{P}^+$, y_t are independently and identically distributed random variables uniformly distributed, independent of $x_{t'}$ with $t' \leq t$, and x_t is a Markov chain with transition matrix P.

Let $V(\theta_0, T)$ be the expectation on the right-hand side of (2.1). Then, it can be written in terms of the probability measure $\mathcal{P}^+$ as

$$V(\theta_0, T) = E^+ \exp \left\{ \mu \sum_{t=0}^{T-1} l(x_t) \right\} \overline{L}_T.$$

Now we define the information state $\sigma_T \in \mathbb{R}^N$ by $\sigma_0 = \theta_0$ and

$$\sigma_T(x) = E^+ \left[I_{\{x_T = x\}} \exp \left\{ \mu \sum_{t=0}^{T-1} l(x_t) \right\} \overline{L}_T | \mathcal{Y}^T \right]. \tag{2.3}$$

It satisfies the recursion

$$\sigma_{T+1} = A(y_{T+1}) \sigma_T, \tag{2.4}$$

where $A(y)$ is the transpose matrix of $A^*(y)$ with entries

$$A^*(y)_{xx'} = M P_{xx'} Q_{x'y} e^{\mu l(x)}. \tag{2.5}$$

Then, relative to the norm $|\sigma| = \sum_{i=1}^N |\sigma^i|$, $\sigma \in \mathbb{R}^N$,

$$V(\theta_0, T) := E^+ |\sigma_T|. \tag{2.6}$$

Let $S := \{\theta \in \mathbb{R}^N : \theta^i \geq 0 \quad \text{and} \quad \sum_{i=1}^N \theta^i = 1\}$. We define $z_t := |\sigma_t|$ and the state information $\theta_t \in S$ by $\theta_t := \frac{\sigma_t}{|\sigma_t|}$, $t = 0, 1, \ldots$. Then, z_t and θ_t are solutions of the recursions

$$\begin{aligned} z_{t+1} &= z_t |A(y_{t+1}) \theta_t|, \\ z_0 &= 1, \end{aligned} \tag{2.7}$$

and

$$\theta_{t+1} = \frac{A(y_{t+1}) \theta_t}{|A(y_{t+1}) \theta_t|} =: F(\theta_t, y_{t+1}). \tag{2.8}$$

Therefore, defining $G(\theta, y) := \log |A(y)\theta|$, (2.6) can be rewritten in terms of $\{\theta_t\}$ and $\{y_t\}$ as

$$\begin{aligned} V(\theta_0, T) &= E^+ z_T \\ &= E^+ \exp \sum_{t=0}^{T-1} G(\theta_t, y_{t+1}). \end{aligned}$$

Thus, we have expressed $V(\theta_0, T)$ in terms of the completely observed state information $\{\theta_t\}$. On the other hand, with $\theta = \theta_0$,

$$V(\theta, T+1) = E^+ \left[e^{G(\theta, y_1)} V(\theta_1, T) \right]. \tag{2.9}$$

From (2.1), we expect that, for T large,

$$V(\theta, T) \approx \exp\{\rho T + W(\theta)\} \tag{2.10}$$

for some number ρ and function W. Then, from substituting formally (2.10) into (2.9), we should have

$$e^{\rho + W(\theta)} = E^+ \left[e^{G(\theta, y) + W(F(\theta, y))} \right]. \tag{2.11}$$

Note that, formally, e^W is a positive eigenfunction corresponding to the eigenvalue e^ρ of the operator

$$T\psi(\theta) = E^+ \left[e^{G(\theta, y)} \psi(F(\theta, y)) \right], \quad \psi \in C(S),$$

where $C(S)$ denotes the set of continuous functions on S.

On the other hand, if

$$\phi(x, T) := E_x \exp\left\{ \mu \sum_{t=0}^{T-1} l(x_t) \right\},$$

then (see [F-HH, Thm. 2.11])

$$\lambda = \lim_{T \to \infty} \frac{1}{\mu} \cdot \frac{1}{T} \log \phi(x, T).$$

In fact, from the Frobenius theory of positive matrices,

$$\phi(x, T) \approx e^{\mu \lambda T} \psi(x)$$

in the sense that the ratio tends to 1 as $T \to \infty$. Here ψ is a positive eigenvector corresponding to the eigenvalue $e^{\mu\lambda}$ of the matrix B with entries $B_{xx'} = e^{\mu l(x)} P_{xx'}$.

Hence,

$$V(\theta, T) = \langle \phi(\cdot, T), \theta \rangle \approx e^{\lambda T} \langle \psi, \theta \rangle,$$

and from (2.9) one gets that (2.11) holds with

$$\rho = \mu\lambda, \quad W(\theta) = \log\langle \psi, \theta \rangle.$$

Equation (2.11) has the following control interpretation. Let $P(Y)$ be the set of probability measures on Y, and define the set $\Pi \subset P(Y)$ by

$$\Pi := \left\{ \pi = (\pi^1, \dots, \pi^M) \in P(Y) : \pi^i > 0,\, i = 1, \dots, M \right\}.$$

Let ν be the uniform distribution on Y. We define the relative entropy function $I(\,\cdot\,\|\nu) : P(Y) \to \mathbb{R} \cup \{+\infty\}$ by

$$I(\pi \| \nu) = \begin{cases} \sum\limits_{i=1}^{M} \log(M \cdot \pi^i) \pi^i & \text{if} \quad \pi \in \Pi, \\ +\infty & \text{otherwise.} \end{cases}$$

Then, it can be proved that (see [D-E]), for any function $\psi : Y \to \mathbb{R}$,

$$\log \int e^{\psi} d\nu = \sup_{\pi \in \Pi} \left\{ \int \psi d\pi - I(\pi \| \nu) \right\}, \tag{2.12}$$

and the supremum is attained at the unique probability measure $\overline{\pi} \in \Pi$ given by

$$\overline{\pi}^i = \frac{\frac{1}{M} e^{\psi(y_i)}}{\int e^{\psi} d\nu}, \qquad i = 1, \ldots, M.$$

Therefore, using (2.12), and taking the logarithmic transformation on both sides of (2.11), the latter equation can be written as

$$\rho + W(\theta) = \sup_{\pi \in \Pi} \sum_{i=1}^{M} \left[W(F(\theta, y_i)) + G(\theta, y_i) - \log(M\pi^i) \right] \pi^i. \tag{2.13}$$

This equation corresponds to the dynamic programming equation of the following stochastic optimal control problem (cf. [F-HH]). The control set is Π, while the reward function is $K(\theta, \pi) = \sum_{i=1}^{M} \left[G(\theta, y_i) - \log(\pi^i M) \right] \pi^i$, with $(\theta, \pi) \in S \times \Pi$. An admissible control is a sequence $\tilde{\pi} = \{\tilde{\pi}_t\}$ of random variables taking values in Π such that $\tilde{\pi}_t$ is $\mathcal{Y}^t$-adapted for each $t = 0, 1, \ldots$. We denote by $\mathcal{A}$ the set of admissible controls. Given $\tilde{\pi} \in \mathcal{A}$, it induces a probability measure $\mathcal{P}^{\tilde{\pi}}$ on $(\Omega, \mathcal{Y}^{\infty})$ in the following way. For each cylinder set $A = [\omega \in \Omega | y_1(\omega) = i_1, \ldots, y_t(\omega) = i_t]$, we define

$$\mathcal{P}^{\tilde{\pi}}(A) := \tilde{\pi}_0(i_1) \cdot \tilde{\pi}_1(i_1)(i_2) \cdots \tilde{\pi}_{t-1}(i_1, \ldots, i_{t-1})(i_t).$$

Here $\tilde{\pi}_t$ is to be interpreted as the conditional probability distribution of y_t given $y_1, \ldots, y_{t-1}$. On the other hand, the stochastic dynamics of the state information θ_t are given by (2.8). Finally, the reward functional (to be maximized) is defined, for $\tilde{\pi} \in \mathcal{A}$ and $\theta \in S$ given, by

$$J^{\tilde{\pi}}(\theta) = \limsup_{T \to \infty} \frac{1}{T} E^{\tilde{\pi}} \sum_{t=0}^{T-1} \left[G(\theta_t, y_{t+1}) - I(\pi_{t+1} \| \nu) \right].$$

Actually, standard dynamic programming arguments show that

$$\rho = \sup_{\tilde{\pi} \in \mathcal{A}} J^{\tilde{\pi}}(\theta) \quad \text{for all} \quad \theta \in S.$$

LEMMA 2.3. *There exists a unique continuous solution (up to an additive constant) $W : S \to \mathbb{R}$ to the equation* (2.11).

Proof. As noted above, $W(\theta) = \log\langle \psi, \theta \rangle$ is one solution. Let $W_1, W_2 \in C(S)$ such that

$$e^{\rho + W_1(\theta)} = E^{+} \left[e^{G(\theta, y) + W_1(F(\theta, y))} \right]$$

and

$$e^{\rho + W_2(\theta)} = E^{+} \left[e^{G(\theta, y) + W_2(F(\theta, y))} \right].$$

Then, dividing the first equation by $e^{\rho + W_2(\theta)}$, we have that $Z(\cdot) := W_1(\cdot) - W_2(\cdot)$ satisfies the equation

$$e^{Z(\theta)} = \int e^{Z(F(\theta, y))} g(\theta, y) \nu(dy), \tag{2.14}$$

with $g(\theta, y) = \exp\{W_2(F(\theta, y)) - W_2(\theta) + G(\theta, y) - \rho\}$.

Let K be the stochastic kernel defined on S by

$$K(\theta, A) = \sum_{y \in Y} I_A(F(\theta, y))\nu(dy), \ \ \theta \in S, \ A \in \mathbb{B}(S).$$

In [K] it was proved that there exists a unique invariant measure $v \in P(S)$ corresponding to K. On the other hand, let $\tilde{K}$ be the stochastic kernel defined by

$$\tilde{K}(\theta, A) = \sum_{y \in Y} I_A(F(\theta, y))g(\theta, y)\nu(dy), \ \ \theta \in S, \ A \in \mathbb{B}(S).$$

Since $\tilde{K}$ is Feller and S is compact, there exists an invariant measure $\tilde{v} \in P(S)$ corresponding to this kernel (see Proposition 8.3.4 in [D-E]), and using the facts that v is unique and that $K(\theta, \cdot)$ is equivalent to $\tilde{K}(\theta, \cdot)$ for all $\theta \in S$, it follows that $\tilde{v}$ is unique. Then, from (2.14), we have that

$$e^{Z(\theta)} = \int e^Z \, d\tilde{v} \quad \text{for all} \quad \theta \in S. \qquad \square$$

3. Small noise limit. Let X and Y be the finite sets defined in section 2. Given the functions $f : X \to X$ and $g : X \to Y$, we consider the deterministic finite state machine

$$\left\{ \begin{array}{lcl} x_{t+1} & = & f(x_t), \ \ t = 0, 1, \ldots, \ x_0 = x, \\ y_{t+1} & = & g(x_{t+1}). \end{array} \right. \tag{3.1}$$

Further, we define a perturbed system $\sum$ by

$$\left\{ \begin{array}{lcl} x_{t+1} & = & b(x_t, \omega_t), \ \ t = 0, 1, \ldots, x_0 = x, \\ y_{t+1} & = & h(x_{t+1}, \omega_t), \end{array} \right. \tag{3.2}$$

where the exogenous inputs ω_t take values in a finite set Z and the functions $b : X \times Z \to X$ and $h : X \times Z \to Y$ are given. Also, in order to measure the magnitude of disturbances ω, we introduce the function $\vartheta : Z \to \mathbb{R}^+$.

Throughout this section we assume the following.

(A1) For each $x, x' \in X$ there exist T_1, $0 < T_1 < \infty$, and $\tilde{\omega} = (\omega_0, \ldots, \omega_{T_1 - 1}) \in Z^{T_1}$ such that for the initial condition $x_0 = x$ and input $\tilde{\omega}$, the system reaches x' after T_1 steps.

(A2) Given $x \in X$, $y \in Y$, there exists $\omega \in Z$ such that $y = h(x, \omega)$.

Let $\varphi_1 : X \times X \to \mathbb{R} \cup \{+\infty\}$ and $\varphi_2 : X \times Y \to \mathbb{R}$ be the functions defined by

$$\varphi_1(x, x') = \min_{\omega \in Z}\{\vartheta(\omega) : x' = b(x, \omega)\}$$

and

$$\varphi_2(x, y) = \min_{\omega \in Z}\{\vartheta(\omega) : y = h(x, \omega)\},$$

respectively, with the standard convention that the minimum over an empty set equals $+\infty$. The values $\varphi_1(x, x')$ and $\varphi_2(x', y)$ represent the minimum "magnitude" associated with the disturbances going from x to x' and x' to y in one time step.

We define the matrices $P^\varepsilon = (P^\varepsilon_{xx'})$, $Q^\varepsilon = (Q_{xy})$ by

$$P^\varepsilon_{xx'} = \frac{1}{Z_1^\varepsilon(x)} \exp \left\{ -\frac{\varphi_1(x,x')}{\varepsilon} \right\}$$

and

$$Q^\varepsilon_{xy} = \frac{1}{Z_2^\varepsilon(x)} \exp \left\{ -\frac{\varphi_2(x,y)}{\varepsilon} \right\},$$

where $\varepsilon > 0$ is a small noise parameter and $Z_1^\varepsilon, Z_2^\varepsilon$ are normalizing constants satisfying $\sum_{x'} P^\varepsilon_{xx'} = 1$ and $\sum_y Q^\varepsilon_{xy} = 1$, respectively.

Remark 3.1. Note that from (A1)–(A2), it follows that the matrices P^ε and Q^ε satisfy assumption (A) in section 2.

Let δ_x, $x \in X$ be the Dirac measure concentrated at x. The risk-sensitive index for $\sum$ is the real number defined by

$$\lambda^{\mu,\varepsilon} = \lim_{T\to\infty} \frac{\varepsilon}{\mu} \cdot \frac{1}{T} \log E_{\delta_x} \exp \left\{ \frac{\mu}{\varepsilon} \sum_{t=0}^{T-1} \ell(x_t) \right\}. \tag{3.3}$$

Then, Lemma 2.3 implies the existence of a unique continuous function $W^{\mu,\varepsilon} : S \to \mathbb{R}$, unique up to an additive constant, such that

$$\exp \left\{ \frac{\mu}{\varepsilon} \cdot \lambda^{\mu,\varepsilon} + W^{\mu,\varepsilon}(\theta) \right\} = E^+ \exp \{ W^{\mu,\varepsilon}(F^{\mu,\varepsilon}(\theta,y)) + G^{\mu,\varepsilon}(\theta,y) \}, \tag{3.4}$$

with $F^{\mu,\varepsilon}(\theta,y) := \frac{A^\varepsilon(y)\theta}{|A^\varepsilon(y)\theta|}$, and $G^{\mu,\varepsilon}(\theta,y) := \log |A^\varepsilon(y)\theta|$. Note that we have made explicit the dependence on the parameters ε, μ.

Now we make the following transformations. Let $\zeta : \mathbb{R}^N \to S$ defined by $\zeta_p := (\exp \frac{\mu}{\varepsilon} p)/|\exp \frac{\mu}{\varepsilon} p|$. Further, define the functions $\overline{W}^{\mu,\varepsilon} : \mathbb{R}^N \to \mathbb{R}$, $\overline{G}^{\mu,\varepsilon} : \mathbb{R}^N \times Y \to \mathbb{R}$ and $L^{\mu,\varepsilon} : \mathbb{R}^N \times Y \to \mathbb{R}^N$ by

$$\left\{ \begin{array}{lcl} \overline{W}^{\mu,\varepsilon}(p) & = & \frac{\varepsilon}{\mu} W^{\mu,\varepsilon}(\zeta_p), \\ \overline{G}^{\mu,\varepsilon}(p,y) & = & \frac{\varepsilon}{\mu} G^{\mu,\varepsilon}(\zeta_p, y), \\ L^{\mu,\varepsilon}(p,y) & = & \frac{\varepsilon}{\mu} \log A^\varepsilon(y) \exp\{\frac{\mu}{\varepsilon} p\}, \end{array} \right.$$

respectively. Here $\log p$ and e^p, for $p \in \mathbb{R}^N$, are understood to be componentwise.

Note that, for each $p \in \mathbb{R}^N$, $y \in Y$,

$$\begin{aligned} F^{\mu,\varepsilon}(\zeta_p, y) &= \frac{A^\varepsilon(y) e^{\frac{\mu}{\varepsilon} p}}{|A^\varepsilon(y) e^{\frac{\mu}{\varepsilon} p}|} \\ &= \frac{\exp\{\frac{\mu}{\varepsilon} L^{\mu,\varepsilon}(p,y)\}}{|\exp\{\frac{\mu}{\varepsilon} L^{\mu,\varepsilon}(p,y)\}|} \\ &= \zeta_{L^{\mu,\varepsilon}(p,y)}. \end{aligned}$$

Therefore, we can rewrite (3.4) as

$$\begin{aligned} \lambda^{\mu,\varepsilon} + \overline{W}^{\mu,\varepsilon}(p) = \frac{\varepsilon}{\mu} \log \sum_{y \in Y} \exp \Big\{ \frac{\mu}{\varepsilon} [\overline{W}^{\mu,\varepsilon}(L^{\mu,\varepsilon}(p,y)) \\ + \overline{G}^{\mu,\varepsilon}(p,y)] \Big\} \frac{1}{M}. \end{aligned} \tag{3.5}$$

Let $(p, q) := \sup_{x \in X}(p(x) + q(x))$ be the "sup-pairing" defined on $\mathbb{R}^N \times \mathbb{R}^N$. We define the functions $L^{\mu} : \mathbb{R}^N \times Y \to \mathbb{R}^N$ and $\overline{G}^{\mu} : \mathbb{R}^N \times Y \to \mathbb{R}$ by

$$L^{\mu}(p, y)(x') = \max_{x \in X} \left\{ \ell(x) - \frac{1}{\mu}[\varphi_1(x, x') + \varphi_2(x', y)] + p(x) \right\}$$

and

$$\overline{G}^{\mu}(p, y) = (L^{\mu}(p, y), 0) - (p, 0),$$

respectively. See [B-J].

Theorem 3.2.

(a) *$\varepsilon \to \lambda^{\mu,\varepsilon}$ is uniformly bounded.*

(b) *$L^{\mu,\varepsilon}(p, y) \to L^{\mu}(p, y)$ as $\varepsilon \to 0$ uniformly on compact subsets of $\mathbb{R}^N \times Y$.*

(c) *$\overline{G}^{\mu,\varepsilon}(p, y) \to \overline{G}^{\mu}(p, y)$ as $\varepsilon \to 0$ uniformly on compact subsets of $\mathbb{R}^N \times Y$.*

(d) *There exist a sequence $\{\varepsilon_n\}$, with $\varepsilon_n \to 0$, and a continuous function $\overline{W}^{\mu} : \mathbb{R}^N \to \mathbb{R}$, such that $\overline{W}^{\mu,\varepsilon_n}$ converges uniformly on compact subsets to $\overline{W}^{\mu}$.*

Proof. Equation (3.3) implies that $0 \leq \lambda^{\mu,\varepsilon} \leq \|\ell\|$, while part (b) follows from Theorem 3.4 in [B-J]. Let $p \in \mathbb{R}^N$, $y \in Y$. Then,

$$\begin{aligned} \overline{G}^{\mu,\varepsilon}(p, y) &= \frac{\varepsilon}{\mu} \log \left| A^{\varepsilon}(y) \cdot \frac{e^{\frac{\mu}{\varepsilon}p}}{|e^{\frac{\mu}{\varepsilon}p}|} \right| \\ &= \frac{\varepsilon}{\mu} \log |A^{\varepsilon}(y) e^{\frac{\mu}{\varepsilon}p}| - \frac{\varepsilon}{\mu} \log |e^{\frac{\mu}{\varepsilon}p}| \\ &\to (L^{\mu}(p, y), 0) - (p, 0) \quad \text{as } \varepsilon \to 0, \end{aligned}$$

where the convergence is due to the Laplace–Varadhan lemma [F-W] and part (b).

Now we shall prove part (d). Let $B^{\varepsilon} = [B^{\varepsilon}_{xx'}]$ be the matrix with entries $B^{\varepsilon}_{xx'} = e^{\frac{\mu}{\varepsilon}l(x)} P^{\varepsilon}_{xx'}$, and let $\psi^{\mu,\varepsilon}$ be the positive eigenvector (see section 2), with $\max_{x \in X} \psi(x) = 1$, corresponding to the dominant eigenvalue of B^{ε}. Then, from section 2, we may take

$$\begin{aligned} W^{\mu,\varepsilon}(\theta) &= \log\langle \psi^{\mu,\varepsilon}, \theta \rangle \\ &= \log\langle e^{\frac{\mu}{\varepsilon}\eta^{\mu,\varepsilon}}, \theta \rangle, \end{aligned} \tag{3.6}$$

with $\eta^{\mu,\varepsilon} = \frac{\mu}{\varepsilon} \log \psi^{\mu,\varepsilon}$. Moreover, Theorem 3.2 in [F-HH] implies that there exist a sequence $\{\varepsilon_n\}$, with $\varepsilon_n \to 0$, and a vector $\eta^{\mu} \in \mathbb{R}^N$ such that

$$\eta^{\mu,\varepsilon_n} \to \eta^{\mu} \quad \text{as } n \to \infty. \tag{3.7}$$

Therefore, using the Laplace–Varadhan lemma, (3.6)–(3.7) imply that

$$\overline{W}^{\mu,\varepsilon}(p) \to \overline{W}^{\mu}(p) := (p, \eta^{\mu}) - (p, 0) \tag{3.8}$$

uniformly on compact subsets of $\mathbb{R}^N$, where (p, q) is the sup-pairing. □

Theorem 3.3. *There exist a number λ^{μ} and a continuous function $\overline{W}^{\mu} : \mathbb{R}^N \to \mathbb{R}$ such that*

$$\lambda^{\mu} + \overline{W}^{\mu}(p) = \max_{y \in Y}\{\overline{W}^{\mu}(L^{\mu}(p, y)) + \overline{G}^{\mu}(p, y)\}. \tag{3.9}$$

Proof. Let $\{\varepsilon_n\}$ be a sequence as in Theorem 3.2(d). Then, part (a) of that theorem implies the existence of a subsequence of $\{\varepsilon_n\}$, which we denote again by

$\{\varepsilon_n\}$, such that $\lambda^{\mu,\varepsilon_n}$ converges to some number λ^μ as $n \to \infty$. Therefore, the Laplace–Varadhan lemma applied on the right-hand side of (3.5) yields that (3.9) holds. □

Remark 3.4. Since l is nonnegative, $\lambda^\mu \geq 0$. As seen in [F-HH, section 2] $\lambda^\mu = 0$ if and only if $\mu \leq \mu^*$, where $\mu^* > 0$ is the H^∞-norm. (Additional assumptions are needed to ensure that $\mu^* < \infty$.) Corresponding results for robust control with partial state information are discussed in section 5.

4. The risk-sensitive control problem. In this section we introduce the risk-sensitive control problem for partially observed Markov models. This problem shall be solved using the dynamic programming method. We prove the existence of a solution to the dynamic programming equation, which is a nonlinear eigenvalue problem. This equation can be transformed, using the variational formula (2.12), into the Isaacs equations for a stochastic dynamic game with average cost per unit time criterion.

Let X and Y be the same sets defined in section 1, and let U be a finite control set. Given an underlying probability space $(\Omega, \mathcal{F}, \mathcal{P})$, we shall consider controlled Markov processes x_t with state space X, initial distribution θ_0, and transition matrix $P(u) = [P(u)_{xx'}]$, with $u \in U$. The state process x_t is not observed directly, and instead, the observation process $\{y_t\}$ is available. This process takes values in Y, and its associated matrix is $Q = [Q_{xy}]$. Intuitively, $P(u)$ is the matrix of transitions from x_t to x_{t+1} if control u_t is used, and Q_{xy} is the probability of observing $y_{t+1} = y$ if $x_{t+1} = x$.

We denote by $\mathcal{U}$ the set of admissible control sequences $\widetilde{u} = \{u_t\}$, where u_t is a U-valued random variable adapted to the observations σ-field $\mathcal{Y}^t$, with $\mathcal{Y}^t = \sigma\{y_1, \dots, y_t\}$, $\mathcal{Y}^0 = \{\emptyset, \Omega\}$. Given an admissible control sequence $\widetilde{u}$, it defines a probability measure on $(\Omega, \mathcal{G}^t)$, with $\mathcal{G}^t = \sigma(x_0, \dots, x_t; y_1, \dots, y_t)$ such that

$$\mathcal{P}_{\theta_0}^{\tilde{u}}(x_0, \dots, x_t; y_1, \dots, y_t) = \prod_{t=0}^{t-1} P(u_t)_{x_t x_{t+1}} Q_{x_{t+1} y_{t+1}} \theta_0(x_0).$$

Throughout this section we assume the following.

(H) (a) There exists $T_1 > 0$ such that for every $u^{T_1} = (u_0, u_1, \dots, u_{T_1-1}) \in U^{T_1}$ and $x,\ x' \in X$ there exists $x^{T_1+1} = (x_0, x_1, \dots, x_{T_1}) \in X^{T_1}$ with $x_0 = x,\ x_{T_1} = x'$, and $\prod_{t=0}^{T_1-1} P(u_t)_{x_t x_{t+1}} > 0$.

(b) The matrix Q is positive.

Given the nonnegative cost per unit time function $l : X \times U \to \mathbb{R}$, for each $\widetilde{u} \in \mathcal{U}$ the cost functional (to be minimized) is the exponential growth criterion

$$\lambda(\tilde{u}) = \limsup_{T\to\infty} \frac{1}{T} \cdot \frac{1}{\mu} \log E_{\theta_0}^{\tilde{u}} \exp \mu \sum_{t=0}^{T-1} l(x_t, u_t), \tag{4.1}$$

where $\mu > 0$ is the risk averse parameter, and analogously to $E_{\theta_0}^{\tilde{u}}$, $\mathcal{P}_{\theta_0}^{\tilde{u}}$ denotes the probability induced by $\{x_t\}, \{y_t\}$, given $\widetilde{u} \in \mathcal{U}$ and the initial distribution θ_0.

Therefore, the goal is to find a control sequence $\widetilde{u}^*$ which minimizes $\lambda(\widetilde{u})$. We let

$$\Lambda = \inf_{\tilde{u} \in \mathcal{U}} \lambda(\widetilde{u}).$$

Paralleling section 2, we introduce an information state $\theta \in S$ and replace the original partially observed risk-sensitive control problem with an equivalent completely observed with state variable θ. Then, the dynamic programming method yields an optimal control for this original problem, which is separated through the information state.

Let $\mathcal{P}^+$ be the analogous probability measure on $\mathcal{G}^T$ given in (2.2). Under this probability measure, the controlled Markov chain $\{x_t\}$ and the observation process $\{y_t\}$ have the same properties mentioned in section 2.

Now we define, for $T = 1, 2, \dots,$

$$\sigma_T(x) = E^+ \left[I_{\{x_T = x\}} \exp\left[\mu \sum_{t=0}^{T-1} l(x_t, u_t) \right] \overline{L}_T | \mathcal{Y}^T \right],$$

and the information state $\theta_T \in S$ by

$$\theta_T = \frac{\sigma_T}{|\sigma_T|}.$$

Both σ_T and θ_T satisfy the recursions (2.4) and (2.8), substituting $A(y)$ with $A(y, u)$, which in this case has entries

$$A(y, u)_{xx'} = MP(u)_{x'x} Q_{xy} e^{\mu l(x', u)}.$$

Applying the same substitution in the definitions of G and F, one gets that (4.1) can be written as

$$\lambda(\widetilde{u}) = \limsup_{T \to \infty} \frac{1}{T} \frac{1}{\mu} \log E^+_\theta \exp\left\{ \sum_{t=1}^{T} G(\theta_{t-1}, u_{t-1}, y_t) \right\}.$$

The corresponding dynamic programming equation (cf. Thm. 4.6 in [F-HH]) for this control problem is given by

$$\exp\{\Upsilon + W(\theta)\} = \min_{u \in U} \left[E^+ \exp\{G(\theta, u, y) + W(F(\theta, u, y))\} \right]. \tag{4.2}$$

As in section 2, let ν denote the uniform distribution on Y.

THEOREM 4.1 (verification theorem). *Let $\Upsilon \in \mathbb{R}$, $W \in C(S)$ be a solution to the dynamic programming equation* (4.2). *Then*

(a) $\Upsilon \leq \mu\lambda(\widetilde{u})$ *for all* $\widetilde{u} \in \mathcal{U}$;

(b) *let* $\widetilde{u}^* = \{u_t^*\} \in \mathcal{U}$ *such that, for each* $t = 0, 1, \dots,$ $\tilde{u}_t^* = u^*(\theta_t)$, *where* $u^* : S \to U$ *is a function with*

$$u^*(\theta) \in \operatorname*{argmin}_{u \in U} [E^+ \exp\{W(F(\theta, u, y)) + G(\theta, u, y)\}].$$

Then $\Upsilon = \mu\lambda(\tilde{u}^*)$.

Proof. Given $\widetilde{u} \in \mathcal{U}$, $T > 0$, the dynamic programming principle implies that

$$\exp\{\Upsilon + W(\theta)\} \leq E^+ \exp\left\{ \sum_{t=0}^{T-1} G(\theta_t, u_t, y_{t+1}) + W(\theta_T) \right\},$$

with equality for the control $\widetilde{u}^*$. Now, take logs on both sides of this inequality, and divide by μT. Then, noting that $W(\theta_T)$ is bounded since θ_T is in the compact simplex S, letting $T \to \infty$ the theorem follows. □

THEOREM 4.2. *There exist a number Υ and a nonnegative continuous function $W : S \to \mathbb{R}$ such that* (4.2) *holds.*

In order to prove this theorem, we shall need some preliminary results. First, we shall introduce an infinite horizon, discounted cost stochastic game. Let W_β denote the upper value function of this game. Then, once we prove that $\{W_\beta; \beta \in (0, 1)\}$ is equicontinuous and $(1 - \beta)W_\beta(\theta^*)$ is uniformly bounded, where $\theta^* \in S$ is a reference point, the theorem follows in a straightforward way using the Arzelà–Ascoli theorem.

Stochastic dynamic game. The stochastic dynamics of the state θ_t for the controlled sequence (two players) is

$$\theta_{t+1} = F(\theta_t, u_t, y_{t+1}). \tag{4.3}$$

Here, the sequence $\{u_t\} \subset U$ represents the controls that the minimizer controller (player 1) chooses at each time step, while the maximizer controller (player 2) chooses a sequence $\{\pi_t\} \subset \Gamma$, where π_t is the distribution of y_{t+1}. The information available at each time t for both players is the history of observations $\{y_1, \ldots, y_t\}$. We shall make this precise using the following strategies.

Let $\mathcal{V} := \{\vec{u} = \{\overline{u}_t\} \mid \overline{u}_0 = u_0 \in U,\ \overline{u}_t : Y^t \times \Gamma^t \to U,\ t \geq 1\}$ be the set of strategies for player 1, and let $\mathcal{W} = \{\vec{\pi} = \{\overline{\pi}_t\} \mid \overline{\pi}_0 = \pi_0(u_0) \in \Gamma,\ \overline{\pi}_t : Y^t \times U^t \to \Gamma,\ t \geq 1\}$ be the set of strategies for player 2. Then, given the strategies $\vec{u}$, $\vec{\pi}$, we obtain the control sequences $\tilde{u} = \{\tilde{u}_t\}$, $\tilde{\pi} = \{\tilde{\pi}_t\}$, with $\tilde{u}_t$, $\tilde{\pi}_t$ $\mathcal{Y}^t$-adapted, determined by $\tilde{u}_0 = u_0$, $\tilde{\pi}_0 = \overline{\pi}_0(u_0)$, $\tilde{u}_1(y_1) = \overline{u}_1(y_1, \pi_1)$, $\tilde{\pi}_1(y_1) = \overline{\pi}_1(y_1, \tilde{u}_0, \tilde{u}_1(y_1)) \ldots$. Note that player 1 chooses first at each time step t (upper game). Moreover, $(\vec{u}, \vec{\pi})$ defines a probability measure $\mathcal{P}^{\vec{u},\vec{\pi}}$ on $\mathcal{Y}^\infty$, where $\mathcal{Y}^\infty$ is the σ-algebra generated by the "cylinder sets," in the following way:

$$P^{\vec{u},\vec{\pi}}(y_1, y_2, \ldots, y_n) = \tilde{\pi}_0(u_0)(y_1)\tilde{\pi}_1(y_1)(y_2) \cdots \tilde{\pi}_{n-1}(y_1, \ldots, y_{n-1})(y_n).$$

We denote by $E^{\vec{u},\vec{\pi}}$ the corresponding expectation operator. Finally, the reward per stage function is $K : S \times U \times \Gamma \to \mathbb{R}$, defined by $K(\theta, u, \pi) := \sum_{j=1}^{M} \pi^j [G(\theta, u, y^j) - \log(\pi^j M)]$.

Then, in order to prove Theorem 4.2, we shall introduce a sequence of infinite horizon discounted games. Let $\beta \in (0, 1)$ denote the discount factor. Given $\theta \in S$, $\vec{u} \in \mathcal{V}$, and $\vec{\pi} \in \mathcal{W}$, define the payoff functional

$$I_\beta(\theta, \vec{u}, \vec{\pi}) = E^{\vec{u},\vec{\pi}} \sum_{t=0}^{\infty} \beta^t K(\theta_t, u_t, \pi_t).$$

DEFINITION 4.3. *When there exists a pair of strategies $\vec{u}^*, \vec{\pi}^*$ such that*

$$I_\beta(\theta, \vec{u}^*, \vec{\pi}) \leq I_\beta(\theta, \vec{u}^*, \vec{\pi}^*) \leq I_\beta(\theta, \vec{u}, \vec{\pi}^*) \ \text{ for all } \ \vec{u}, \vec{\pi}, \tag{4.4}$$

the value $I_\beta(\theta, \vec{u}^*, \vec{\pi}^*)$ is called the upper value of the game, and $\vec{u}^*, \vec{\pi}^*$ are referred to as a saddle point. The upper value function of this game is denoted by $V_\beta(\theta)$.

Remark 4.4. In section 5 an Elliott–Kalton-type definition of value shall be used, since it is natural for the robust control interpretation. For stochastic differential systems the analogous definition was given by Fleming and Souganidis [F-S]. However, in this case, our present definition allows us to work out the problem without using the Elliott–Kalton definition of strategy.

LEMMA 4.5. *There is a unique concave continuous solution to the Isaacs equation*

$$\begin{aligned} V_\beta(\theta) \ &= \ \inf_{u \in U} \sup_{\pi \in \Gamma} \sum_{j=1}^{M} \pi^j \left[\beta V_\beta\left(F\left(\theta, u, y^j\right)\right)\right. \\ &\qquad \left. + G(\theta, u, y^j) - \log(M\pi^j)\right], \end{aligned} \tag{4.5}$$

and it is the upper value function defined above. Furthermore, the pair of strategies $\vec{u} = \{\overline{u}_t^*\}$, $\vec{\pi} = \{\overline{\pi}_t^*\}$, with $\overline{u}_t^*(y_1, \ldots, y_t, \pi_0, \ldots, \pi_{t-1}) = u^*(\theta_t)$ and $\overline{\pi}_t^*(y_1, \ldots, y_t, u_0, \ldots, u_t) = \pi^*(\theta_t, u_t)$, where

$$u^*(\theta) \in \operatorname*{argmin}_{u \in U} \int e^{\beta V_\beta(F(\theta,u,y)) + G(\theta,u,y)} \nu(dy)$$

and

$$\pi^*[\theta,u]^j = \frac{\frac{1}{M}\, e^{\beta V_\beta(F(\theta,u,y^j))+G(\theta,u,y^j)}}{\displaystyle\int e^{\beta V_\beta(F(\theta,u,y))+G(\theta,u,y)}\nu(dy)}$$

is a saddle point.

Sketch of proof. First, note that, using (2.12), we can write equation (4.5) as

$$V_\beta = \inf_{u\in U} \log \int \exp\{\beta V_\beta(F(\theta,u,y)) + G(\theta,u,y)\}\nu(dy).$$

Now define, for $\phi \in C(S)$, the operator

$$T\phi(\theta) = \inf_{u\in U} \log \int \exp\{\beta \phi(F(\theta,u,y)) + G(\theta,u,y)\}\nu(dy).$$

Then, straightforward calculations show that T is monotonic and contractive. Therefore, by the fixed point theorem, there exists a unique $V_\beta \in C(S)$ such that $TV_\beta = V_\beta$. Moreover, standard dynamic programming arguments show that V_β is the upper value function in Definition 4.3. It remains to see that V_β is concave. Let $\psi_0 = 1$, and define recursively

$$\psi_{n+1}(\theta) = \inf_{u\in U} \int |A(y,u)\theta|\psi_n^\beta(F(\theta,u,y))\nu(dy).$$

Then, from the above arguments, $\psi_n \nearrow \psi$ uniformly, with $\psi = e^{V_\beta}$. Thus, in order to prove that V_β is concave, it is sufficient to prove that ψ_n is concave for each $n = 0, 1, \ldots$. We shall prove this by induction. For $n = 0$ this is obvious. Then, provided ψ_n is concave, it follows, from Lemmas 1 and 2 in [A], that ψ_{n+1} is also concave. □

LEMMA 4.6. (a) *For each $\beta \in (0,1)$ and $\theta \in S$,*

$$|(1-\beta)V_\beta(\theta)| \le \|G\|.$$

(b) *The family of functions $\{V_\beta\}_{\beta\in(0,1)}$ is equicontinuous.*

Proof. Let $\vec{u}^* \in \mathcal{V}$ be as in Lemma 4.5, and take $\vec{\pi} = \{\tilde{\pi}_t\} \in \mathcal{W}$, with $\tilde{\pi}_t \equiv \nu$ for all $t = 0, 1 \ldots$. Then, from (4.4), it follows that

$$V_\beta(\theta) \ge -\frac{\|G\|}{1-\beta} \quad \text{for all } \theta \in S.$$

On the other hand, since $I(\cdot\,\|\nu) \ge 0$, $V_\beta(\theta) \le \frac{\|G\|}{1-\beta}$. This proves (a).

Let $T > 0$, and let Y^T, U^T be the set of multi-indices of length T on Y and U, respectively. Given $y^T = (y_1, \ldots, y_t) \in Y^T$, $u^T = (u_0, \ldots, u_{T-1}) \in U^T$, we define

$$h(\theta, u^T, y^T) := \frac{A(y_T, u_{T-1}) \cdots A(y_1, u_0)\theta}{|A(y_T, u_{T-1}) \cdots A(y_1, u_0)\theta|}.$$

Note that

$$\begin{aligned}\theta_T &= F(\theta_{T-1}, u_{T-1}, y_T)\\ &= h(\theta, u^T, y^T).\end{aligned}$$

The asymptotic behavior of the state process θ_T relies on estimations for random products of matrices. In particular, from assumption (A), it follows that there exist constants C and $r \in (0,1)$ such that

$$\|h(\theta, u^T, y^T) - h(\widetilde{\theta}, u^T, y^T)\| \leq Cr^T \|\theta - \widetilde{\theta}\| \tag{4.6}$$

for all $y^T \in Y^T$, $u^T \in U^T$, and $\theta, \widetilde{\theta} \in S$. We refer to [F-K] (see also [K, Lemma 6.2] and [A-M, Lemma 2.2]) for its proof.

Let $\vec{u} \in \mathcal{U}$, $\vec{\pi} \in \mathcal{W}$, and fix $\theta, \widetilde{\theta} \in S$. Then,

$$\begin{aligned}\|I_\beta(\theta, \vec{u}, \vec{\pi}) - I_\beta(\widetilde{\theta}, \vec{u}, \vec{\pi})\| &\leq E^{\vec{u},\vec{\pi}} \sum_{t=0}^{\infty} \beta^t |G(\theta_t, u_t, y_{t+1}) - G(\widetilde{\theta}_t, u_t, y_{t+1})| \\ &\leq \left\| \frac{dG}{d\theta} \right\| \cdot E^{\vec{u},\vec{\pi}} \sum_{t=0}^{\infty} \beta^t \|\theta_t - \widetilde{\theta}_t\| \\ &= \left\| \frac{dG}{d\theta} \right\| \left[\|\theta - \widetilde{\theta}\| + E^{\vec{u},\vec{\pi}} \sum_{t=1}^{\infty} \beta^t |h(\theta, u^t, y^t) - h(\widetilde{\theta}, u^t, y^t)| \right].\end{aligned}$$

Here $\tilde{\theta}_t$ is the solution of (4.3) with initial state $\tilde{\theta}_0 = \tilde{\theta}$.

Then, from (4.6), it follows that

$$\|I_\beta(\theta, \vec{u}, \vec{\pi}) - I_\beta(\widetilde{\theta}, \vec{u}, \vec{\pi})\| \leq \frac{c' \| \frac{dG}{d\theta} \|}{1 - r} \|\theta - \widetilde{\theta}\|,$$

for some suitable constant c' independent of $\vec{u}$, $\vec{\pi}$. Therefore, for each $\theta, \widetilde{\theta} \in S$,

$$|V_\beta(\theta) - V_\beta(\widetilde{\theta})| \leq \frac{c' \| \frac{dG}{d\theta} \|}{1 - r} \|\theta - \widetilde{\theta}\|.$$

This completes the proof of the lemma. □

Proof of Theorem 4.2. First note that, from (4.5), (2.15) can be rewritten as

$$e^{V_\beta(\theta)} = \inf_{u \in U} \int e^{\beta V_\beta(F(\theta,u,y)) + G(\theta,y)} \nu(dy). \tag{4.7}$$

Now, let $\theta_\beta \in \operatorname{argmin}_{\theta \in S}\{V_\beta(\theta\}$, and define $Z_\beta(\theta) := V_\beta(\theta) - V_\beta(\theta_\beta)$. From Lemma 4.5 we have that $\beta \to |(1-\beta)V_\beta(\theta_\beta)|$ is uniformly bounded, and the functions $\{Z_\beta\}$ are equicontinuous. Therefore, the Arzelà–Ascoli theorem implies the existence of a sequence $\beta_n \uparrow 1$ along which $(1-\beta_n)V_\beta(\theta_{\beta_n})$ converges to a limit Υ, and Z_{β_n} converges to a limit W uniformly. Then, writing (4.7) as

$$e^{(1-\beta)V_\beta(\theta_\beta) + Z_\beta(\theta)} = \inf_{u \in U} \int e^{\beta Z_\beta(F(\theta,u,y)) + G(\theta,u,y)} \nu(dy),$$

the theorem follows from an application of the dominated convergence theorem. □

5. Deterministic dynamic game and robust control. It is well known that the output feedback robust control problem can be recast as a partially observed dynamic game problem when the state space formulation is used. This dynamic game was studied by James and Baras [J-B] for finite state machines and finite horizon. Introducing an information state, they define an alternate deterministic dynamic game with completely observed information state. In this section we shall consider this dynamic game on an infinite horizon and average cost per unit time criterion and study its relationship with the robust control problem.

Consider the deterministic finite state controlled machine defined by

$$\left\{\begin{array}{lcl} x_{t+1} & = & f(x_t, u_t), \quad t = 0, \ldots;\ x_0 = x, \\ y_{t+1} & = & g(x_{t+1}), \end{array}\right. \tag{5.1}$$

where the state x_t evolves in the finite set X, the output y_t takes values in Y, u_t takes values in U, and the functions $f : X \times U \to X$ and $g : X \to Y$ are given. Also, in order to model the influence of disturbances, we introduce a deterministic perturbation of the system (5.1). Let $b : X \times U \times Z \to X$ and $h : X \times W \to Y$ be given functions that define the dynamics of the system $\sum^u$

$$\left\{\begin{array}{lcl} x_{t+1} & = & b(x_t, u_t, \omega_t), \quad t = 0, \ldots; x_0 = x, \\ y_{t+1} & = & h(x_{t+1}, \omega_{t+1}), \end{array}\right. \tag{5.2}$$

where ω_t takes values in the finite set Z, and $x_t,\ y_t$, and u_t evolve in $X,\ Y$, and U, respectively. We assume that there is a "null disturbance" $\omega_\phi \in Z$ such that, for all $x \in X,\ u \in U$,

$$b(x, u, \omega_\phi) = f(x, u),$$
$$h(x, \omega_\phi) = g(x).$$

We also assume the following.

(H1) There exists $T_1,\ 0 \leq T_1 < \infty$, with the following property: for each $x, x' \in X$ and $\tilde{u} = (u_0, \ldots, u_{T_1 - 1}) \in U^{T_1}$, there exists $\tilde{\omega} = (\omega_0, \ldots, \omega_{T_1-1}) \in Z^{T_1}$ such that $x_0 = x$ and $x_{T_1} = x'$.

(H2) Given $x \in X,\ \ y \in Y$, there exists $\omega \in Z$ such that $y = h(x, \omega)$.

In particular, if (H1) holds with $T_1 = 1$, the following stronger condition holds.

(H1$'$) Given $x,\ x' \in X$, and $u \in U$, there exists $\omega \in Z$ such that $x' = b(x, u, \omega)$.

Let $\vartheta_1, \vartheta_2 : Z \to \mathbb{R}^+$ be functions that measure the magnitude of disturbances, and define the functions $\varphi_1 : X \times U \times X \to \mathbb{R}$ and $\varphi_2 : X \times Y \to \mathbb{R}$ by

$$\varphi_1(x, u; x') = \min_{\omega \in Z}\{\vartheta_1(\omega) : \quad x' = b(x, u, \omega)\}$$

and

$$\varphi_2(x, y) = \min_{\omega \in Z}\{\vartheta_2(\omega) : \quad y = h(x, \omega)\},$$

respectively. We assume that

$$\vartheta_1(\omega_\phi) = \vartheta_2(\omega_\phi) = 0.$$

As in section 3 we use the convention that the minimum over the empty set equals $+\infty$. Assumption (H2) implies that $\varphi_2(x, y) < \infty$, and $\varphi_1(x, u; x')$ is always finite if (H1$'$) holds.

Let $l : X \times U \to \mathbb{R}^+$ be the cost per stage function, and define the function $\text{\L}^\mu : \mathbb{R}^N \times U \times Y \to \mathbb{R}^N$ by

$$\text{\L}^\mu(p, u, y)(x') = \max_{x \in X}\left\{ l(x, u) - \frac{1}{\mu}[\varphi_1(x, u; x') + \varphi_2(x', y)] + p(x) \right\}.$$

Assumptions (H1) and (H2) imply that $L^\mu(p, u, y)(x')$ is finite (not $-\infty$).

Consider the difference equation

$$\left\{ \begin{array}{l} p_{t+1} = \text{Ł}^{\mu}(p_t, u_t, y_{t+1}),\ t = 0, 1, \ldots, \\ p_0 = p \in \mathbb{R}^N, \end{array} \right. \tag{5.3}$$

where $\tilde{u} = \{u_t\}_{t=0}^{\infty}$ and $\tilde{y} = \{y_t\}_{t=1}^{\infty}$ are sequences on U and Y and play the role of controls for the minimizer and maximizer controllers respectively. We associate the finite-time payoff functional

$$J^{\mu}(p_0, T; \tilde{u}, \tilde{y}) = (p_T, 0), \quad T > 0,$$

with each pair of control sequences.

Now, following the Elliott–Kalton-type definition of upper value, we say that $\Gamma : Y^{\infty} \to U^{\infty}$ is a strategy for the minimizer controller if, given $\tilde{u} \in U^{\infty}, \tilde{y} \in Y^{\infty}$ such that $\Gamma(\tilde{y}) = \tilde{u}$, $\Gamma(\tilde{y})_t$ depends only on $y_1, \ldots, y_t$ for each $t \geq 1$, while u_0 does not depend on $\tilde{y}$. Let $V(p_0, T)$ denote the (upper) value of this dynamic game, defined by

$$V(p_0, T) = \inf_{\Gamma} \sup_{\tilde{y}} J^{\mu}(p_0, T; \Gamma, \tilde{y}).$$

In [B-J, Thm. 4.1] it was proved that, given a strategy for the minimizer controller,

$$\max_{\tilde{y} \in Y^{\infty}} (p_T, 0) = \max_{x_0, \tilde{\omega}} \left\{ p_0(x_0) + \sum_{t=0}^{T-1} \left[l(x_t, u_t) - \frac{1}{\mu} [\vartheta_1(\omega_t) + \vartheta_2(\omega_{t+1})] \right] \right\}. \tag{5.4}$$

Remark 5.1.

(a) In section 4 we considered *normalized* information states θ_t, with $\Sigma_{x \in X} \theta_t(x) = 1$, rather than unnormalized information states as in [B-J]. Similarly, in the deterministic case we can consider also information states p_t normalized by $(p_t, 0) = \max_{x \in X} p_t(x) = 0$.

(b) *Notation.* For $p \in \mathbb{R}^N,\ c \in \mathbb{R}$, we shall denote by $p + c \in \mathbb{R}^N$ the vector with components $p(x) + c,\ x \in X$.

Let $\Delta = \{\alpha \in \mathbb{R}^N : \max_{x \in X} \alpha(x) = 0\}$, and define the *normalized* information state $\alpha_t \in \Delta$ by

$$\alpha_t(x) = p_t(x) - (p_t, 0).$$

Moreover, let $G^{\mu} : \mathbb{R}^N \times U \times Y \to \mathbb{R}$ be defined by

$$G^{\mu}(p, u, y) = (\text{Ł}^{\mu}(p, u, y), 0) - (p, 0).$$

Then, noting that, for each $p \in \mathbb{R}^N,\ c \in \mathbb{R},\ G^{\mu}(p + c) = G^{\mu}(p)$, we get

$$J(\alpha_0, T; \tilde{u}, \tilde{y}) = \sum_{t=0}^{T-1} G^{\mu}(\alpha_t, u_t, y_{t+1}), \tag{5.5}$$

whenever $p_0 = \alpha_0 \in \Delta$. Further, defining $\overline{\text{Ł}}^{\mu} : \Delta \times U \times Y \to \Delta$ by

$$\overline{\text{Ł}}^{\mu}(\alpha, u, y) = \text{Ł}^{\mu}(\alpha, u, y) - G^{\mu}(\alpha, u, y),$$

we have that

$$\begin{aligned} \alpha_{t+1} &= \text{Ł}^{\mu}(\alpha_t, u_t, y_{t+1}) - G^{\mu}(\alpha_t, u_t, y_{t+1}) \\ &= \overline{\text{Ł}}^{\mu}(\alpha_t, u_t, y_{t+1}). \end{aligned} \tag{5.3$'$}$$

Let $\|\cdot\|$ denotes the supremum norm. The next lemma summarizes several elementary but important properties of G^{μ}, $Ł^{\mu}$, and $\overline{Ł}^{\mu}$.

LEMMA 5.2. *Assume* (H1), (H2). *For each* $p,\ \hat{p} \in \mathbb{R}^N,\ c \in \mathbb{R},\ y \in Y,\ u \in U,\ \alpha \in \Delta$, *the following holds.*

(a)

$$Ł^{\mu}(p + c, u, y) = Ł^{\mu}(p, u, y) + c.$$

(b)

$$\|Ł^{\mu}(p, u, y) - Ł^{\mu}(\hat{p}, u, y)\| \leq \|p - \hat{p}\|,$$

$$-\frac{1}{\mu}\|\varphi_2\| \leq G^{\mu}(p, u, y) \leq \|l\|,$$

$$0 \leq \max_{y \in Y} G^{\mu}(p, u, y) \leq \|l\|.$$

(c) *If* (H1$'$) *and* (H2) *hold, then*

$$(p, 0) - \frac{1}{\mu}[\|\varphi_1\| + \|\varphi_2\|] \leq Ł^{\mu}(p, u, y)(x) \leq \|l\| + (p, 0),$$

$$-[\|l\| + \frac{1}{\mu}(\|\varphi_1\| + \|\varphi_2\|)] \leq \overline{Ł}^{\mu}(\alpha, u, y)(x) \leq 0.$$

Sketch of proof. Part (a) and the first inequality in (b) are immediate from the definition of L^{μ}. For the rest of (b) it suffices to assume that $(p, 0) = 0$. Then

$$G^{\mu}(p, u, y) = \max_{x, x'} \left\{ l(x, u) - \frac{1}{\mu}[\varphi_1(x, u; x') + \varphi_2(x', y)] + p(x) \right\}.$$

One gets $-\mu^{-1}\|\varphi_2\| \leq G^{\mu}(p, u, y)$ by choosing x such that $p(x) = 0$ and $x' = b(x, u, \omega_{\phi})$. If $y = h(x', \omega_{\phi})$, then $0 \leq G^{\mu}(p, u, y)$. Clearly, $G^{\mu} \leq \|l\|$, and the inequalities in (c) are proved similarly. □

Note that, from part (b) of Lemma 5.2, it follows that, for each $T \geq 1$, $p_0, \tilde{p}_0 \in \mathbb{R}^N$, $\tilde{y}$, Γ,

$$\|p_T - \tilde{p}_T\| \leq \|p_0 - \tilde{p}_0\|, \tag{5.6}$$

$$\|J(p_0, T; \Gamma, \tilde{y}) - J(\tilde{p}_0, T; \Gamma, \tilde{y})\| \leq \|p_0 - \tilde{p}_0\|,$$

and therefore

$$|V(p_0, T) - V(\tilde{p}_0, T)| \leq \|p_0 - \tilde{p}_0\|. \tag{5.7}$$

Here $\tilde{p}_t$ is the solution to (5.3) with initial state $\tilde{p}_0$.

Now, in order to study the above dynamic game on an infinite horizon and average cost per unit time criterion, we shall consider first the discounted cost infinite horizon payoff.

Let $0 < \beta < 1$ be the discount factor, and define, for $p_0 \in \mathbb{R}^N, \tilde{y}, \Gamma$,

$$J_{\beta}(p_0; \Gamma, \tilde{y}) = \sum_{t=0}^{\infty} \beta^t[(p_{t+1}, 0) - (p_t, 0)] + (p_0, 0). \tag{5.8}$$

Then, using "summation by parts," for each $T > 0$,

$$\begin{aligned} J_\beta(p_0, T; \Gamma, \tilde{y}) &:= \sum_{t=0}^{T-1} \beta^t [(p_{t+1}, 0) - (p_t, 0)] + (p_0, 0) \\ &= \beta^T (p_T, 0) - \sum_{t=0}^{T-1} (\beta^{t+1} - \beta^t)(p_{t+1}, 0) \\ &= \beta^T (p_T, 0) + (1-\beta) \sum_{t=0}^{T-1} \beta^t (p_{t+1}, 0). \end{aligned}$$

Further, by (5.6),

$$\begin{aligned} |J_\beta(p_0, T; \Gamma, \tilde{y}) - J_\beta(\tilde{p}_0, T; \Gamma, \tilde{y})| &\le \beta^T \|p_0 - \tilde{p}_0\| + (1-\beta) \sum_{t=0}^{T-1} \beta^t \|p_0 - \tilde{p}_0\| \\ &\le (\beta^T + 1) \|p_0 - \tilde{p}_0\|. \end{aligned}$$

Therefore, letting $T \to \infty$, we get

$$|J_\beta(p_0; \Gamma, \tilde{y}) - J_\beta(\tilde{p}_0; \Gamma, \tilde{y})| \le \|p_0 - \tilde{p}_0\|. \tag{5.9}$$

Also, if $p_0 = \alpha_0 \in \Delta$, then, by (5.8) and the fact $(p_{t+1}, 0) - (p_t, 0) = G^\mu(p_t, u_t, y_{t+1})$,

$$(1-\beta)|J_\beta(\alpha_0)| \le \|G^\mu\|. \tag{5.10}$$

Let $W_\beta(\alpha)$ be the (upper) value of the game with discounted infinite horizon payoff. Then, (5.9)–(5.10) imply

$$|W_\beta(\alpha) - W_\beta(\tilde{\alpha})| \le \|\alpha - \tilde{\alpha}\| \quad \text{for all } \alpha, \alpha_0 \in \Delta, \tag{5.11}$$

and

$$0 \le (1-\beta) W_\beta(\cdot) \le \|G^\mu\|, \tag{5.12}$$

where the left-hand side inequality follows from Lemma 5.2(b).

The corresponding Isaacs equation for this game is

$$W_\beta(\alpha) = \min_{u \in U} \max_{y \in Y} [\beta W_\beta(\overline{\mathrm{L}}^\mu(\alpha, u, y)) + G^\mu(\alpha, u, y)], \tag{5.13}$$

and standard dynamic programming arguments show that the upper-value function W_β is the unique continuous solution to this equation.

If (H1$'$) and (H2) hold, then from Lemma 5.2(c), for arbitrary $\alpha_0 \in \Delta$, $\alpha_t \in K$ for all $t \ge 1$, where K is a fixed compact set. Hence, it suffices to consider $\alpha_0 \in K$. In fact, a similar property holds under the weaker assumptions (H1)–(H2).

Lemma 5.3. *Assume* (H1)–(H2). *Then there exists a compact set K such that, for arbitrary $\alpha_0 \in \Delta$, $\alpha_t \in K$ for all $t \ge T_1$.*

Proof. We write $x \to x'$ under u if $x' = b(x, u, \omega)$ for some ω. Let

$$C = \max_{x, x', u} \{\varphi_1(x, u; x') \mid x \to x' \text{ under } u\}.$$

Given x, x', and $\tilde{u}$, let $\tilde{\omega}$ be as in (H1). Then $x_t \to x_{t+1}$ under u_t for $t = 0, 1, \dots, T-1$. From (5.3)

$$p_t(x_t) - \frac{1}{\mu}(C + \|\varphi_2\|) \leq p_{t+1}(x_{t+1}) \leq \|l\| + (p_t, 0),$$

from which, since $x_0 = x,\ x_{T_1} = x'$,

$$p_0(x) - \frac{T_1}{\mu}(C + \|\varphi_2\|) \leq p_{T_1}(x') \leq T_1\|l\| + (p_0, 0).$$

Take $p_0 = \alpha_0$ and x such that $\alpha_0(x) = 0$. Then $\|p_{T_1}(x')\| \leq C_1$ which implies $\alpha_{T_1} \in K_1$ for some fixed compact set K_1. Then also $\alpha_{mT_1} \in K_1$ for $m = 1, 2, \dots,$ and $\alpha_t \in K$ for all $t \geq T_1$ for some compact K. □

Let $\alpha_\beta \in \operatorname{argmin}\{\ W_\beta(\alpha) \mid \alpha \in K\ \}$. Thus, (5.11)–(5.12) imply that, for a sequence $\beta_n \nearrow 1$,

$$\begin{gathered}(1-\beta_n)W_{\beta_n}(\alpha_{\beta_n}) \to \lambda^\mu,\\ W_{\beta_n}(\alpha) - W_{\beta_n}(\alpha_{\beta_n}) \to W^\mu(\alpha),\end{gathered}$$

uniformly on K, with $\min_K W(\alpha) = 0$.

Further, rewriting (5.13), with $\beta = \beta_n$, as

$$\begin{aligned}W_{\beta_n}(\alpha) - W_{\beta_n}(\alpha_{\beta_n}) = \min_u \max_y [&\beta_n(W_{\beta_n}(\bar{\mathbb{L}}^\mu(\alpha, u, y)) - W_{\beta_n}(\alpha_{\beta_n}))\\ &+ G^\mu(\alpha, u, y)] - (1-\beta_n)W_{\beta_n}(\alpha_{\beta_n}),\end{aligned}$$

and letting $n \to \infty$, we get

$$\lambda^\mu + W^\mu(\alpha) = \min_u \max_y [W^\mu(\bar{\mathbb{L}}^\mu(\alpha, u, y)) + G^\mu(\alpha, u, y)]. \tag{5.14}$$

We have proved the following.

THEOREM 5.4. *There exist a number* λ^μ *and a nonnegative continuous function* $W^\mu : K \to \mathbb{R}$ *such that* (5.14) *holds.*

On the other hand, a standard dynamic programming argument implies that, for $0 < T < \infty$,

$$W^\mu(\alpha) = \mathcal{UV}\left[\sum_{t=0}^{T-1}[G^\mu(\alpha_t, u_t, y_{t+1}) - \lambda^\mu] + W^\mu(\alpha_T)\right], \tag{5.15}$$

where $\mathcal{UV}$ denotes the upper value of the finite time game with running cost $G^\mu - \lambda^\mu$ and terminal cost $W^\mu(\alpha_T)$. Further, we also have by (5.4) that

$$V(\alpha, T) = \mathcal{UV}\left[\sum_{t=0}^{T-1} G^\mu(\alpha_t, u_t, y_{t+1})\right].$$

Since α_t belongs to the compact set K, we get, by dividing by T and letting $T \to \infty$, that

$$\lambda^\mu = \lim_{T\to\infty} \frac{V(\alpha, T)}{T}. \tag{5.16}$$

This shows, in particular, that λ^μ does not depend on the particular sequence $\{\beta_n\}$.

The next theorem follows in a way similar to [F-McE, sec. 8], and we omit its proof.

THEOREM 5.5. *The upper-value function of the dynamic game with payoff functional*

$$P(\alpha, \Gamma, \tilde{y}) = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} G^{\mu}(\alpha_t, u_t, y_{t+1})$$

is equal to λ^{μ}.

Optimal stationary control policy. Let $u^* : K \to U$ such that

$$u^*(\alpha) \in \operatorname*{argmin}_{u}\{\max_{y}[W^{\mu}(\overline{Ł}^{\mu}(\alpha, u, y)) + G^{\mu}(\alpha, u, y)]\}. \tag{5.17}$$

Given $\alpha_0 \in K$, the policy $u^*(\cdot)$ determines a strategy Γ^* as follows. Given $\tilde{y} = (y_1, y_2, \ldots)$, define α_t^* by

$$\begin{aligned} \alpha_{t+1}^* &= \overline{Ł}^{\mu}(\alpha_t^*, u^*(\alpha_t^*), y_{t+1}), \\ \alpha_0^* &= \alpha_0. \end{aligned}$$

Now take

$$\Gamma^*(\tilde{y})_t = u_t^* = u^*(\alpha_t^*).$$

Then from (5.14), for all y, α,

$$\lambda^{\mu} + W^{\mu}(\alpha) \geq W^{\mu}(\overline{Ł}(\alpha, u^*(\alpha), y)) + G^{\mu}(\alpha, u^*(\alpha), y).$$

By iterating this inequality, we get, for all $\tilde{y}$, $T > 0$,

$$\lambda^{\mu} T + W^{\mu}(\alpha_0) \geq \sum_{t=0}^{T-1} G^{\mu}(\alpha_t^*, u_t^*, y_{t+1}) + W^{\mu}(\alpha_t^*).$$

Further, by (5.5), this can be rewritten as

$$\lambda^{\mu} T + W^{\mu}(\alpha_0) \geq (p_T^*, 0) + W^{\mu}(\alpha_T^*),$$

where

$$\begin{aligned} p_{t+1}^* &= Ł^{\mu}(p_t^*, u_t^*, y_{t+1}), \\ p_0^* &= \alpha_0. \end{aligned}$$

Hence, by (5.4), for all $x_0 \in X$, $\tilde{\omega} \in Z^{\infty}$, $T > 0$,

$$\begin{aligned} \sum_{t=0}^{T-1} l(x_t^*, u_t^*) + W^{\mu}(\alpha_T^*) \leq -\alpha_0(x_0) + \frac{1}{\mu} \sum_{t=0}^{T-1} [\vartheta_1(\omega_t) + \vartheta_2(\omega_{t+1})] \\ + \lambda^{\mu} T + W^{\mu}(\alpha_0), \end{aligned} \tag{5.18}$$

where

$$\begin{aligned} x_{t+1}^* &= b(x_t^*, u_t^*, \omega_t), \\ x_0^* &= x_0. \end{aligned}$$

Now, let us assume that $\lambda^{\mu} = 0$. In this case (5.18) becomes a kind of "dissipation inequality" (cf. (5.16) in [F-McE]). Suppose, moreover, that $W^{\mu}(\alpha_0) = \min_K W^{\mu}(\alpha)$. Then, inequality (5.18) implies that, for all $x_0,\ \tilde{\omega},\ T > 0$,

$$\sum_{t=0}^{T-1} l(x_t^*, u_t^*) \leq -\alpha_0(x_0) + \frac{1}{\mu}\sum_{t=0}^{T-1}[\vartheta_1(\omega_t) + \vartheta_2(\omega_{t+1})]. \tag{5.19}$$

Thus, the output feedback robust control is achieved with μ^{-1}, according to [B-J, Definition 2.12], provided $\beta = -\alpha_0$ (in their notation for $\beta(\cdot)$). If we assume that $l(x_\phi, u_\phi) = 0$ and $l(x, u) > 0$ for all $x \neq x_\phi$, then $\alpha_0(x_\phi) = 0$. To see this recall that $-\alpha_0(x_0) \geq 0$ with equality for some $\overline{x}_0$. Take $x_0 = \overline{x}_0$ and $\omega_t = 0$ for all t. By (5.19), $l(x_0^*, u_0^*) = l(\overline{x}_0, u_0^*) = 0$ and hence $\overline{x}_0 = x_\phi$.

On the other hand, since $V(p, T) \leq V(p, T+1)$,

$$\overline{W}(p) := \lim_{T\to\infty} V(p, T)$$

exists (maybe $+\infty$). In the case of $\overline{W}(p) < \infty$,

$$\overline{W}(p) = \min_u \max_y \overline{W}(\text{Ł}^{\mu}(p, u, y)).$$

This implies, particularly, since $\overline{W}(p + c) = \overline{W}(p) + c$ for all $p \in \mathbb{R}^N,\ c \in \mathbb{R}$, that

$$\begin{aligned}\overline{W}(\alpha_0) &= \mathcal{U}\mathcal{V}[\overline{W}(p_T)]\\ &= \mathcal{U}\mathcal{V}[\overline{W}(\alpha_T) + (p_T, 0)]\end{aligned}$$

whenever $p_0 = \alpha_0 \in \Delta$.

Replacing $\overline{W}(\alpha)$ by $W^{\mu}(\alpha)$, we can argue as before, with $\lambda^{\mu} = 0$ and u^* as in (5.17).

In order to ensure that $\lambda^{\mu} = 0$ for all sufficiently small μ, let us make (as in [B-J] and [F-HH]) additional assumptions about the existence of a "null" state and control $x_\phi,\ u_\phi$.

(H3)

$$\left\{\begin{array}{rcll} f(x_\phi, u_\phi) & = & x_\phi, & \\ l(x_\phi, u_\phi) & = & 0, & \\ l(x, u) & > & 0 & \text{for } (x, u) \neq (x_\phi, u_\phi). \end{array}\right.$$

(H4) There exists an integer T_0 such that for any initial condition $x_0 = x$, the solution to $x_{t+1} = f(x_t, u_\phi)$ reaches x_ϕ in T_0 steps.

If any constant control u is chosen, with $u_t = u$ for $t = 0, 1, \ldots$, then [F-HH, sect. 2] and section 3 above give a corresponding λ_u^{μ}. By either (5.16) or Theorem 5.5, $\lambda_u^{\mu} \geq \lambda^{\mu}$. In particular, take $u = u_\phi$. Then (H3) and (H4) imply that $\lambda_{u_\phi}^{\mu} = 0$ for $0 < \mu \leq \mu_\phi$, for suitable μ_ϕ. Hence $\lambda^{\mu} = 0$ for $0 < \mu \leq \mu_\phi$. Let

$$\mu^* = \sup\{\mu\ :\ \lambda^{\mu} = 0\}.$$

Then the crucial dissipation inequality (5.19) of robust control with partial state information can be achieved using the optimal control policy $u^*(\alpha)$, provided $\mu \leq \mu^*$.

Now let us assume that $\mu > \mu^*$, and hence $\lambda^{\mu} > 0$. In this case, (5.18) implies, since $W(\alpha_T^*)$ is bounded, that

$$\limsup_{T\to\infty}\left\{\frac{1}{T}\sum_{t-1}^{T-1} l(x_t^*, u_t^*) - \frac{1}{T}\cdot\frac{1}{\mu}\sum_{t=0}^{T-1}[\vartheta_1(\omega_t) + \vartheta_2(\omega_{t+1})]\right\} \leq q\lambda^{\mu}.$$

In terms of Definition 2.1 in [D-J], the above inequality means that the system Σ^{u^*} has finite power gain less than or equal to μ^{-1}.

Small noise limit. For each $u \in U$, analogously to section 3, define the matrices $P^\varepsilon(u)$ and Q^ε.

Then, replacing the matrices $P(u), Q$ in section 4 by $P^\varepsilon(u)$, Q^ε, Theorem 4.2 implies (in view of (H1-H2)) the existence of a number $\Lambda^{\mu,\varepsilon}$ and a nonnegative continuous function $W^{\mu,\varepsilon} : S \to \mathbb{R}$ such that

$$\exp\left\{\frac{\mu}{\varepsilon}\cdot\Lambda^{\mu,\varepsilon} + W^{\mu,\varepsilon}(\theta)\right\} = \min_{u\in U} E^{+}\exp\left\{G^{\mu,\varepsilon}(\theta,u,y) + W^{\mu,\varepsilon}(F^{\mu,\varepsilon}(\theta,u,y))\right\}, \tag{5.20}$$

where now $F^{\mu,\varepsilon}(\theta,u,y) = \frac{A^\varepsilon(u,y)\theta}{|A^\varepsilon(u,y)\theta|}$ and $G^{\mu,\varepsilon}(\theta,u,y) = \log|A^\varepsilon(u,y)\theta|$.

Now we proceed as in section 3 and define the functions $\overline{W}^{\mu,\varepsilon} : \mathbb{R}^N \to \mathbb{R}$, $\overline{G}^{\mu,\varepsilon} : \mathbb{R}^N \times U \times Y \to \mathbb{R}$, and $\text{Ł}^{\mu,\varepsilon} : \mathbb{R}^N \times U \times Y \to \mathbb{R}^N$ by

$$\left\{\begin{array}{l} \overline{W}^{\mu,\varepsilon}(p) = \frac{\varepsilon}{\mu}\overline{W}^{\mu,\varepsilon}(\zeta_p), \\ \overline{G}^{\mu,\varepsilon}(p,u,y) = \frac{\varepsilon}{\mu}\overline{G}^{\mu,\varepsilon}(\zeta_p,u,y), \\ \text{Ł}^{\mu,\varepsilon}(p,u,y) = \frac{\varepsilon}{\mu}\log A^\varepsilon(u,y)e^{\frac{\mu}{\varepsilon}p}, \end{array}\right.$$

Note that, from the above transformations, we can write (5.19) as

$$\begin{aligned} \Lambda^{\mu,\varepsilon} + \overline{W}^{\mu,\varepsilon}(p) \quad = \quad \min_{u\in U}\frac{\varepsilon}{\mu}\log\sum_{y\in Y}\exp\{&\overline{W}^{\mu,\varepsilon}(\text{Ł}^{\mu,\varepsilon}(p,u,y)) \\ &+ \overline{G}^{\mu,\varepsilon}(p,u,y)\}\cdot\frac{1}{M}. \end{aligned}$$

Then, using the same kind of arguments as in the proof of the Theorem 3.2, it follows that

(a) $\varepsilon \to \Lambda^{\mu,\varepsilon}$ is uniformly bounded,

(b) $\text{Ł}^{\mu,\varepsilon}(p,u,y) \to \text{Ł}^{\mu}(p,u,y)$ as $\varepsilon \to 0$ uniformly on compact sets of $\mathbb{R}^N \times U \times Y$,

(c) $\overline{G}^{\mu,\varepsilon}(p,u,y) \to G^{\mu}(p,u,y)$ as $\varepsilon \to 0$ uniformly on compact sets of $\mathbb{R}^N \times U \times Y$.

In fact, if there exists a sequence $\{\varepsilon_n\}$, with $\varepsilon_n \to 0$ as $n \to \infty$, such that $\Lambda^{\mu,\varepsilon_n}$ tends to some Λ^μ and $\overline{W}^{\mu,\varepsilon_n}$ converges uniformly on compact sets t o some function $W^\mu : \mathbb{R}^N \to \mathbb{R}$, the existence of a solution to (5.14) follows straightforwardly in view of (a)–(c) above. In order to obtain such a sequence ε_n, it would suffice to prove equicontinuity of the functions $\overline{W}^{\mu,\varepsilon}$, which we have not succeeded in doing. For the simpler situation in section 3 this difficulty was avoided by using the explicit form (3.6) of $W^{\mu,\varepsilon}$ and a corresponding convergence result [F-HH, Thm. 3.2] for the state feedback case.

Acknowledgments. We wish to thank Professor Steven Marcus for bringing to our attention several reprints of his work. D. Hernández-Hernández also expresses his gratitude to the Division of Applied Mathematics, Brown University, and the Institute for Systems Research, University of Maryland, for their hospitality and support.

REFERENCES

[A] K. J. Astrom, *Optimal control of Markov processes with incomplete state information, II. The convexity of the loss function*, J. Math. Anal. Appl., 10 (1965), pp. 403–406.

[A-M] A. Arapostathis and S. I. Marcus, *Analysis of an identification algorithm arising in the adaptive estimation of Markov chains*, Math. Control Signals Systems, 3 (1990), pp. 1–29.

[B-J] J. S. Baras and M. R. James, *Robust and risk-sensitive output feedback control for finite state machines and hidden Markov models* (summary), J. Math. Systems Estim. Control, 7 (1997), pp. 371–374.

[D-J] P. M. Dower and M. R. James, *Dissipativity and nonlinear systems with finite power gain*, Internat. J. Robust Nonlinear Control, submitted.

[D-E] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, John Wiley, New York, 1997.

[FG-M] E. Fernández-Gaucherand and S. I. Marcus, *Risk sensitive optimal control of hidden Markov models: Structural results*, IEEE Trans. Automat. Control, to appear.

[F-HH] W. H. Fleming and D. Hernández-Hernández, *Risk-sensitive control of finite state machines on an infinite horizon* I, SIAM J. Control Optim., 35 (1997), pp. 1790–1810.

[F-McE] W. H. Fleming and W. M. McEneaney, *Risk sensitive control on an infinite time horizon*, SIAM J. Control Optim., 33 (1995), pp. 1881–1915.

[F-McE1] W. H. Fleming and W. M. McEneaney, *Risk sensitive and differential games*, in Stochastic Theory and Adptive Control, Lecture Notes in Control and Inform. Sci. 184, Springer-Verlag, Berlin, 1992, pp. 185–197.

[F-S] W. H. Fleming and P. E. Souganidis, *On the existence of value functions of two-player zero-sum stochastic differential games*, Indiana Univ. Math. J., 38 (1989), pp. 239–314.

[F-W] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems*, Springer-Verlag, New York, 1984.

[F-K] H. Furstenberg and H. Kesten, *Products of random matrices*, Ann. Math. Statist., 31 (1960), pp. 457–469.

[HH-M] D. Hernández-Hernández and S. I. Marcus, *Risk sensitive control of Markov processes in countable state space*, Systems Control Lett., 29 (1996), pp. 147–155; *Corrigendum*, Systems Control Lett., 34 (1998), pp. 105–106.

[J] D. H. Jacobson, *Optimal stochastic linear systems with exponential criteria and their relation to deterministic differential games*, IEEE Trans. Automat. Control, 18 (1973), pp. 124–131.

[J-B] M. R. James and J. S. Baras, *Robust H_∞ output feedback control for nonlinear systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 1007–1017.

[J-B-E] M. R. James, J. S. Baras, and R. Elliott, *Risk sensitive control and dynamic games for partially observed discrete time nonlinear systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 750–792.

[K] T. Kaijser, *A limit theorem for partially observed Markov chains*, Ann. Probab., 3 (1975), pp. 677–696.

[R-S] W. J. Runggaldier and L. Stettner, *Approximations of Discrete Time Partially Observed Control Problems*, Applied Mathematics Monographs, Gardini Editorie Stampatori, Pisa, Italy.

[W] P. Whittle, *Risk Sensitive Optimal Control*, John Wiley, New York, 1990.

SIAM J. Control Optim.
Vol. 37, No. 4, pp. 1070–1088

NEWTON'S MESH INDEPENDENCE PRINCIPLE FOR A CLASS OF OPTIMAL SHAPE DESIGN PROBLEMS*

MANFRED LAUMEN†

Abstract. Many optimal shape design problems can be stated as infinite dimensional minimization problems. For deriving an implementable algorithm it has to be decided either to discretize the problem and to use a finite algorithm to solve the discrete problem or to state an algorithm in function space and to discretize this algorithm. This issue has yet to be addressed in the field of optimal shape design research. One big advantage for the latter procedure is a mesh independence behavior as it has been proven by Allgower et al. [*SIAM J. Numer. Anal.*, 23 (1986), pp. 160–169]. Since their assertions are not directly applicable to these specific kinds of problems, a modified version of their mesh independence principle is given here in order to derive more efficient algorithms for the resulting large scale problems.

Key words. Newton's method, optimal shape design, mesh independence principle

AMS subject classifications. 65K10, 49K20, 49M15, 49M05

PII. S0363012996303529

1. Introduction.

In 1986, Allgower et al. [3] published a general mesh independence proof for Newton's method in the context of nonlinear equations. The applications of their theoretical results to various mathematical problems are presented in the original paper and in the follow-up publication [2]. Heinkenschloss stated a similar mesh independence result for the Gauss–Newton method, which was applied to a parameter identification problem [21], [22], [23]. For quasi-Newton methods with Broyden update, Kelley and Sachs analyzed the influence of discretizations in [27].

Besides the advantage of the mesh independence principle for predicting the convergence of the computable method on the basis of the analyzed infinite dimensional convergence, there is a further important point for practical implementations. The mesh independence lays the theoretical foundation for the justification of refinement strategies and helps to design this refinement process (see, e.g., [2]). Since the focus is the infinite dimensional solution, a fine discretization scheme has to be chosen, so that the discrete solution approximates the infinite dimensional solution appropriately. However, a fine discretization also means that the finite problem consists of many variables, and therefore, an increased amount of work per iteration has to be expected. Keeping the numerical method fixed, the only possibility of reducing the total workload to obtain the discrete solution is to improve the starting point of the nonlinear iteration. Apparently, this can be done by using information from the coarse grid discretization leading to the concept of mesh refinement (respectively, nested iteration). Several numerical computations confirm the efficiency of this strategy for various algorithms (see, e.g., [17], [18]). The intention of this paper is to lay the foundation for using mesh refinement to solve the resulting large scale problems efficiently. This is done by proving an appropriately modified mesh independence principle.

Shape optimization is described by finding the geometry of a structure which is optimal in the sense of a given minimized cost function with respect to certain

*Received by the editors May 15, 1996; accepted for publication (in revised form) April 20, 1998; published electronically April 27, 1999.

http://www.siam.org/journals/sicon/37-4/30352.html

†Schueleinplatz 9, 81673 Muenchen, Germany (Manfred.Laumen@feilmeier.de).

constraints. Many-faceted problems naturally arise in engineering applications with the goal of designing a specific structure in an optimal sense, or alternatively, of understanding and determining the shape of a given structure. Typical applications are the design of a nozzle [35], a thermal diffuser [14], an airfoil boundary [36], or various beams and plates [20], [38], with respect to specific optimality conditions.

Several methods, e.g., the speed method [38], the boundary element method [35], the fictitious domain method [19], and the mapping method [8], have been developed for such problems in the past. In particular, in recent years the homogenization method has been the focus of research for solving such kinds of optimal shape design problems (see, e.g., [1], [9], [11], [25], [33]). The computation of the solution is a time consuming task for all of them because it normally leads to a large scale optimization problem, which requires the subsequent solution of many boundary value problems. We restrict our presentation to the mapping method which has the advantage that the theory also covers a class of optimal control problems, where the coefficients of the variational equation are influenced by the control. However, the mesh independence assertions presented here can also be applied to the other kinds of methods.

The considered class of optimal shape design problems that are extensively investigated in [30], [31] can be written as

$$\min_{u \in \mathcal{U}_{ad}} \tilde{J}(u, \tilde{y}, \tilde{z}). \tag{1.1}$$

The design, respectively, control, function $u \in \mathcal{U}_{ad}$ parametrizes the bounded domain

$$\tilde{\Omega} = \{\tilde{x} = (\tilde{x}_1, \tilde{x}_2)^T \in \mathbb{R}^2 \,|\, \tilde{x}_2 \in I := (0,1) \wedge \tilde{x}_1 \in (0, u(\tilde{x}_2))\} \tag{1.2}$$

and $\mathcal{U}_{ad}$ is a suitable subset of a function space $\mathcal{U}$. The state $\tilde{y} \in \tilde{\mathcal{V}}$, $\tilde{\mathcal{V}}$ Hilbert space, is the solution of an elliptic boundary value problem of the second kind on the domain $\tilde{\Omega}$, and the function $\tilde{z} \in \tilde{\mathcal{Z}}$, $\tilde{\mathcal{Z}}$ Hilbert space, has to be introduced to describe a desired state of the function $\tilde{y}$ or for handling inhomogeneous Dirichlet boundary conditions.

Since the boundary value problem is solved with the finite element method, the weak formulation in terms of a variational equation on the moving domain $\tilde{\Omega}$ is used from the beginning. Based on the boundary partition $\partial\tilde{\Omega} = \tilde{\Gamma}_0 \cup \tilde{\Gamma}_1$ with $\tilde{\Gamma}_0 \cap \tilde{\Gamma}_1 = \emptyset$, the space of the state is defined by

$$\tilde{\mathcal{V}} = \{\phi \in \mathcal{H}^1(\tilde{\Omega}) |\, \gamma_0 \phi|_{\tilde{\Gamma}_0} = 0\},$$

where the trace map $\gamma_0 \in \mathcal{L}(\mathcal{H}^1(\tilde{\Omega}), \mathcal{H}^{\frac{1}{2}}(\partial\tilde{\Omega}))$ of order zero is further omitted if the meaning is obvious.

The mapping method transforms the moving domain $\tilde{\Omega}$ to a fixed domain Ω leading to an optimal control problem that is defined on fixed Ω. Therefore, we assume that a suitable transformation $T = T(u)$ completely determined by a function $u \in \mathcal{U}_{ad}$ exists, which is the case for several kinds of transformations, as presented, for example, by Banks and Kojima [5]. Thus, by using the generalized substitution rule the optimal shape design problem (1.1) is modified to the optimal control problem

$$\min_{u \in \mathcal{U}_{ad}} J(u, y, z) \tag{1.3}$$

with a general variational equation

$$a(u; y, \eta) = l(u; \eta) \qquad \forall \eta \in \mathcal{V}, \tag{1.4}$$

where the $\mathcal{V}$-elliptic and continuous bilinear form $a(u;\cdot,\cdot):\mathcal{V}\times\mathcal{V}\to\mathbb{R}$ is given by

$$a(u;y,\eta)=\sum_{|i|,|j|\leq 1}\langle a_{ij}(u)D^i y, D^j\eta\rangle_{\mathcal{L}^2(\Omega)}+\langle b(u)y,\eta\rangle_{\mathcal{L}^2(\Gamma_1)},$$

and the linear functional $l(u;\cdot):\mathcal{V}\to\mathbb{R}$ is defined by

$$l(u;\eta)=\sum_{|i|\leq 1}\langle f_i(u),\eta\rangle_{\mathcal{L}^2(\Omega)}+\langle f(u),\eta\rangle_{\mathcal{L}^2(\Gamma_1)}.$$

This variational equation is described on the fixed domain Ω, where the coefficient functions now depend nonlinearly on the parameter function $u\in\mathcal{U}_{ad}$. The tilde indicating a function, boundary, etc., on the moving domain $\tilde{\Omega}$ is omitted, if the symbols are defined analogously on the fixed region Ω.

The derived minimization problem is further simplified by supposing initially the nonactivity of the constraint $u\in\mathcal{U}_{ad}$; i.e., the problem (1.3), respectively, (1.4), can be treated with the modified cost function

$$\min_{u\in\mathcal{U}} J(u,y,z). \tag{1.5}$$

This simplification is justified if the solution u_* is supposed to be an interior point of $\mathcal{U}_{ad}$, and if the starting point is near the solution, which is guaranteed, for instance, if a nested iteration is used.

By using the Lax–Milgram lemma, the equality constraint (1.4) can be eliminated by a solution operator $y=S(u)$. Hence, the constrained minimization problem (1.5) is modified to the unconstrained minimization problem

$$\min_{u\in\mathcal{U}} J(u,S(u),z(u)). \tag{1.6}$$

To handle the possible ill-posedness of the problems, a Tikhonov regularization term (see, e.g., [6], [7]) is sometimes added to this cost function if necessary. Thus, (1.6) is finally written as

$$\min_{u\in\mathcal{U}} F(u) \tag{1.7}$$

with $F(u):=J(u,S(u),z(u))+\frac{\varepsilon}{2}\|u-u_{\mathcal{T}}\|^2_{\mathcal{T}},\ \ \varepsilon\in\mathbb{R}$, and a function $u_{\mathcal{T}}$.

There are two different procedures for stating numerical methods based on the differentiability properties. One way is to derive the derivatives in function spaces and to discretize the algorithm afterward as will be presented here. The other procedure that is commonly used in the context of optimal shape design problems is to discretize the problem first and then to state numerical methods for the discrete problem.

It is a well-known fact that both procedures could lead to different algorithms. This fact is discussed in detail, for instance, by Chenais [12] in the context of optimal shape design problems and by Kelley and Sachs [26], [28] for solving continuous problems with quasi-Newton methods.

In contrast to the commonly used procedure for solving optimal shape design problems, the discretization of the infinite dimensional algorithm offers the possibility of improving approximation results by comparing the computable iterates with the underlying infinite dimensional ones. A typical question for the research in this context is the provability of a mesh independence principle. In other words, do all quantities of the method, such as iterates, solutions, and convergence constants, depend continuously on the discretization?

This illustrates impressively the demand of a mesh independence principle for this specific class of problems. The research of Allgower et al. concerning Newton's method is stated in the context of nonlinear equations. Their work can be transferred to our minimization problem by the necessary first order condition $F'(u) = 0$ in $\mathcal{U}'$. Then, Newton's method is defined by ($n \in \mathbb{N}$)

$$\begin{aligned} F''(u_{n-1})(w)(v) &= -F'(u_{n-1})(v) \qquad \forall v \in \mathcal{U}, \\ u_n &= u_{n-1} + w, \end{aligned} \tag{1.8}$$

where F, F', and F'' also depend on functions defined on the infinite dimensional Hilbert space $\mathcal{V}$. This method is discretized by the replacement of the infinite dimensional spaces $\mathcal{V}$ and $\mathcal{U}$ with the finite dimensional subspaces $\mathcal{V}^N$ and $\mathcal{U}^M$, leading to the discretized algorithm

$$\begin{aligned} F''_N(u^M_{n-1})(w^M)(v^M) &= -F'_N(u^M_{n-1})(v^M) \qquad \forall v^M \in \mathcal{U}^M, \\ u^M_n &= u^M_{n-1} + w^M. \end{aligned} \tag{1.9}$$

At first glance, one difficulty for applying their results directly to the optimal shape design problems under consideration arises from the fact that the uniform boundedness of a projection operator Π is required. This operator should map the infinite dimensional space $\mathcal{U}$ to the finite element space $\mathcal{U}^M$. Unfortunately, for our problems $\mathcal{U}$ is chosen to be only a Banach and not a Hilbert space, which means that the existence of such a projection cannot be assured. The lack of a Hilbert space can be overcome in two ways.

First, we could try to extend the formulation of the problem from the Banach space $\mathcal{U}$ to any Hilbert space, for example, $\mathcal{H}^1(I)$. However, the Fréchet-differentiability has to be proven in this larger space, which is far more complicated. Also, the extension strategy of Tröltzsch [39], who stated optimality conditions in a Hilbert space assuming the differentiability only with respect to a smaller Banach space, cannot be applied since it is not guaranteed that the extension of the derivative with respect to $\mathcal{H}^1(I)$ even exists.

Second, the Banach space $\mathcal{U}$ could be restricted to a smaller Hilbert space, for example, $\mathcal{H}^2(I)$. However, the mesh independence theorem has to be stated under rather strong regularity conditions limiting the application to real problems.

On closer inspection of the Allgower et al. mesh independence theorem, one can see that their assumptions on the projection operator $\Pi : \mathcal{U} \to \mathcal{U}^M$ are required only since they do not assume $\mathcal{U}^M \subset \mathcal{U}$. Since this inclusion is naturally satisfied for the finite element method that is commonly used for solving optimal shape design problems, their theorem could be modified without using such a projection operator.

Another problem for applying the results of Allgower et al. is caused by their consistency condition with respect to the derivatives of the cost function. For the considered problems it is not reasonable to expect such a required order of consistency.

These are the reasons for presenting a new modified mesh independence principle in the next section, which could be applied to these optimal shape design problems. The main assertion of the paper is given in Theorem 2.2. Since the derivation of the underlying complicated theory for the construction of Newton's method to solve the optimal shape design problems would be beyond the scope of this paper, we give only evidence for the necessity of the modified mesh independence theorem in section 3 and refer to the paper [31] for the theory. Finally, the discretization of the algorithm is given by the finite element method, and numerical results confirm the practical importance of the derived assertions.

2. Mesh independence principle. The modification of Newton's mesh independence theorem is adjusted to the notation of minimization problems and could be applied to the optimal shape design problems under consideration. For the nth iteration of Newton's method it leads to the estimate of the error $\|u_n - u_n^M\|_{\mathcal{U}}$ instead of the weaker assertion of the standard theorem based on the error $\|\Pi u_n - u_n^M\|_{\mathcal{U}}$. Throughout the proofs advantage is taken of the fact that the approximated function F_N is defined for all $u \in \mathcal{U}$, although it will be evaluated only for some points $u^M \in \mathcal{U}^M$.

Corresponding to the infinite dimensional convergence assertion, u_0 can be chosen to be in the ball $U_* := U(u_*, \rho_*) := \{u \in \mathcal{U} : \|u - u_*\|_{\mathcal{U}} \le \rho_*\}$ in order to guarantee the convergence to the solution u_*. Some further assumptions concerning the cost function F_N, which are assumed to hold on a possibly smaller ball $\hat{U}_* := \hat{U}(u_*, \hat{\rho}_*) := \{u \in \mathcal{U} : \|u - u_*\|_{\mathcal{U}} \le \hat{\rho}_*\}$ with $\hat{\rho}_* \le \rho_*$, will be stated in the following.

Assumption C1. Let the assumptions of the theorem on the q-quadratic convergence of Newton's method for the infinite dimensional problem be satisfied. In particular, with appropriate constants L and δ

$$\begin{aligned} \|F''(u) - F''(v)\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} &\le L\|u - v\|_{\mathcal{U}} \qquad \forall u, v \in U_*, \\ \|F''(u_*)^{-1}\|_{\mathcal{L}(\mathcal{L}(\mathcal{U},\mathbb{R}),\mathcal{U})} &\le \delta. \end{aligned}$$

Assumption C2. There exist uniformly bounded Lipschitz constants $L_N^{(i)}$, $i = 1, 2$, such that

$$\begin{aligned} \|F_N'(u) - F_N'(v)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})} &\le L_N^{(1)}\|u - v\|_{\mathcal{U}} \ \ \forall u, v \in \hat{U}_*, \forall N \in \mathbb{N}, \\ \|F_N''(u) - F_N''(v)\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} &\le L_N^{(2)}\|u - v\|_{\mathcal{U}} \ \ \forall u, v \in \hat{U}_*, \forall N \in \mathbb{N}. \end{aligned}$$

Without loss of generality, we assume $L_N^{(i)} \le L$, $i = 1, 2$, for all N.

Assumption C3. There exists a sequence $\zeta_N^{(1)}$ with $\zeta_N^{(1)} \overset{N\to\infty}{\longrightarrow} 0$, such that

$$\begin{aligned} \|F_N'(u) - F'(u)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})} &\le \zeta_N^{(1)} \qquad \forall u \in \hat{U}_*, \forall N \in \mathbb{N}, \\ \|F_N''(u) - F''(u)\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} &\le \zeta_N^{(1)} \qquad \forall u \in \hat{U}_*, \forall N \in \mathbb{N}. \end{aligned}$$

Assumption C4. There exists a sequence $\zeta_M^{(2)}$ with $\zeta_M^{(2)} \overset{M\to\infty}{\longrightarrow} 0$, such that for all $M \in \mathbb{N}$ there exists a $\hat{u}^M \in \mathcal{U}^M \cap \hat{U}_*$, such that $\|\hat{u}^M - u_*\|_{\mathcal{U}} \le \zeta_M^{(2)}$.

Assumption C5. F_N' and F_N'' correspond to the derivatives of F_N.

Since the cost function F is assumed to be twice continuously Fréchet-differentiable, its first derivative is also Lipschitz continuous

$$\|F'(u) - F'(v)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})} \le \hat{L}\|u - v\|_{\mathcal{U}} \qquad \forall u, v \in U_*,$$

where, without loss of generality, we also assume $\hat{L} \le L$.

Since $\|F''(u_*)^{-1}\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} \le \delta$ and

$$\begin{aligned} &\delta\|F''(u_*) - F_N''(\hat{u}^M)\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} \\ &\le \delta\|F''(u_*) - F''(\hat{u}^M)\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} + \delta\|F''(\hat{u}^M) - F_N''(\hat{u}^M)\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} \\ &\le \delta L\zeta_M^{(2)} + \delta\zeta_N^{(1)} \le \delta\hat{\zeta} < 1 \end{aligned}$$

hold for a constant $\hat{\zeta} \in \mathbb{R}$ if M and N are sufficiently large, the Banach lemma [24, Thm. V.2.4] yields the existence of $F''(\hat{u}^M)^{-1}$ with

$$\|F''(\hat{u}^M)^{-1}\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} \leq \frac{\delta}{1-\delta\hat{\zeta}} =: \hat{\delta}.$$

Analogously, since it is known that $\rho_* \leq \frac{2}{3\delta L} < \frac{1}{\delta L}$ [37], the inequalities

$$\delta\|F''(u_i) - F''(u_*)\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} \leq \delta L\|u_i - u_*\|_{\mathcal{U}} \leq \delta L\rho_* < 1$$

hold yielding the existence of $F''(u_i)^{-1}$ with

$$\|F''(u_i)^{-1}\|_{\mathcal{L}(\mathcal{L}(\mathcal{U},\mathbb{R}),\mathcal{U})} \leq \frac{\delta}{1-\delta L\rho_*} =: \hat{\delta}.$$

Therefore, possibly by redefining δ by the constant $\hat{\delta}$, it can be assumed that

$$\|F''(u_i)^{-1}\|_{\mathcal{L}(\mathcal{L}(\mathcal{U},\mathbb{R}),\mathcal{U})} \leq \delta \qquad \forall i \in \mathbb{N}, \tag{2.1}$$

$$\|F''_N(\hat{u}^M)^{-1}\|_{\mathcal{L}(\mathcal{L}(\mathcal{U},\mathbb{R}),\mathcal{U})} \leq \delta \qquad \forall \hat{u}^M, \forall N \in \mathbb{N}, \tag{2.2}$$

for M and N satisfying $\delta L\zeta_M^{(2)} + \delta\zeta_N^{(1)} \leq \delta\hat{\zeta} < 1$.

The next theorem presents sufficient conditions for the existence of a solution of the minimization problem

$$\min_{u^M \in \mathcal{U}^M} F_N(u^M) \tag{2.3}$$

and describes the convergence behavior of Newton's method for $M, N \to \infty$.

Theorem 2.1. *Let* (C1)–(C5) *be satisfied. Assume that the discretization parameters M and N fulfill the condition*

$$\zeta_{MN} := 2\delta\left(\max\{1,L\} + \frac{1}{2\delta}\right)\left(\zeta_N^{(1)} + \zeta_M^{(2)}\right) \leq \min\left\{\hat{\rho}_*, \frac{1}{\delta L}\right\}. \tag{2.4}$$

Then the discretized Newton's method has a local solution $u_*^M \in \hat{U}_*$, and the discretization error satisfies

$$\|u_*^M - u_*\|_{\mathcal{U}} \leq \zeta_{MN}.$$

Proof. The basic idea is to apply Kantorowitsch's theorem [24] to Newton's method with starting point $u_0^M = \hat{u}^M \in \hat{U}_*$ in order to obtain the existence of a solution u_*^M of the infinite dimensional minimization problem. The corresponding geometrical situation is sketched in Figure 2.1.

First, Assumptions C2–C4, equation (2.2), and the condition (2.4) yield the inequalities

$$\begin{aligned} 2h &:= 2\delta L\|F''_N(\hat{u}^M)^{-1}F'_N(\hat{u}^M)\|_{\mathcal{U}} \\ &\leq 2\delta L\|F''_N(\hat{u}^M)^{-1}\|_{\mathcal{L}(\mathcal{L}(\mathcal{U},\mathbb{R}),\mathcal{U})}\|F'_N(\hat{u}^M)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})} \\ &\leq 2\delta^2 L\left(\|F'_N(\hat{u}^M) - F'(\hat{u}^M)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})} + \|F'(\hat{u}^M) - F'(u^*)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})}\right) \\ &\leq 2\delta^2 L\left(\zeta_N^{(1)} + L\|\hat{u}^M - u^*\|_{\mathcal{U}}\right) \\ &\leq 2\delta^2 L\max\{1,L\}\left(\zeta_N^{(1)} + \zeta_M^{(2)}\right) \qquad (2.5) \\ &\leq 2\delta^2 L\left(\max\{1,L\} + \frac{1}{2\delta}\right)\left(\zeta_N^{(1)} + \zeta_M^{(2)}\right) \leq \delta L\zeta_{MN} \leq 1, \end{aligned}$$

which imply the required assumption $h \leq \frac{1}{2}$.

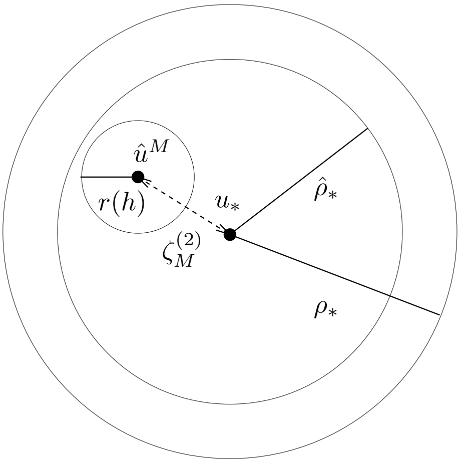


FIG. 2.1. *Balls used for the local convergence investigations.*

To be able to apply the Kantorowitsch theorem only the condition $U(\hat{u}^M, r(h)) \subset U(u_*, \hat{\rho}_*)$ remains to be checked. Because of Assumption C4 this condition can be rewritten as

$$r(h) := \frac{1}{\delta L}(1 - \sqrt{1-2h}) \leq \hat{\rho}_* - \zeta_M^{(2)}, \tag{2.6}$$

which is proven by using the given assumptions and inequality (2.6):

$$\begin{aligned} r(h) &\leq \frac{1}{\delta L}\left(1 - \sqrt{1 - 2\delta^2 L \max\{1, L\}\left(\zeta_N^{(1)} + \zeta_M^{(2)}\right)}\right) \\ &\leq \frac{2\delta \max\{1, L\}(\zeta_N^{(1)} + \zeta_M^{(2)})}{1 + \sqrt{1 - 2\delta^2 L \max\{1, L\}(\zeta_N^{(1)} + \zeta_M^{(2)})}} \\ &\leq 2\delta \max\{1, L\}\left(\zeta_N^{(1)} + \zeta_M^{(2)}\right) \\ &\leq 2\delta\left(\max\{1, L\} + \frac{1}{2\delta}\right)\left(\zeta_N^{(1)} + \zeta_M^{(2)}\right) - \zeta_M^{(2)} \\ &\leq \zeta_{MN} - \zeta_M^{(2)} \leq \hat{\rho}_* - \zeta_M^{(2)}. \end{aligned}$$

Finally, this yields the existence of a solution $u_*^M \in U(\hat{u}^M, r(h))$ and

$$\|u_*^M - u_*\|_{\mathcal{U}} \leq \|u_*^M - \hat{u}^M\|_{\mathcal{U}} + \|\hat{u}^M - u_*\|_{\mathcal{U}} \leq \zeta_{MN} - \zeta_M^{(2)} + \zeta_M^{(2)} = \zeta_{MN}. \qquad \square$$

Now we have proven that a solution $u_*^M \in U(\hat{u}^M, r(h)) \subset U(u_*, \hat{\rho}_*)$ of the discretized minimization problem exists. Next, the main theorem of the paper shows that the discretized Newton's method converges to the solution u_*^M for any $u_0^M \in U(u_*, \rho_1)$ for sufficiently small ρ_1.

THEOREM 2.2. *Let Assumptions* C1–C5 *be satisfied, and assume that the discretization parameters M and N satisfy the condition $\zeta_{MN} \leq \frac{1}{6}\min\left\{\frac{\hat{\rho}_*}{4}, \frac{1}{6L\delta+1}\right\}$. Then the discretized Newton's method converges to u_*^M for all $u_0^M \in U(u_*, \rho_1)$ with*

$\rho_1 = \frac{3}{4}\min\{\frac{1}{3L\delta}, \frac{\hat{\rho}_*}{2}\}$. *In addition, for all starting points satisfying the condition*

$$\|u_0^M - u_0\|_{\mathcal{U}} \le \tau$$

with $\tau = \frac{2(\frac{1}{2}+\|u_0-u_*\|_{\mathcal{U}})\zeta_{MN}}{b^2+\sqrt{b^2-6L\delta(\frac{1}{2}+\|u_0-u_*\|_{\mathcal{U}})\zeta_{MN}}}$ *and* $b = 1 + \frac{1}{2}\zeta_{MN} - 2\delta L\|u_0^M - u_*\|_{\mathcal{U}}$, *the following convergence properties hold* ($c_1, c_2, c_3, c_4 \in \mathbb{R}, n \in \mathbb{N}$)*:*

$$\|u_{n+1}^M - u_*^M\|_{\mathcal{U}} \le c_1\|u_n^M - u_*^M\|_{\mathcal{U}}^2, \tag{2.7}$$

$$\|u_n^M - u_n\|_{\mathcal{U}} \le c_2\zeta_{MN}, \tag{2.8}$$

$$\|F_N'(u_n^M) - F'(u_n)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})} \le c_3\zeta_{MN}, \tag{2.9}$$

$$\|u_n^M - u_*^M\|_{\mathcal{U}} \le \|u_n - u_*\|_{\mathcal{U}} + c_4\zeta_{MN}. \tag{2.10}$$

Proof. The proof is divided into three parts. First, the convergence of the discretized Newton's method is proven for all u_0^M in a suitable ball around u_*. Then, the convergence of the iteration function u_n^M to the infinite dimensional iteration u_n is shown, and third, the proof is finished.

1. Since the assumptions of Theorem 2.1 are fulfilled, the existence of a solution $u_*^M \in \hat{U}_*$ is guaranteed. Now, it will be shown that the convergence of the discretized Newton's method to this solution is guaranteed, if u_0^M is chosen to be in $U(u_*, \rho_2)$ with $\rho_2 = \min\left\{\frac{1}{3L\delta}, \frac{\hat{\rho}_*}{2}\right\}$ and $\zeta_{MN} \le \min\left\{\frac{\hat{\rho}_*}{4}, \frac{1}{6L\delta+1}\right\}$.

   The inequalities

$$\begin{aligned}\|u_*^M - u_*\|_{\mathcal{U}} + \|u_0^M - u_*^M\|_{\mathcal{U}} &\le 2\|u_*^M - u_*\|_{\mathcal{U}} + \|u_0^M - u_*\|_{\mathcal{U}} \\ &\le 2\zeta_{MN} + \rho_2 \le \hat{\rho}_*\end{aligned}$$

   imply that $U(u_*^M, \|u_0^M - u_*^M\|_{\mathcal{U}}) \subset \hat{U}_*$, which means that Assumptions C1–C5 are also valid in $U(u_*^M, \|u_0^M - u_*^M\|_{\mathcal{U}})$.

   Since $\|F''(u_*)^{-1}\|_{\mathcal{L}(\mathcal{L}(\mathcal{U},\mathbb{R}),\mathcal{U})} \le \delta$ and since the definition of ζ_{MN} implies $\zeta_N^{(1)} \le \frac{1}{2\delta}\zeta_{MN}$, the following inequalities are obtained using Theorem 2.1:

$$\begin{aligned}&\delta\|F_N''(u_*^M) - F''(u_*)\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} \\ &\le \delta\|F_N''(u_*^M) - F_N''(u_*)\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} + \delta\|F_N''(u_*) - F''(u_*)\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} \\ &\le \delta L\|u_*^M - u_*\|_{\mathcal{U}} + \delta\zeta_N^{(1)} \le \delta L\zeta_{MN} + \frac{1}{2}\zeta_{MN} \le \frac{\delta L + \frac{1}{2}}{6L\delta + 1} < 1.\end{aligned}$$

   Now the Banach lemma yields $\|F_N''(u_*^M)^{-1}\|_{\mathcal{L}(\mathcal{L}(\mathcal{U},\mathbb{R}),\mathcal{U})} \le \frac{\delta}{1-(\delta L+\frac{1}{2})\zeta_{MN}}$.

   So all assumptions of the theorem on the q-quadratic convergence of Newton's method are fulfilled, yielding the convergence to u_*^M for all u_0^M in a neighborhood of u_*^M. By applying a refined formulation of this theorem, which is given by Rheinboldt [3], [37], the convergence is even guaranteed for all $u_0^M \in U(u_*^M, r_*)$ with $r_* = \frac{2}{3L\|F_N''(u_*^M)^{-1}\|_{\mathcal{L}(\mathcal{L}(\mathcal{U},\mathbb{R}),\mathcal{U})}}$.

   Only $U(u_*, \rho_2) \subset U(u_*^M, r_*)$ remains to be proven in order to ensure the convergence for any $u_0^M \in U(u_*, \rho_2)$, which is implied by

$$\begin{aligned}\|u_0^M - u_*^M\|_{\mathcal{U}} &\leq \|u_0^M - u_*\|_{\mathcal{U}} + \|u_* - u_*^M\|_{\mathcal{U}} \\ &\leq \rho_2 + \zeta_{MN} \\ &\leq \frac{1 + 3L\delta\zeta_{MN}}{3L\delta} \\ &< \frac{2(1 - L\delta\zeta_{MN} - \frac{1}{2}\zeta_{MN})}{3L\delta} \\ &\leq \frac{2}{3L\|F_N''(u_*^M)^{-1}\|_{\mathcal{L}(\mathcal{L}(\mathcal{U},\mathbb{R}),\mathcal{U})}} =: r_*.\end{aligned}$$

To sum up, the discretized Newton's method converges to u_*^M for all $u_0^M \in U(u_*, \rho_2)$ with the q-quadratic convergence rate

$$\|u_{n+1}^M - u_*^M\|_{\mathcal{U}} \leq \delta L \|u_n^M - u_*^M\|_{\mathcal{U}}^2,$$

where $c_1 = \delta L$ is independent of the discretization parameters M and N.

2. In this part a proof by induction is used to show

$$\|u_n^M - u_n\|_{\mathcal{U}} \leq \tau \leq c_2\zeta_{MN} \tag{2.11}$$

for all $u_0^M \in U(u_*, \rho_1)$, $\rho_1 = \frac{3}{4}\rho_2$, and for all discretization parameters M and N fulfilling $\zeta_{MN} \leq \frac{1}{6}\min\left\{\frac{\hat{\rho}_*}{4}, \frac{1}{6L\delta+1}\right\}$, where τ is given by

$$\begin{aligned}\tau &= \frac{2(\frac{1}{2} + \|u_0 - u_*\|_{\mathcal{U}})\zeta_{MN}}{b^2 + \sqrt{b^2 - 6L\delta(\frac{1}{2} + \|u_0 - u_*\|_{\mathcal{U}})\zeta_{MN}}} \\ &\leq \frac{2(\frac{1}{2} + \|u_0 - u_*\|_{\mathcal{U}})\zeta_{MN}}{b^2} =: c_2\zeta_{MN}\end{aligned}$$

with $b = 1 + \frac{1}{2}\zeta_{MN} - 2\delta L\|u_0^M - u_*\|_{\mathcal{U}}$. The constant τ is well defined, since the inequalities

$$6L\delta\left(\frac{1}{2} + \|u_0 - u_*\|_{\mathcal{U}}\right)\zeta_{MN} \leq \frac{2L\delta + 1}{4(6L\delta + 1)} < \frac{1}{4} \text{ and } b \geq 1 - 2\delta L\rho_1 \geq \frac{1}{2}$$

imply $b^2 \geq \frac{1}{4} \geq 6L\delta\left(\frac{1}{2} + \|u_0 - u_*\|_{\mathcal{U}}\right)\zeta_{MN}$.

While the assertion (2.11) is fulfilled by assumption for $n = 0$, the induction step is based on the simple decomposition

$$\begin{aligned}u_{i+1}^M - u_{i+1} = F_N''(u_i^M)^{-1}\{&[F_N''(u_i^M)(u_i^M - u_i) - F_N'(u_i^M) + F_N'(u_i)] \\ &+[(F_N''(u_i^M) - F_N''(u_i))(F''(u_i)^{-1}F'(u_i))] \\ &+[F_N''(u_i)(F''(u_i)^{-1}F'(u_i)) - F'(u_i)] \\ &+[F'(u_i) - F_N'(u_i)]\}.\end{aligned} \tag{2.12}$$

Assumptions C1–C4, equation (2.1), and the definition of ζ_{MN} imply

$$\begin{aligned}&\delta\|F_N''(u_i^M) - F''(u_i)\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} \\ &\leq \delta\|F_N''(u_i^M) - F_N''(u_i)\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} + \delta\|F_N''(u_i) - F''(u_i)\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} \\ &\leq \delta(L\tau + \zeta_N^{(1)}) \leq \delta L\tau + \frac{1}{2}\zeta_{MN} \leq \frac{\delta L\zeta_{MN} + 2\|u_0 - u_*\|_{\mathcal{U}}\delta L\zeta_{MN}}{1 - 2\delta L\|u_0 - u_*\|_{\mathcal{U}}} + \frac{1}{2}\zeta_{MN} \\ &\leq \frac{\frac{1}{3}\delta L + \frac{1}{4}}{6L\delta + 1} < 1\end{aligned}$$

resulting in the inequality $\|F_N''(u_i^M)^{-1}\|_{\mathcal{L}(\mathcal{L}(\mathcal{U},\mathbb{R}),\mathcal{U})} \leq \frac{\delta}{1-(L\delta\tau+\frac{1}{2}\zeta_{MN})}$. Using a standard argument [16, Lemma 2.4.2] we obtain

$$\|F_N''(u_i^M)(u_i^M-u_i)-F_N'(u_i^M)+F_N'(u_i)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})} \leq \frac{1}{2}L\|u_i^M-u_i\|_{\mathcal{U}}^2 \leq \frac{1}{2}L\tau^2,$$

and the derived convergence assertion $\|u_i-u_*\|_{\mathcal{U}} \leq \|u_0-u_*\|_{\mathcal{U}}$ yields

$$\begin{aligned}&\|(F_N''(u_i^M)-F_N''(u_i))(F''(u_i)^{-1}F'(u_i))\|_{\mathcal{L}(\mathcal{U},\mathbb{R})}\\ &\leq L\|u_i^M-u_i\|_{\mathcal{U}}\|u_i-u_{i+1}\|_{\mathcal{U}} \leq 2L\tau\|u_0-u_*\|_{\mathcal{U}}.\end{aligned}$$

The assumptions of the theorem lead to

$$\begin{aligned}&\|F_N''(u_i)(F''(u_i)^{-1}F'(u_i))-F'(u_i)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})}\\ &\leq \|-F_N''(u_i)(u_{i+1}-u_i)+F''(u_i)(u_{i+1}-u_i)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})}\\ &\leq \|F_N''(u_i)-F''(u_i)\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))}\|u_{i+1}-u_i\|_{\mathcal{U}}\\ &\leq \zeta_N^{(1)}2\|u_0-u_*\|_{\mathcal{U}} \leq \frac{1}{\delta}\zeta_{MN}\|u_0-u_*\|_{\mathcal{U}}\end{aligned}$$

and $\|F'(u_i)-F_N'(u_i)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})} \leq \zeta_N^{(1)} \leq \frac{1}{2\delta}\zeta_{MN}$. Using the decomposition (2.12) the last inequalities complete the induction proof by

$$\begin{aligned}&\|u_{i+1}^M-u_{i+1}\|_{\mathcal{U}}\\ &\leq \frac{\delta}{1-(L\delta\tau+\frac{1}{2}\zeta_{MN})}\left\{\frac{1}{2}L\tau^2+2L\|u_0-u_*\|_{\mathcal{U}}\tau+\left(\frac{1}{2}+\|u_0-u_*\|_{\mathcal{U}}\right)\frac{\zeta_{MN}}{\delta}\right\}\\ &=\tau.\end{aligned}$$

The last equality is based on the fact that τ is equal to the smallest solution of the quadratic equation $3L\delta\tau^2-2b\tau+2\zeta_{MN}(\frac{1}{2}+\|u_0-u_*\|_{\mathcal{U}})=0$.

3. Finally, inequality (2.9) is shown by

$$\begin{aligned}&\|F_N'(u_n^M)-F'(u_n)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})}\\ &\leq \|F_N'(u_n^M)-F_N'(u_n)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})}+\|F_N'(u_n)-F'(u_n)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})}\\ &\leq L\|u_n^N-u_n\|_{\mathcal{U}}+\zeta_{MN} \leq (Lc_2+1)\zeta_{MN} =: c_3\zeta_{MN},\end{aligned}$$

and inequality (2.10) results from

$$\begin{aligned}&\|(u_n^M-u_*^M)-(u_n-u_*)\|_{\mathcal{U}} \leq \|u_n^M-u_n\|_{\mathcal{U}}+\|u_*^M-u_*\|_{\mathcal{U}}\\ &\leq c_2\zeta_{MN}+\zeta_{MN} \leq (c_2+1)\zeta_{MN} =: c_4\zeta_{MN}. \qquad \square\end{aligned}$$

At the end of this section, the influence of the derived convergence results on the required number of iterations is analyzed with respect to the stopping criterion

$$\begin{aligned}t_i &:= \|F''(u_i)(u_{i+1}-u_i)-F'(u_i)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})} \leq \text{TOL},\\ t_i^{MN} &:= \|F_N''(u_i^M)(u_{i+1}^M-u_i^M)-F_N'(u_i^M)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})} \leq \text{TOL}.\end{aligned}$$

The number $i(\text{TOL})$, respectively, $i^{MN}(\text{TOL})$, is defined to be the smallest iteration number satisfying the stopping criterion, i.e., $i^{MN}(\text{TOL}) := \min\{i : t_i^{MN} \leq \text{TOL}\}$, respectively, $i(\text{TOL}) := \min\{i : t_i \leq \text{TOL}\}$.

LEMMA 2.3. *Let the assumptions of Theorem* 2.2 *be satisfied. Then for every* $\text{TOL} > 0$ *and* $\epsilon > 0$, *there exist parameters* $\hat{M}, \hat{N} \in \mathbb{N}$ *such that*

$$\begin{aligned} &i(\text{TOL} + \epsilon) \leq i^{MN}(\text{TOL}) \leq i(\text{TOL}) \qquad \forall M \geq \hat{M}, N \geq \hat{N}, \text{ and} \\ &i^{MN}(\text{TOL}) = i(\text{TOL}) \qquad \forall M \geq \hat{M}, N \geq \hat{N}. \end{aligned}$$

Proof. By using the different assumptions and Theorem 2.2, the inequality

$$|t_i - t_i^{MN}| \leq c\zeta_{MN} \tag{2.13}$$

can easily be derived. The remaining part corresponds to the proof of Corollary 4.2.6 in [21]. Based on (2.13) it is possible to choose $\hat{M}$ and $\hat{N}$ such that the inequalities

$$\begin{aligned} |t_{i(\text{TOL})} - t_{i(\text{TOL})}^{MN}| \leq \text{TOL} - t_{i(\text{TOL})} \qquad \forall M \geq \hat{M}, N \geq \hat{N}, \\ |t_i - t_i^{MN}| \leq \epsilon \qquad \forall M \geq \hat{M}, N \geq \hat{N}, \end{aligned}$$

are fulfilled. Then, for all $M \geq \hat{M}, N \geq \hat{N}$,

$$t_{i(\text{TOL})}^{MN} \leq t_{i(\text{TOL})} + |t_{i(\text{TOL})} - t_{i(\text{TOL})}^{MN}| \leq \text{TOL}$$

implies $i^{MN}(\text{TOL}) \leq i(\text{TOL})$, and $i(\text{TOL} + \epsilon) \leq i^{MN}(\text{TOL})$ is derived using

$$t_{i^{MN}(\text{TOL})} \leq t_{i^{MN}(\text{TOL})}^{MN} + |t_{i^{MN}(\text{TOL})} - t_{i^{MN}(\text{TOL})}^{MN}| \leq \text{TOL}.$$

The definition of $i(\text{TOL})$ ensures that $t_{i(\text{TOL})-1} > \text{TOL}$. By choosing $\epsilon = \frac{t_{i(\text{TOL})-1} - \text{TOL}}{2}$ we get $\text{TOL} + \epsilon < t_{i(\text{TOL})-1}$. But to obtain a residual that is less than $t_{i(\text{TOL})-1}$, at least one more iteration is necessary, i.e.,

$$i(\text{TOL} + \epsilon) \geq (i(\text{TOL}) - 1) + 1. \qquad \square$$

This explains the frequently observed numerical behavior that the required iteration number is independent of the discretization parameters if these are large enough.

3. Optimal shape design problems. The computing of the first two derivatives of the cost function is based on the adjoint $p \in \mathcal{V}$ given by the solution of the equation

$$a(u; \eta, p) = J_y(u, S(u), z(u))(\eta) \qquad \forall \eta \in \mathcal{V}. \tag{3.1}$$

Some assumptions are listed here for the infinite dimensional problems in order to adapt the theorem of Newton's q-quadratic convergence rate. Although they could be weakened, for the sake of simplicity we have stated them in a rather general version.

ASSUMPTION A1. $\Omega \subset \mathbb{R}^2$ is Lipschitz continuous.

ASSUMPTION A2. $a(u; \cdot, \cdot)$ is a $\mathcal{V}$-elliptic bilinear form for all $u \in \mathcal{U}_{ad}$.

ASSUMPTION A3. The second Fréchet-derivatives of $f(u) \in \mathcal{V}'$, $a_{ij}(u) \in \mathcal{L}^\infty(\Omega)$, and $b(u) \in \mathcal{L}^\infty(\Gamma_1)$ exist and are Lipschitz continuous in u.

ASSUMPTION A4. $J : \mathcal{U}_{ad} \times \mathcal{V} \times \mathcal{Z} \otimes \mathbb{R}$ is twice continuously Fréchet-differentiable and $J_y(u, S(u), z(u)) \in \mathcal{V}'$, $J_{yu}(u, S(u), z(u))(v) \in \mathcal{V}'$ for all $v \in \mathcal{U}$, $J_{yy}(u, S(u), z(u)) \in (\mathcal{V} \times \mathcal{V})'$, $J_{yz}(u, S(u), z(u)) \in (\mathcal{V} \times \mathcal{Z})'$ are Lipschitz continuous with respect to all components.

Now, the well-known theorem can be written in the following version.

THEOREM 3.1. *Let Assumptions* A1–A4 *be satisfied. Let $z(u) \in \mathcal{Z}$ be twice continuously Fréchet-differentiable in an open neighborhood U of the solution u_*, and $z''(u)$, as well as all partial derivatives of J, be Lipschitz continuous. The existence of $F''(u_*)^{-1} \in \mathcal{L}(\mathcal{L}(\mathcal{U}, \mathbb{R}), \mathcal{U})$ and of the solution u_* in the interior of $\mathcal{U}_{ad}$ is further ensured.*

Then there exists a $\rho_* > 0$, such that the iterative sequence $\{u_n\}$ generated by Newton's method converges to the solution u_* for every initial value u_0 with $\|u_0 - u_*\|_{\mathcal{U}} \leq \rho_*$. Moreover, there is a suitable constant C depending on U and on the existing Lipschitz constant L of the second derivative of the cost function, such that

$$\|u_n - u_*\|_{\mathcal{U}} \leq C\|u_{n-1} - u_*\|_{\mathcal{U}}^2. \tag{3.2}$$

Proof. The theorem is stated in such a way that all the assumptions for stating the first derivative of the cost function are fulfilled (see [32], [31]). Hence, Newton's method is well defined and the desired assertion is derived as an application of the general theorem on Newton's q-quadratic convergence rate [40, p. 208]. □

The finite element method is used for the discretization of the variational equations in $\Omega \subset \mathbb{R}^2$, as well as for the calculation of Newton's equation in $I \subset \mathbb{R}$. Let $\{\psi_k\}_{k=1}^M, \psi_k \in \mathcal{U}^M$, be the one-dimensional spline basis functions and let $\{\phi_\tau\}_{\tau=1}^N, \phi_\tau \in \mathcal{V}^N$, be the two-dimensional spline basis functions, which are used to solve the distinct variational equations. In the following a vector consisting of the N or M basis coefficient is always denoted with an arrow on top of the same letter as the coefficient, e.g., $\vec{y}^N$.

For the optimal shape design problems Newton's equation is now written as

$$\langle F_N''(u^M)(w^M), v^M\rangle = \langle -F_N'(u^M), v^M\rangle \qquad \forall v^M \in \mathcal{U}^M, \tag{3.3}$$

which is equivalent to the linear system $H\vec{w}^M = d$, with the Hessian matrix $H = \langle F_N''(u^M)(\psi_i), \psi_j\rangle_{i,j=1}^M \in \mathbb{R}^{M\times M}$ and the vector $d = \langle -F_N'(u^M), \psi_i\rangle_{i=1}^M \in \mathbb{R}^M$. In [31] it has been proven that the right-hand side vector can be computed by

$$\begin{aligned} d = {} & \langle -J_u(u^M, y^N, z(u^M)) - J_z(u^M, y^N, z(u^M))z'(u^M), \psi_i\rangle_{i=1}^M \\ & -H_1\vec{p}^N - H_4(\vec{u}^M - \vec{u}_{\mathcal{T}}^M) \end{aligned}$$

and the required result of the matrix-vector multiplication $H\vec{w}^M$ is given by

$$H\vec{w}^M = H_1\vec{\hat{p}}^N + H_2\vec{\hat{y}}^N + H_3\vec{w}^M + H_4\vec{w}^M$$

for any vector $\vec{w}^M$, where $H_i, i = 1, \ldots, 4$, are specific sparse auxiliary matrices. The state y^N and the adjoint p^N, as well as the derivative of the state $\hat{y}^N = \hat{y}^N(w^M)$ and of the adjoint $\hat{p}^N = \hat{p}^N(w^M)$, are given by the solution of a variational problem of the kind (1.4).

Now, the question will be answered as to how the discretization parameter N influences the state, the adjoint, and the derivatives of the cost function. A detailed analysis is necessary, since the corresponding errors are influenced not only explicitly but also implicitly; for example, the error of the adjoint depends also on the error of the state. The term semidiscretized problem is used in the sense that the discretization is done only with respect to the parameter N, but not with respect to M. Nevertheless, since $\mathcal{U}^M \subset \mathcal{U}$, the approximation properties are also true for a fixed $u^M \in \mathcal{U}^M$.

THEOREM 3.2. *Let Assumptions* A1–A4 *be satisfied, $z(u) \in \mathcal{Z}$ be twice continuously Fréchet-differentiable, $J \in \mathcal{C}^2(\mathcal{U}, \mathcal{V}, \mathcal{Z})$, and $z''(u)$, as well as all partial derivatives of J, be further locally Lipschitz continuous with respect to all components, such that Newton's method is well defined.*

If the infinite dimensional solution $\chi \in \mathcal{V}$ of a variational equation with respect to the bilinear form $a(u; \chi, \eta)$, $u \in \mathcal{U}$ fixed, and the corresponding semidiscretized solution $\chi^N \in \mathcal{V}^N$ with respect to $a(u; \chi^N, \eta^N)$, satisfy the convergence condition ($c_\chi \in \mathbb{R}$)

$$\|\chi - \chi^N\|_{\mathcal{V}} \leq c_\chi \zeta(N) \qquad \text{as } N \to \infty,$$

then the following assertions for $N \to \infty$ are true ($c_1, c_2, c_3, c_4 \in \mathbb{R}$):

$$\begin{aligned} \|y - y^N\|_{\mathcal{V}} &\leq c_1 \zeta(N), \\ \|p - p^N\|_{\mathcal{V}} &\leq c_2 \zeta(N), \\ \|F'(u) - F'_N(u)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})} &\leq c_3 \zeta(N), \\ \|F''(u) - F''_N(u)\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} &\leq c_4 \zeta(N). \end{aligned}$$

Proof. As a direct implication of the given assumptions, the inequality

$$\|y - y^N\|_{\mathcal{V}} \leq c_y \zeta(N) \tag{3.4}$$

is derived for the exact and discretized state. Analogously, the inequality

$$\|p - \bar{p}^N\|_{\mathcal{V}} \leq c_p \zeta(N) \tag{3.5}$$

is obtained for the infinite adjoint p, and the solution, $\bar{p}^N$, of the equation

$$a(u^M; \bar{p}^N, \eta^N) = J_y(u^M; y)(\eta^N) \qquad \forall \eta^N \in \mathcal{V}^N.$$

Our main interest is not the solution $\bar{p}^N$, but the solution $p^N \in \mathcal{V}^N$ of the equation

$$a(u^M; p^N, \eta^N) = J_y(u^M; y^N)(\eta^N) \qquad \forall \eta^N \in \mathcal{V}^N,$$

which is completely discretized with respect to N. The last two equations imply

$$a(u^M; \bar{p}^N - p^N, \eta^N) = J_y(u^M; y)(\eta^N) - J_y(u^M; y^N)(\eta^N) \qquad \forall \eta^N \in \mathcal{V}^N.$$

Using the $\mathcal{V}$-ellipticity of the bilinear form yields the inequalities (L Lipschitz constant)

$$\begin{aligned} \|\bar{p}^N - p^N\|_{\mathcal{V}} &\leq c\|J_y(u^M; y) - J_y(u^M; y^N)\|_{\mathcal{V}'} \\ &\leq cL\|y - y^N\|_{\mathcal{V}} \\ &\leq c c_y L \zeta(N). \end{aligned} \tag{3.6}$$

This leads to an estimate of the infinite adjoint and of the semidiscretized adjoint

$$\begin{aligned} \|p - p^N\|_{\mathcal{V}} &\leq \|p - \bar{p}^N\|_{\mathcal{V}} + \|\bar{p}^N - p^N\|_{\mathcal{V}} \\ &\leq c_p \zeta(N) + c c_y L \zeta(N) \\ &=: c_2 \zeta(N) \end{aligned} \tag{3.7}$$

by exploiting (3.5) and (3.6). Finally, since every summand of $F'(u^M)$ depends Lipschitz continuously on at least y or p, the result is verified using the inequalities (3.4) and (3.7).

Since the proof for the second derivative of the cost function is nearly the same, it is omitted here. □

Theorem 3.2 reduces the approximation properties of the cost function derivatives to the approximation property of the discretized variational equation. This issue is well investigated in the research of finite elements. If the underlying mesh satisfies certain conditions, then the convergence is guaranteed (see, e.g., Ciarlet [13]).

COROLLARY 3.3. *Let Assumptions* A1–A4 *be satisfied,* $z(u^M) \in \mathcal{Z}$ *be twice continuously Fréchet-differentiable,* $J \in \mathcal{C}^2(\mathcal{U}, \mathcal{V}, \mathcal{Z})$*, and* $z''(u)$*, as well as all partial derivatives of* J*, be further locally Lipschitz continuous with respect to all components such that Newton's method is well defined.*

Consider a regular triangulation, where all finite elements are affine-equivalent to a reference element in the sense of Ciarlet [13]*. Then, the use of linear spline basis functions implies the following assertions for* $N \to \infty$*:*

$$\begin{aligned} \|y - y^N\|_{\mathcal{V}} &\to 0, \\ \|p - p^N\|_{\mathcal{V}} &\to 0, \\ \|F'(u) - F'_N(u)\|_{\mathcal{L}(\mathcal{U},\mathbb{R})} &\to 0, \\ \|F''(u) - F''_N(u)\|_{\mathcal{L}(\mathcal{U},\mathcal{L}(\mathcal{U},\mathbb{R}))} &\to 0. \end{aligned}$$

Proof. The assumptions concerning the finite element mesh and the finite element functions are defined in such a way that Theorem 18.2 of Ciarlet [13] can be applied. Thus, for fixed $u \in \mathcal{U}$, the solution of the semidiscretized variational equation converges to the infinite dimensional solution in the $\mathcal{V}$ norm. In other words, $\zeta(N) \to 0$ as $N \to \infty$, and a slight modification of Theorem 3.2 yields the desired result. □

For the optimal shape design problems in question, Assumptions C1 and C3 can now be replaced with the modified Assumptions D1 and D3 by using the implications of Theorem 3.1, Theorem 3.2, and Corollary 3.3, respectively.

ASSUMPTION D1. Let A1–A4 be satisfied. Let $z(u) \in \mathcal{Z}$ be twice continuously Fréchet-differentiable in U_*, and let $z''(u)$, as well as all partial derivatives of J, be Lipschitz continuous. Furthermore, let the existence of $F''(u_*)^{-1} \in \mathcal{L}(\mathcal{L}(\mathcal{U},\mathbb{R}),\mathcal{U})$ and of the solution u_* in the interior of $\mathcal{U}_{ad}$ be guaranteed.

ASSUMPTION D3. The infinite dimensional solution $\chi \in \mathcal{V}$ of a variational equation with respect to the bilinear form $a(u;\chi,\eta)$, $u \in \mathcal{U}$ fixed, and the corresponding semidiscretized solution $\chi^N \in \mathcal{V}^N$ with respect to $a(u;\chi^N,\eta^N)$, satisfy the convergence condition ($\zeta_N^{(1)} \overset{N\to\infty}{\longrightarrow} 0$): $\|\chi - \chi^N\|_{\mathcal{V}} \le \zeta_N^{(1)}$.

Assumption D3 means that a convergent discretization scheme has to be used to solve the variational equations. Corollary 3.3 translates this to the case of linear spline basis functions with a regular triangulation that is often used in computation.

It should further be mentioned that in the special case of $u_* \in \mathcal{C}^2(I)$, the inequality

$$\|I_M u_* - u_*\|_{\mathcal{W}^{1,\infty}} \le \frac{1}{M-1}\|u_*\|_{\mathcal{W}^{2,\infty}}$$

(see, e.g., [4, p. 218]) with respect to the linear interpolation operator $I_M : \mathcal{U} \to \mathcal{U}^M$ already guarantees C4 for linear spline basis functions by choosing $\hat{u}^M = I_M u_*$.

Assumption C5 has to be verified for the considered kind of discretization scheme as it has been demonstrated by Chenais [12] for a specific class of problems. This assumption could also be weakened by supposing the convergence of F'_N and F''_N to the derivatives of the discretized cost function with respect to $N, M \to \infty$. Several convergence assertions are given as an immediate consequence of Theorem 2.2.

LEMMA 3.4. *Let the assumptions of Theorem* 2.2 *be satisfied. Then for all starting points fulfilling* $\|u_0^M - u_0\|_{\mathcal{U}} \le \tau$ *the following assertions hold for the state* y*:*

$$\begin{aligned}
&\|y_*^N - y_*\|_{\mathcal{V}} \le \hat{c}_1\zeta_{MN},\\
&\|y_n^N - y_n\|_{\mathcal{V}} \le \hat{c}_2\zeta_{MN}, \ \text{and}\\
&\|y_n^N - y_*^N\|_{\mathcal{V}} \le \|y_n - y_*\|_{\mathcal{V}} + \hat{c}_3\zeta_{MN}.
\end{aligned}$$

Analogous results are true for the adjoint p*, as well as for the derivatives* $\hat{y}$ *and* $\hat{p}$*.*

Proof. Since the corresponding solution operators are proven to be Lipschitz continuous with respect to u, this is an easy implication of Theorem 2.2. □

4. Numerical results. The numerical results presented next correspond to the considered class of optimal shape design problems with the cost function ($\varepsilon = 10^{-5}$)

$$\min_{u\in\mathcal{U}} \int_\Omega (y-z)^2 u dx + \frac{\varepsilon}{2}\|u\|^2_{\mathcal{H}^1(I)}.$$

Based on the definition of the initial and exact control

$$u_0(x_2) = 1.1 \text{ and } u_*(x_2) = \begin{cases} 1 - 0.1x_2, & x_2 \le 0.5,\\ 0.9 + 0.1x_2, & x_2 > 0.5,\end{cases}$$

z is defined by $z = u_* x_1 e^{u_* x_1 x_2}\sin(u_* x_1\pi)\sin(x_2\pi)$. Since the state y is the solution of a boundary value problem with inhomogeneous Dirichlet condition, the splitting $y = \hat{y} + g_D(u)$ must be introduced. The function $g_D(u) = u x_1 e^{u x_1 x_2}\sin(u x_1 \pi)\sin(x_2\pi)$ is defined on the boundary and $\hat{y} = S(u)$ is the solution of the variational equation

$$a(u;\hat{y},\eta) = l(u;\eta) \qquad \forall \eta \in \mathcal{H}_0^1(\Omega),$$

where the bilinear form and the linear functional are defined by

$$\begin{aligned}
a(u;\hat{y},\eta) &= \int_\Omega \frac{1}{u}\left(e^{-ux_1x_2} + e^{ux_1x_2}x_1^2u'^2\right)\hat{y}_{x_1}\eta_{x_1} - e^{ux_1x_2}x_1u'\hat{y}_{x_1}\eta_{x_2}\\
&\quad - e^{ux_1x_2}x_1u'\hat{y}_{x_2}\eta_{x_1} + ue^{ux_1x_2}\hat{y}_{x_2}\eta_{x_2}\,dx,\\
l(u;\eta) &= \int_\Omega f(ux_1,x_2)u\,dx - a(u;g_D(u),\eta) \text{ with}\\
f(\tilde{x}_1,x_2) &= (\pi^2\tilde{x}_1 - x_2)\sin(\tilde{x}_1\pi)\sin(x_2\pi) - \pi(2+\tilde{x}_1x_2)\cos(\tilde{x}_1\pi)\sin(x_2\pi)\\
&\quad - e^{2\tilde{x}_1x_2}\left[(2\tilde{x}_1^3 - \tilde{x}_1\pi^2)\sin(\tilde{x}_1\pi)\sin(x_2\pi) + 3\pi\tilde{x}_1^2\sin(\tilde{x}_1\pi)\cos(x_2\pi)\right].
\end{aligned}$$

Each linearized optimal control problem has been solved by the SYMMLQ algorithm (see, e.g., [34]) with the stopping criterion $\|H\vec{w}^M - d\|_2 \le 10^{-9}$. Since for our example the variational equations consist of a symmetric bilinear form, the CG method with a hierarchical and a diagonal preconditioner is implemented to accelerate the expected convergence rate. This iterative method terminates if the $\mathcal{L}^2(\Omega)$ norm of the residual is less than or equal to 10^{-11}. This stopping criterion is rather small in order to exclude the influence of these errors on the observed convergence rate. However, the algorithm could be accelerated by taking advantage of the inexact Newton's method [15]. The $\|\cdot\|_{\mathcal{H}^1(I)}$ norm is used for the Tikhonov regularization term ($u_{\mathcal{T}} \equiv 0$, $\varepsilon = 10^{-4}$) and the nonlinear iteration is stopped if the norm of F_N' is less than 10^{-8}.

TABLE 4.1
Newton's method with various parameters M and N.

IT	SYM	TIME	$\|w^M\|$	$F_N(u_i^M)$	$\|F_N'(u_i^M)\|$	$\|u_i^M - u_*\|$
$M = \sqrt{N} = 17$						
0	0	1	$0.000E+00$	$0.351E-01$	$0.514E-01$	$0.126E+00$
1	22	4	$0.120E+00$	$0.280E-02$	$0.104E-01$	$0.601E-01$
2	24	4	$0.316E-01$	$0.274E-03$	$0.213E-02$	$0.406E-01$
3	20	4	$0.225E-01$	$0.642E-04$	$0.340E-03$	$0.214E-01$
4	19	4	$0.141E-01$	$0.390E-04$	$0.177E-03$	$0.101E-01$
5	19	4	$0.572E-02$	$0.283E-04$	$0.113E-03$	$0.659E-02$
$M = \sqrt{N} = 33$						
0	0	5	$0.000E+00$	$0.341E-01$	$0.255E-01$	$0.126E+00$
1	55	52	$0.105E+00$	$0.215E-02$	$0.455E-02$	$0.413E-01$
2	60	55	$0.305E-01$	$0.959E-04$	$0.693E-03$	$0.189E-01$
3	61	56	$0.127E-01$	$0.124E-04$	$0.816E-04$	$0.945E-02$
4	41	38	$0.642E-02$	$0.762E-05$	$0.261E-04$	$0.464E-02$
5	39	35	$0.269E-02$	$0.641E-05$	$0.138E-04$	$0.233E-02$
$M = \sqrt{N} = 65$						
0	0	20	$0.000E+00$	$0.339E-01$	$0.127E-01$	$0.126E+00$
1	142	599	$0.104E+00$	$0.178E-02$	$0.205E-02$	$0.347E-01$
2	168	712	$0.279E-01$	$0.641E-04$	$0.286E-03$	$0.109E-01$
3	257	1043	$0.876E-02$	$0.619E-05$	$0.274E-04$	$0.393E-02$
4	85	339	$0.205E-02$	$0.499E-05$	$0.533E-05$	$0.206E-02$
5	73	287	$0.138E-02$	$0.488E-05$	$0.216E-05$	$0.926E-03$
$M = \sqrt{N} = 129$						
0	0	82	$0.000E+00$	$0.338E-01$	$0.636E-02$	$0.126E+00$
1	360	7412	$0.104E+00$	$0.158E-02$	$0.961E-03$	$0.322E-01$
2	481	9423	$0.268E-01$	$0.517E-04$	$0.126E-03$	$0.848E-02$
3	1064	19609	$0.910E-02$	$0.586E-05$	$0.121E-04$	$0.341E-02$
4	116	2095	$0.328E-02$	$0.482E-05$	$0.116E-05$	$0.435E-03$
5	73	1327	$0.782E-03$	$0.480E-05$	$0.369E-06$	$0.646E-03$
$M = \sqrt{N} = 257$						
0	0	343	$0.000E+00$	$0.338E-01$	$0.318E-02$	$0.126E+00$
1	894	84506	$0.105E+00$	$0.148E-02$	$0.463E-03$	$0.310E-01$
2	1266	112849	$0.263E-01$	$0.461E-04$	$0.590E-04$	$0.763E-02$
3	4869	413652	$0.678E-02$	$0.534E-05$	$0.571E-05$	$0.198E-02$
4	671	54504	$0.155E-02$	$0.517E-05$	$0.503E-05$	$0.768E-03$
5	76	6384	$0.424E-03$	$0.485E-05$	$0.168E-05$	$0.670E-03$

The convergence behavior of Newton's method is illustrated in Table 4.1. The discretization parameters M and N are increased simultaneously, where five levels and five iterations per level are considered. The first column gives the number of the nonlinear iterations. The second gives the required number of iterations of the SYMMLQ algorithm. Third, the required time, measured in seconds, for all iterations is specified. Right next to it, the $\mathcal{L}^2(I)$-norm of the step, the value of the cost function, the $\mathcal{L}^2(\Omega)$-norm of the gradient, and the $\mathcal{L}^2(I)$-norm of the control error $u_i - u_*$ are tabulated. Attention is drawn to the fact that u_i is the finite ith iteration, whereas u_* is the solution of the infinite dimensional problem without regularization. On account of this, q-quadratic convergence rate of these values cannot be expected.

The number of SYMMLQ-iterations increases to about two to three times that of the previous coarse level. In other words, the condition number of the Hessian matrix increases with the parameter M, which means that more variational equations must be solved. A preconditioner for Newton's equation could improve this behavior.

The convergence behavior is also improved for each increase in the discretization

Table 4.2
Nested iteration of Newton's method ($u \geq 0.3$, $u_0 = 1.8$).

IT	TIME$_\Sigma$	$\|w^M\|$	$F_N(u_i^M)$	$\|F_N'(u_i^M)\|$	$\|u_i^M - u_*\|$
M=9, $\varepsilon = 10^{-2}$ (Projected gradient method)					
0	0	$0.000E+00$	$0.406E+01$	$0.784E+00$	$0.825E+00$
10	5	$0.250E-01$	$0.521E-02$	$0.304E-02$	$0.605E-01$
20	10	$0.115E-02$	$0.483E-02$	$0.200E-03$	$0.323E-01$
30	14	$0.173E-03$	$0.483E-02$	$0.311E-04$	$0.289E-01$
40	19	$0.306E-04$	$0.483E-02$	$0.570E-05$	$0.285E-01$
50	23	$0.241E-05$	$0.483E-02$	$0.434E-06$	$0.285E-01$
60	27	$0.444E-06$	$0.483E-02$	$0.832E-07$	$0.285E-01$
69	31	$0.465E-07$	$0.483E-02$	$0.846E-08$	$0.285E-01$
M=17, $\varepsilon = 10^{-3}$ (Newton's method)					
0	32	$0.000E+00$	$0.646E-03$	$0.139E-02$	$0.285E-01$
1	36	$0.256E-01$	$0.504E-03$	$0.220E-03$	$0.795E-02$
2	39	$0.473E-02$	$0.492E-03$	$0.487E-04$	$0.731E-02$
3	42	$0.181E-02$	$0.491E-03$	$0.186E-04$	$0.766E-02$
4	45	$0.445E-03$	$0.490E-03$	$0.844E-05$	$0.779E-02$
5	48	$0.117E-03$	$0.490E-03$	$0.449E-05$	$0.782E-02$
M=33, $\varepsilon = 10^{-4}$ (Newton's method)					
0	53	$0.000E+00$	$0.547E-04$	$0.144E-03$	$0.782E-02$
1	82	$0.607E-02$	$0.491E-04$	$0.872E-05$	$0.221E-02$
2	112	$0.580E-03$	$0.489E-04$	$0.477E-05$	$0.241E-02$
3	141	$0.125E-03$	$0.489E-04$	$0.282E-05$	$0.242E-02$
4	171	$0.579E-04$	$0.489E-04$	$0.172E-05$	$0.242E-02$
5	197	$0.315E-04$	$0.489E-04$	$0.108E-05$	$0.242E-02$
M=65, $\varepsilon = 10^{-5}$ (Newton's method)					
0	216	$0.000E+00$	$0.514E-05$	$0.172E-04$	$0.242E-02$
1	502	$0.180E-02$	$0.487E-05$	$0.514E-06$	$0.732E-03$
2	762	$0.130E-03$	$0.486E-05$	$0.249E-06$	$0.806E-03$
3	1002	$0.294E-04$	$0.486E-05$	$0.129E-06$	$0.811E-03$
4	1198	$0.126E-04$	$0.486E-05$	$0.733E-07$	$0.812E-03$
5	1353	$0.587E-05$	$0.486E-05$	$0.459E-07$	$0.812E-03$
10	1697	$0.305E-06$	$0.486E-05$	$0.114E-07$	$0.812E-03$
11	1753	$0.216E-06$	$0.486E-05$	$0.933E-08$	$0.812E-03$

parameters. However, the discretization error apparently influences the convergence rate. Although the q-quadratic convergence rate in the infinite dimensional setting cannot be directly observed, the fast reduction of the gradient norm and the cost function within the first iterations points to such a property. After the first iterations, the convergence rate is dominated by the large discretization error.

The analyzed behavior of Newton's method for increasing parameters M and N naturally leads to the idea of nested iteration. Since Newton's iteration on a coarse grid requires less time than on a fine grid, it could be expected that this yields a more efficient algorithm. Important issues in this context are how to control the increasing parameters and how to choose the ratio between the parameters M and N. Within the limitation of this report we deal with this problem only in passing.

The results of Table 4.2 are obtained by using four levels, $M = 9, 17, 33, 65$, with $N = M^2$, and each approximation is transferred to the finer grid by linear interpolation. The regularization parameter is decreased with respect to the different levels in order to handle the ill-posedness of the problem ($\varepsilon = 10^{-j}$, $j = 2, \ldots, 5$). Now, the accumulated time is listed in the second column.

To illustrate the applicability of the derived algorithm the starting approximation is chosen to be $u_0 = 1.8$, which is far away from the solution u_*. A further constraint

$u \geq 0.3$ has to be added since otherwise the control would become negative, which makes no sense for a domain. For simplicity, the projected gradient method with Armijo rule [10] is implemented on the coarse grid, where the stopping criterion is kept.

After the iterations on the coarse grid, the approximation is interpolated to the finer grid to carry out five Newton iterations. Then, the approximation is interpolated to the next finer grid and once again improved by five Newton iterations. This process proceeds until the finest grid ($M = 65$) is reached in order to compare the results with the presented method without nested iteration.

Although the starting point is further away from the solution this nested iteration method is faster than Newton's method on the fine grid. The bound 10^{-3} of the control is satisfied after the first iteration on the finest grid and the computation requires only about eight minutes. This has to be compared to 50 minutes for Newton's method without nested iteration, which is approximately required for computing a competitive approximation.

Finally, it can be observed that the control error at each level is unchanged at the last few iterations. Therefore, the refinement strategy could even be improved by using a modified stopping criterion at each level. Since the issue of choosing an appropriate stopping criterion is well discussed for various methods (see, e.g., [2], [23], [29]), we will not deal further with this topic.

To sum up, the computational experiments give evidence for the convergence assertions of the derived modified mesh independence theorem. It has been further illustrated that one key to constructing more efficient methods is to exploit the infinite dimensional information and to use a mesh refinement strategy.

REFERENCES

[1] G. Allaire and R. V. Kohn, *Optimal design for minimum weight and compliance in plane stress using extremal microstructures*, European J. Mech. A Solids, 12 (1993), pp. 839–878.

[2] E. L. Allgower and K. Böhmer, *Application of the mesh independence principle to mesh refinement strategies*, SIAM J. Numer. Anal., 24 (1987), pp. 1335–1351.

[3] E. L. Allgower, K. Böhmer, F.-A. Potra, and W. C. Rheinboldt, *A mesh-independence principle for operator equations and their discretizations*, SIAM J. Numer. Anal., 23 (1986), pp. 160–169.

[4] O. Axelsson and V. A. Barker, *Finite Element Solution of Boundary Value Problems*, Academic Press, Orlando, FL, 1984.

[5] H. T. Banks and F. Kojima, *Boundary identification for 2-d parabolic problems arising in thermal testing of materials*, in Proc. of the 27th Conference on Decision and Control, Austin, TX, 1988, pp. 1678–1683.

[6] H. T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Systems & Control: Foundations & Applications. Birkhäuser–Verlag, Boston, 1989.

[7] J. Baumeister, *Stable Solution of Inverse Problems*, Advanced Lectures in Mathematics, Friedr. Vieweg & Sohn, Braunschweig, 1987.

[8] D. Begis and R. Glowinski, *Application de la méthode des éléments finis à l'approximation d'un problème de domain optimal*, Appl. Math. Comput., 2 (1975), pp. 130–169.

[9] M. Bendsoe and C. Mota-Soares, eds., *Topology Design of Structure*, Kluwer, Amsterdam, 1992.

[10] D. P. Bertsekas, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Trans. Automat. Control, 21 (1976), pp. 174–184.

[11] E. Bonnetier and C. Conca, *Approximation of Young measures by functions and application to a problem of optimal design for plates with variable thickness*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 399–422.

[12] D. Chenais, *Discrete gradient and discretized continuum gradient method for shape optimization of shells*, Mech. Structures Mach., 22 (1994), pp. 73–115.

[13] P. G. Ciarlet, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Vol. II, P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, 1991, pp. 17–351.

[14] M. Delfour, G. Payre, and J.-P. Zolésio, *Optimal design of a minimum weight thermal diffuser with constraint on the output thermal power flux*, Appl. Math. Optim., 9 (1983), pp. 225–262.

[15] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[16] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Nonlinear Equations and Unconstrained Optimization*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[17] W. Hackbusch, *Multi-Grid Methods and Applications.* Springer-Verlag, Berlin, 1985.

[18] W. Hackbusch, *Integralgleichungen*, Teubner Studienbücher Mathematik, B.G. Teubner, Stuttgart, 1989.

[19] J. Haslinger, K.-H. Hoffmann, and M. Kočvara, *Control/fictitious domain method for solving optimal shape design problems*, RAIRO Modél. Math. Anal. Numér., 27 (1993), pp. 157–182.

[20] J. Haslinger and P. Neittaanmäki, *Finite Element Approximation for Optimal Shape Design: Theory and Applications*, John Wiley, New York, 1988.

[21] M. Heinkenschloss, *Gauss-Newton Methods for Infinite Dimensional Least Squares Problems with Norm Constraints*, Ph.D. thesis, Fb IV - Mathematik, Universität Trier, Germany, 1991.

[22] M. Heinkenschloss, *Mesh independence for nonlinear least squares problems with norm constraints*, SIAM J. Optim., 3 (1993), pp. 81–117.

[23] M. Heinkenschloss, M. Laumen, and E. W. Sachs, *Gauss-Newton methods with grid refinement*, in Estimation and Control of Distributed Parameter Systems, Internat. Ser. Numer. Math. 100, W. Desch, F. Kappel, and K. Kunisch, eds., Birkhäuser, Basel, 1991, pp. 161–174.

[24] L. W. Kantorowitsch and G. P. Akilow, *Funktionalanalysis in normierten Räumen*, Verlag Harri Deutsch, Thun-Frankfurt am Main, 1978.

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

[26] C. T. Kelley and E. W. Sachs, *Mesh independence of Newton-like methods for infinite-dimensional problems*, J. Integral Equations Appl., 3 (1991), pp. 549–573.

[27] C. T. Kelley and E. W. Sachs, *Broyden's method for approximate solution of nonlinear integral equations*, J. Integral Equations Appl., 9 (1985), pp. 25–44.

[28] C. T. Kelley and E. W. Sachs, *Quasi-Newton methods and unconstrained optimal control problems*, SIAM J. Control Optim., 25 (1987), pp. 1503–1516.

[29] C. T. Kelley and E. W. Sachs, *Approximate quasi-Newton methods*, Math. Programming, 48 (1990), pp. 41–70.

[30] M. Laumen, *A comparison of numerical methods for optimal shape design problems*, Optim. Methods Softw., to appear.

[31] M. Laumen, *Newton's method for a class of optimal shape design problems*, SIAM J. Optim., to appear.

[32] M. Laumen, *Numerical Methods for Optimal Shape Design Problems*, Ph.D. thesis, FB IV-Mathematik, Universität Trier, Germany, 1996.

[33] F. Murat and L. Tartar, *Calcul des variations et homogénéisation*, in Les méthodes de l'homogénéisation: Théorie et applications en physique, D. Bergman et al., eds., Collection Dir. Etudes et Rech. Elec. France 57, Eyrolles, Paris, pp. 319–369.

[34] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.

[35] O. Pironneau, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, New York, 1984.

[36] O. Pironneau and A. Vossinis, *Comparison of some optimization algorithms for optimum shape design in aerodynamics*, Programme 6: Calcul scientifique, Modélisation et Logiciels numériques 1392, INRIA, 1991.

[37] W. C. Rheinboldt, *An adaptive continuation process for solving systems of nonlinear equations*, in Mathematical Models and Numerical Methods, Banach Center Publ. 3, PWN, Warsaw, 1978, pp. 129–142.

[38] J. Sokolowski and J. P. Zolesio, *Introduction to Shape Optimization: Shape Sensitivity Analysis*, Springer-Verlag, Berlin, 1992.

[39] F. Tröltzsch, *Optimality Conditions for Parabolic Control Problems and Applications*, Teubner, Leipzig, 1984.

[40] E. Zeidler, *Nonlinear Functional Analysis and its Applications* I: *Fixed-Point Theorems*, Springer-Verlag, New York, 1986.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 4, pp. 1089–1122

ABSOLUTE STABILIZATION AND MINIMAX OPTIMAL CONTROL OF UNCERTAIN SYSTEMS WITH STOCHASTIC UNCERTAINTY*

V. A. UGRINOVSKII† AND I. R. PETERSEN†

Abstract. This paper is concerned with existence and optimality properties of so-called guaranteed cost controllers for an uncertain system subject to structured uncertainty. The uncertainty in the system is assumed to have a stochastic character and to satisfy certain stochastic integral constraints. It is shown that a minimax optimal guaranteed cost state feedback controller for a stochastic system can be synthesized as a state feedback controller absolutely stabilizing this system. For each initial state of the system, this controller can be found by parametric optimization of solutions of a parameter-dependent generalized matrix Riccati equation arising in stochastic H_∞ theory.

Key words. robust control, robust performance, structured uncertainty, stochastic H_∞ control, stochastic absolute stability, multiplicative noise

AMS subject classifications. 93E20, 93E15, 93D21, 93D22, 49J35, 60G35

PII. S0363012996309964

1. Introduction.

Recently, Petersen and James [16] introduced a new general framework to allow for both stochastic and deterministic uncertainty in a *discrete*-time system. The motivation for this has been to combine two alternative uncertainty models used commonly in control theory and applications. The one model considers uncertainty arising from unmodelled dynamics, modelling and linearization errors, and slow parameter deviations which are deterministic in nature. Another approach involves the use of stochastic processes to model dynamics driven by noise signals and uncertainties due to fast parameters variations. This framework involved defining a new class of stochastic uncertain systems in which the uncertainty was described by a *stochastic uncertainty constraint.*

This paper addresses the problem of developing a similar framework for *continuous*-time systems. We introduce a rather general model of stochastic uncertainty which naturally extends the deterministic structured uncertainty model (cf. [12, 15, 19, 20, 21, 22, 28]), allowing for stochastic perturbations. Thus, our uncertainty model is a continuous-time counterpart of the discrete-time stochastic uncertainty models introduced in [16]. However, we should emphasize the difference between the results of this paper and those of [16]. The results in [16] concern uncertain systems with additive noise perturbations. This paper deals with a multiplicative noise perturbation model. Structured stochastic uncertainties with multiplicative noise perturbations arise naturally in many control problems and are considered extensively in the literature (cf. [5, 7, 8, 9, 25, 30]). In what follows, we describe two examples that motivate the class of problems considered in this paper.

Technically and in its content, this paper is closely related to the paper of Savkin and Petersen [20]. The main problem addressed is finding a linear static state feedback controller yielding a prescribed level of performance in the face of stochastic structured uncertainty in the system. In [20], such a controller is found as a minimax optimal

*Received by the editors September 29, 1996; accepted for publication (in revised form) June 8, 1998; published electronically April 27, 1999. The research was supported by the Australian Research Council.

http://www.siam.org/journals/sicon/37-4/30996.html

†Department of Electrical Engineering, Australian Defence Force Academy, Canberra ACT 2600, Australia (valu@ee.adfa.oz.au, irp@ee.adfa.oz.au).

controller which minimizes the maximum (over all admissible uncertainties) value of a cost functional. The class of controllers considered are static state feedback controllers, which absolutely stabilize the system. However, the key point of the controller proposed here is that its construction is based on the stabilizing solution of a *generalized* Riccati equation related to a stochastic H_∞ control problem and to a stochastic differential game considered recently in [26]. Our set up has required the results of [26] to be slightly modified. Therefore, we present these modifications in sections 2 and 3. Then, in section 4 we consider the auxiliary problem of stochastic absolute stabilizability of a system whose uncertainty satisfies a certain stochastic integral quadratic constraint. The main result is given in section 5. It establishes that the search for an minimax optimal controller can be reduced to a finite-dimensional optimization problem over the solutions to a generalized Riccati equation.

One more important feature which distinguishes stochastic problems from deterministic ones is as follows. The stochastic Ito equations which serve to give a mathematically rigorous description for system dynamics are nonautonomous equations in their nature. Technically, this is reflected in the fact that we have to take into account the difference between the effect of initial conditions posed at different times; i.e., roughly speaking, we cannot reduce a problem with initial condition $x(s) = h$, $s > 0$, to a problem with an initial condition imposed of time zero. That is why all systems and processes in this paper are considered with respect to the starting instant s, not 0. In particular, this allows us to readily integrate problems with random initial conditions [17] into our framework.

Notation. We use the notation $\mathbf{R}^n$, $\mathbf{R}^{n\times q}$ to denote the n-dimensional real Euclidian vector space and the space of real $n\times q$-matrices equipped with the Euclidian matrix norm. We shall use symbols $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ to denote, respectively, the norm of vectors and matrices and the inner product of vectors. Furthermore, given a positive symmetrical matrix $Q \in \mathbf{R}^{q\times q}$, $\mathrm{tr}\Theta_1 Q\Theta_2'$ defines an inner product on $\mathbf{R}^{n\times q}$. Hence, this space can be considered a subspace in the Hilbert space of Hilbert–Schmidt operators.

Let $\{\Omega, \mathcal{F}, \mathcal{P}\}$ be a complete probability space, and let $w_1(t)$, $w_2(t)$ be two mutually independent Wiener processes in $\mathbf{R}^{q_1}$, $\mathbf{R}^{q_2}$ with covariance matrices Q_1, Q_2, respectively. Let $\mathcal{F}_t$ denote the increasing sequence of Borel sub-σ-fields of $\mathcal{F}$, generated by $\{w_{1,2}(s), 0 \le s < t\}$. Also, let $\mathbf{E}$ and $\mathbf{E}\{\cdot|\mathcal{F}_t\}$ be the corresponding unconditional and conditional expectation operators, respectively, where the latter is the expectation with respect to $\mathcal{F}_t$.

Let L_2^s denote the Hilbert space $L_2(\Omega, \mathcal{F}_s, \mathcal{P}; \mathbf{R}^n)$ of $\mathcal{F}_s$-measurable random variables $\Omega \to \mathbf{R}^n$, which is complete with respect to the norm $\left(\mathbf{E}\|\cdot\|^2\right)^{1/2}$. For $T \le \infty$, let $L_2(s, T; \mathbf{R}^n)$ denote the Hilbert space generated by the (t,ω)-measurable $\{\mathcal{F}_t, t \ge 0\}$-nonanticipating processes $x(t,\omega)\colon [s,T] \times \Omega \to \mathbf{R}^n$ and complete with respect to the norm $|||\cdot||| = (\int_s^T \mathbf{E}\|\cdot\|^2 dt)^{1/2}$. We shall write $L_2(s; \mathbf{R}^n)$ for $L_2(s, +\infty; \mathbf{R}^n)$.

Given a symmetric positive definite $q \times q$-matrix Q, let $\mathbf{R}_Q^{n\times q}$ denote the Hilbert space of $n \times q$-matrices, with the inner product $\mathrm{tr}\Theta_1 Q\Theta_2'$.

We consider an uncertain stochastic system described by the following stochastic differential Ito equation:

$$\begin{aligned}(1)\quad & dx = (Ax(t) + B_1u(t) + B_2\xi(t))dt + (Hx(t) + P_1u(t))dw_1(t) + P_2\xi(t)dw_2(t),\\ & z(t) = Cx(t) + Du(t),\end{aligned}$$

where $x(t) \in \mathbf{R}^n$ is the state, $u(t) \in \mathbf{R}^{m_1}$ is the control input, $z(t) \in \mathbf{R}^p$ is a vector assembling all uncertainty outputs, and $\xi(t) \in \mathbf{R}^{m_2}$ is a vector assembling all uncertainty inputs. Here A, B_1, B_2, C, D are matrices of corresponding dimensions, and H, P_1, P_2 are linear bounded operators $\mathbf{R}^n \to \mathbf{R}_{Q_1}^{n\times q_1}$, $\mathbf{R}^{m_1} \to \mathbf{R}_{Q_1}^{n\times q_1}$, $\mathbf{R}^{m_2} \to \mathbf{R}_{Q_2}^{n\times q_2}$, respectively. In the sequel, we shall use the adjoint operators H^*, P_1^*, P_2^* defined by the following:

$$\begin{aligned}&\langle H^*\Theta_1, x\rangle = \mathrm{tr}\Theta_1 Q_1 (Hx)', \quad \langle P_1^*\Theta_1, u\rangle = \mathrm{tr}\Theta_1 Q_1 (P_1 u)',\\ &\langle P_2^*\Theta_2, \xi\rangle = \mathrm{tr}\Theta_2 Q_2 (P_2\xi)'\\ &\forall\, x\in\mathbf{R}^n,\ u\in\mathbf{R}^{m_1},\ \xi\in\mathbf{R}^{m_2},\ \Theta_1\in\mathbf{R}^{n\times q_1},\ \Theta_2\in\mathbf{R}^{n\times q_2}.\end{aligned}$$

1.1. System uncertainty. The uncertainty in the above system (1) is described by the equation

$$\xi(t) = \xi_\phi(t) := \phi(t, x(\cdot)|_0^t, u(\cdot)|_0^t). \tag{2}$$

We suppose this uncertainty to satisfy the following stochastic integral quadratic constraint.

Definition 1. *Let $\bar{R}\ge 0$, $\bar{G} > 0$, $W > 0$ be given matrices. Then an uncertainty of the form* (2) *is said to be admissible if the following conditions hold.*

1. *For any $s \ge 0$, if $u(\cdot) \in L_2(s, T; R^{m_1})$, then there exists a unique solution to* (1), (2) *that lies in $L_2(s, T; R^n)$;*

2. *There exists a sequence $\{t_j\}_{j=1}^\infty$ such that $t_j > s$, $t_j \to \infty$ as $j \to \infty$ and the following condition holds. If $u(\cdot) \in L_2([s, t_j]; R^{m_1})$ and $x(\cdot) \in L_2([s, t_j]; R^n)$, then $\xi_\phi(\cdot) \in L_2([s, t_j]; R^{m_2})$ and*

$$\int_s^{t_j} \mathbf{E}\left(\langle z(t), \bar{R}z(t)\rangle - \langle \xi_\phi(t), \bar{G}\xi_\phi(t)\rangle\right) dt \ge -\mathbf{E}\langle h, Wh\rangle, \quad h = x(s). \tag{3}$$

We use the notation $\Phi(\bar{R}, \bar{G}, W)$ to denote the set of admissible uncertainties. However, we will write Φ wherever it produces no confusion.

Observe that the trivial uncertainty $\phi \equiv 0$ satisfies the above constraint. In the sequel, we shall refer to the system corresponding to this uncertainty as the *nominal* system.

In a typical situation, the plant may contain several uncertain feedback loops. In our notation, this is described by the decomposition of uncertainty output vector z and uncertainty input vector ξ_ϕ into several blocks of reduced dimensions as follows:

$$z = [z_1', \dots, z_k']', \quad \xi_\phi = [\xi_{\phi,1}', \dots, \xi_{\phi,k}']'.$$

This in turn induces a corresponding block decomposition on the matrices C, D, B_2, and P_2 in (1). The version of Definition 1 that accounts for this structure of uncertainty proceeds from the assumption that each uncertainty loop satisfies its own stochastic integral quadratic constraint of the form of (3):

$$\int_s^{t_j} \mathbf{E}\left(\langle z_i(t), \bar{R}_i z_i(t)\rangle - \langle \xi_{\phi,i}(t), \bar{G}_i\xi_{\phi,i}(t)\rangle\right) dt \ge -\mathbf{E}\langle h, W_i h\rangle, \quad i = 1, \dots, k, \tag{4}$$

where $\bar{R}_i \ge 0$, $\bar{G}_i > 0$ and $W_i > 0$. Then for any numbers $\tau_1 > 0, \dots, \tau_k > 0$, we replace all of the constraints by a single stochastic integral quadratic constraint in the form of (3):

$$\int_s^{t_j} \mathbf{E}\left(\langle z(t), \bar{R}_\tau z(t)\rangle - \langle \xi_\phi(t), \bar{G}_\tau\xi_\phi(t)\rangle\right) dt \ge -\mathbf{E}\langle h, W_\tau h\rangle, \quad h = x(s), \tag{5}$$

where $\bar{R}_\tau$, $\bar{G}_\tau$ are the block diagonal matrices

$$\bar{R}_\tau = \mathrm{diag}[\tau_1\bar{R}_1, \ldots, \tau_k\bar{R}_k], \quad \bar{G}_\tau = \mathrm{diag}[\tau_1\bar{G}_1, \ldots, \tau_k\bar{G}_k],$$

and $W_\tau = \sum_{i=1}^k \tau_i W_i$.

A specific feature of the uncertainty description in the form of the integral quadratic constraint (3) or the structured integral quadratic constraints (4) is that this description employs a certain sequence $\{t_j\}_{j=1}^\infty$ of times. Obviously, in the particular case where $u(\cdot)$ is a *stabilizing* control input which guarantees that the uncertainty input $\xi(\cdot)$ of the form (2) and the solution $x(\cdot)$ to the system (1), driven by the control input $u(\cdot)$ and uncertainty input $\xi(\cdot)$, both exist on $[0,\infty)$ and belong to $L_2[0,\infty)$, there is no need in employing the sequence $\{t_j\}_{j=1}^\infty$ to describe the uncertainty. Indeed, given the constraint in the form (3) and stabilizing control input $u(\cdot)$, by passing to the limit in (3) as $t_j \to \infty$, one can replace the integral over $[0,t_j]$ in (3) by the integral over the infinite interval $[0,\infty)$. Furthermore, by then making use of the Parseval identity, one may proceed to consider the frequency domain version of the integral quadratic constraint (3); cf. [15]. However, at this stage, we have yet not determined that any control input $u(\cdot)$ is stabilizing. We wish to define the class of admissible uncertainties for a generic control input; therefore we ought to avoid referring to any particular stabilizing properties of the control input when the constraints on the uncertainty are being defined. As in the deterministic case (cf. [20, 21, 22]), this can be achieved by considering control inputs, uncertainty inputs, and the corresponding solutions defined on a sequence of expanding finite intervals $[0,t_j]$. To give the reader an idea on how conservative the uncertainty model described by Definition 1 is, note that if for a certain uncertainty input ξ_ϕ, there exists no sequence $\{t_j\}$ such that the constraint (3) is satisfied for suitable $\bar{R}$, $\bar{G}$, and $\bar{W}$, then this means that the considered uncertainty input is not locally (and hence globally) square integrable.

The uncertainty constraint given by (3) or (4) extends the integral quadratic constraints such as those given in [20, 21, 22] to stochastic systems with multiplicative noise. As in these references, this uncertainty description allows for the uncertainty input ξ to depend dynamically on the uncertainty outputs. Also, a constraint in the form of (3) or (4) represents an extension of the discrete stochastic sum constraint of [16] to the case of continuous-time stochastic systems. However, the results of [16] allow for additive noise rather than multiplicative noise. Note also that the uncertainty description in the form of the constraint (3) encompasses the standard norm-bounded uncertainty description. Indeed, if

$$\xi(t) = \Delta(t)z, \quad \|\Delta(t)\| \le 1,$$

i.e., $\phi(t,x,u) = \Delta(t)(Cx + Du)$, then the constraint (3) is satisfied with $\bar{R} = I$, $\bar{G} = I$, and any matrix $W > 0$ and sequence $\{t_j\}_{j=1}^\infty$, provided that $x(\cdot)$ and $u(\cdot)$ lie in the corresponding L_2 spaces. A corresponding observation is also true in the case of structured norm-bounded uncertainty.

Stochastic extensions to integral quadratic constraints may provide a possible approach to the problem of nonworst-case robust control design. Recall that the standard deterministic worst-case robust control design presumes that all uncertainties have an equal chance of occurring, so that one does not expect that certain uncertainty inputs are more or less likely than others. Although the worst-case design methodology has proved its efficacy in various engineering problems, it suffers from the disadvantage that the designer lacks the opportunity to discriminate between "expected" uncertainties and those uncertainties which are known to seldom occur. In

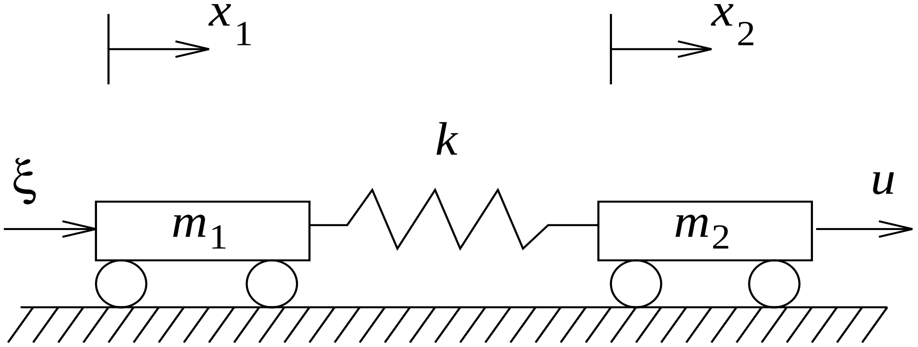


Fig. 1. *The two carts system.*

other words, the standard worst-case design methodology proceeds from the assumption that the values that the uncertainty may take, are equally likely; i.e., one can think of the uncertainty arising from a uniformly distributed random variable taking its values in the space of uncertainty inputs. However, this may not accurately represent the uncertainty in the system under consideration. For example, it may have been determined that the uncertainty inputs have a distribution other than the uniform distribution, and the designer may wish to make use of as much realistic a priori information about the distribution of uncertainty values as possible. The following example illustrates a situation in which this idea can be applied to go beyond the worst-case robust control design in the case where the distribution of uncertainties is Gaussian. In this case, we arrive at an underlying system of the form (1).

Suppose that the system to be controlled consists of two carts connected by a spring as shown in Figure 1. There is a disturbance ξ of the form (2) acting on the first cart. A control force u drives the second cart. The spring constant k has a specific nominal value $k_0 = 1.25$ but can vary and is considered uncertain. This may reflect the nonlinear nature of the true spring. A series of experiments was undertaken to determine values of the spring constant in various conditions. It was revealed that for each time instant t, the histogram of observed values is consistent with a stationary Gaussian distribution, with the mean k_0. It was also found in the experiments that k ranged over the interval $[0.5, 2]$. Assume that the masses of carts are $m_1 = m_2 = 1$. Then, the system is described by the equation

$$\dot{x} = (A + F\Delta(t)C)x + B_1 u + B_2 \xi,$$

where $x = [x_1\ x_2\ \dot{x}_1\ \dot{x}_2]' \in \mathbf{R}^4$, $\Delta(t) = k(t) - k_0$, and

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1.25 & 1.25 & 0 & 0 \\ 1.25 & -1.25 & 0 & 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix},$$

$$F = \begin{bmatrix} 0 \\ 0 \\ -1 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & -1 & 0 & 0 \end{bmatrix}. \tag{6}$$

Assume that for all $t \geq 0$, $\Delta(t)$ is the Gaussian white noise process with zero

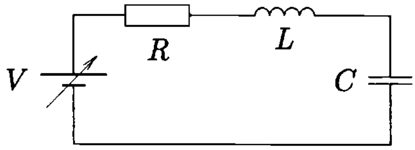


FIG. 2. *The uncertain electric circuit.*

mean and $E\Delta^2(t) = \sigma^2$. We can then choose the value of the parameter σ such that $k(t) = k_0 + \Delta(t)$ obeys the bounds $0.5 \leq k(t) \leq 2$ with a sufficiently high probability. For example, for $\sigma = 0.25$, we have $P(|k(t) - k_0| \leq 0.75) \geq 0.997$. This model of spring rate variations leads us to the uncertain stochastic system[1] of the form (1),

$$\begin{aligned} dx &= (Ax + B_1 u + B_2 \xi)dt + Hx dw_1(t), \\ z &= Cx + Du, \end{aligned}$$

where $H = \sigma FC$, and $w_1(\cdot)$ is the scalar Wiener process. Note that in this model, the probability $P(|k(t) - k_0| \leq 0.75)$ is increasing as $\sigma^2 \downarrow 0$. However, we shall always have $P(|k(t) - k_0| \leq 0.75) < 1$; i.e., the value of the spring rate $k(t)$ may exceed the presumed bounds on uncertainty $[0.5, 2]$ with nonzero probability. This phenomenon indicates the "soft" norm bound on the uncertainty.

Another example showing how the above uncertainty description may arise is given by the electric circuit shown in Figure 2. The differential equations describing the current and voltage dynamics in this circuit are the following:

$$\begin{aligned} \frac{di}{dt} &= -\frac{R}{L} i + \frac{1}{L}(V - V_C), \\ \frac{dV_C}{dt} &= \frac{1}{C} i, \end{aligned}$$

where i is the current flowing in the circuit and V_C is the capacitor voltage. One can control the system by applying the appropriate voltage V. In the circuit, the resistance R of the resistor and the inductance L of the inductor may vary as follows.

It is known that the resistance R may slowly vary from R^- to R^+ due to, e.g., the resistor temperature variations. That is, $R = R(t) = R_0 + R_1 \Delta R(t)$, where $R_0 = (R^+ + R^-)/2$ is the nominal value of the resistor, $R_1 = (R^+ - R^-)/2$, and $\Delta R(t)$ satisfies the standard norm-bounded uncertainty constraint, $|\Delta R(t)| \leq 1$. Also, it is known that nearby electrical devices may induce changes in the inductance value. That is, one may suppose that the inductance $L = L(t)$ is given by $L^{-1} = L_0^{-1} + L_1^{-1}\zeta(t)$, where $\zeta(t)$ is the Gaussian white noise with the mean 0 and the covariance 1, L_0 is the mean inductance. The value of L_1 can be determined by estimating the covariance of the reciprocal inductance.

Letting $x = [i, V_C] \in \mathbf{R}^2$, $z_1 = i$, $z_2 = V_C - V$, one can write a formal Langevin

[1] In this example, one relies on the fact that $(FC)^2 = 0$ when proceeding from the formal Langevin equation to a corresponding mathematically rigorous Ito equation. Conditions under which a solution of the Langevin equation satisfies also a corresponding Stratonovich equation and, consequently, an Ito equation are given in [24].

equation for the system in Figure 2:

$$\begin{aligned}
\dot{x} &= \begin{bmatrix} -R_0/L_0 & -1/L_0 \\ -1/C & 0 \end{bmatrix} x + \begin{bmatrix} 1/L_0 \\ 0 \end{bmatrix} V + \begin{bmatrix} -R_1/L_0 \\ 0 \end{bmatrix} \xi_1 \\
&\quad + \left(\begin{bmatrix} -R_0/L_1 & -1/L_1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 1/L_1 \\ 0 \end{bmatrix} V + \begin{bmatrix} -R_1/L_1 \\ 0 \end{bmatrix} \xi_1 \right) \zeta(t), \\
z &= x + \begin{bmatrix} 0 \\ -1 \end{bmatrix} V, \qquad \xi_1 = \Delta R(t) z_1.
\end{aligned}$$

The rigorous mathematical description of the system is then given by the following Ito stochastic differential equation of the form (1):

$$\begin{aligned}
dx &= \left(\begin{bmatrix} -\frac{R_0}{L_0} + \frac{R_0^2}{2L_1^2} & -\frac{1}{L_0} + \frac{R_0}{2L_1^2} \\ -1/C & 0 \end{bmatrix} x + \begin{bmatrix} \frac{1}{L_0} - \frac{R_0}{2L_1^2} \\ 0 \end{bmatrix} V \right. \\
&\quad \left. + \begin{bmatrix} -\frac{R_1}{L_0} + \frac{R_0 R_1}{L_1^2} & \frac{R_1}{2L_1^2} & \frac{R_1^2}{2L_1^2} \\ 0 & 0 & 0 \end{bmatrix} \xi \right) dt \\
&\quad + \left(\begin{bmatrix} -R_0/L_1 & -1/L_1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 1/L_1 \\ 0 \end{bmatrix} V + \begin{bmatrix} -R_1/L_1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \xi \right) dw(t), \\
z &= x + \begin{bmatrix} 0 \\ -1 \end{bmatrix} V, \qquad \xi = \Delta(t) z, \qquad \Delta(t) = \begin{bmatrix} \Delta R(t) & 0 \\ 0 & \Delta R(t) \\ (\Delta R(t))^2 & 0 \end{bmatrix}.
\end{aligned}$$

It is easy to see that $\Delta'\Delta \leq 2I$. As we have already mentioned, this constraint can be converted into a certain integral quadratic constraint of the form (3). It is also worth noting that in this example, the uncertainty is structured. Hence, the integral quadratic constraints on structured uncertainty of the form (4) can be used to describe the uncertainty in this example. It is known that the use of the structured integral quadratic constraints uncertainty description may lead to a potentially less conservative guaranteed robust performance. This motivates us to consider an optimal guaranteed cost control problem in section 5 for the system (1) with structured uncertainty.

1.2. Guaranteed cost control problem.

As has been mentioned above, the main problem addressed in this paper is to find a linear static state feedback controller resulting in an optimal performance in the face of stochastic uncertainty in the system (1). In this section, we set up this problem.

Let $R \in \mathbf{R}^{n \times n}$, $G \in \mathbf{R}^{m_1 \times m_1}$, $R' = R > 0$, $G' = G > 0$, be given matrices. Associated with the uncertain system (1), (2), consider the cost functional,

$$J^{s,h}(u(\cdot)) = \int_s^\infty \mathbf{E} \left(\langle x(t), Rx(t) \rangle + \langle u(t), Gu(t) \rangle \right) dt, \tag{7}$$

where $x(t)$ is the solution to (1), (2) satisfying the initial condition $x(s) = h$.

Definition 2. *Given a constant* $\gamma > 0$ *and the cost functional* (7), *the state feedback controller*

$$u^0 = Kx \tag{8}$$

is said to be a guaranteed cost controller *for the uncertain system* (1), (2), *with cost* *functional* (7) *and initial condition h, if it satisfies the following conditions:*

(i) *This controller stabilizes the nominal system, i.e., the resulting closed-loop nominal system,*

$$dx(t) = [A + B_1K]x(t)dt + [H + P_1K]x(t)dw_1(t), \quad x(s) = h, \tag{9}$$

is exponentially stable in a mean-square sense. That is, there exist constants $C > 0$, $\alpha > 0$ such that

$$\mathbf{E}\|x(t)\|^2 \leq C\mathbf{E}\|h\|^2 e^{-\alpha(t-s)}.$$

(ii) *For all $s > 0$ and $h \in L_2^s$, the corresponding solution to the closed-loop uncertain system* (1), (3), (8),

$$dx(t) = ([A + B_1K]x(t) + B_2\xi(t))\,dt + [H + P_1K]x(t)dw_1(t) + P_2\xi(t)dw_2(t), \tag{10}$$

$$\begin{pmatrix} z \\ u \end{pmatrix} = \begin{bmatrix} C + DK \\ K \end{bmatrix} x, \qquad x(s) = h,$$

with any admissible uncertainty input (2), *lies in $L_2(s, \mathbf{R}^n)$. Furthermore, as a consequence, the corresponding control input $u(\cdot)$ and admissible uncertainty input $\xi_\phi(\cdot)$ lie in $L_2(s, \mathbf{R}^{m_1})$, $L_2(s, \mathbf{R}^{m_2})$, respectively.*

(iii) *The corresponding value of the cost functional* (7) *is bounded by the constant γ for all admissible uncertainties:*

$$\sup_{\phi(\cdot)\in\Phi} J^{s,h}(u(\cdot)) \leq \gamma. \tag{11}$$

Note that in the case of the structured uncertainty, the definition remains virtually the same, with the obvious replacement of the constraint (3) by the structured constraints (4) (or their matrix version (5)).

We now introduce the set $\mathcal{K}$ of guaranteed cost controllers of the form (8). Note that the constant γ in Definition 2 describes the prespecified required level of robust performance of the closed-loop system. In the (nonoptimal) guaranteed cost control problem, it is satisfactory to obtain any controller satisfying the condition (11). In this paper, we address the optimal version of the guaranteed cost control problem, in which we seek to find a control guaranteeing the minimum upper bound on the worst-case performance of the uncertain closed-loop system driven by the uncertainty input ξ_ϕ, $\phi \in \Phi$:

$$\inf_{u(\cdot)\in\mathcal{K}} \sup_{\phi\in\Phi} J^{s,h}(u(\cdot)). \tag{12}$$

Note that in this minimax optimization problem, the admissible maximizers are those which satisfy the constraint (3) (or the constraints (4)). Thus, the optimization problem (12) is a constrained minimax optimization problem. The controller solving the constrained minimax optimization problem (12) will be referred to as a minimax optimal guaranteed cost controller.

2. Stochastic differential game. In this section we consider the stochastic linear quadratic differential game associated with (1). The set up is similar to that in [26], where the special case of $P_1 = P_2 = 0$ is considered. The results presented below extend in a straightforward manner those of [26]. This extension is of primary importance for the results of this paper and will be extensively used in what follows. Hence, for the sake of completeness, we include these extensions here.

As in [26], the consideration of the stochastic linear quadratic differential game in this section assumes that the underlying system has certain stability properties. In particular, we shall require in this section that the matrix A is stable. In the subsequent sections this assumption will be significantly weakened; see Assumption 2 in the next section.

Assumption 1. *The linear system*

$$dx(t) = Ax(t)dt + Hxdw_1(t), \qquad x(s) = h, \tag{13}$$

that corresponds to the system (1), *driven by the control input* $u(\cdot) = 0$ *and the uncertainty input* $\xi(\cdot) = 0$, *is exponentially stable in mean-square sense*; *i.e.*, *there exist constants* $C > 0,\ \alpha > 0$ *such that*

$$\mathbf{E}\|x(t)\|^2 \le Ce^{-\alpha(t-s)}\mathbf{E}\|h\|^2. \tag{14}$$

As a consequence, *the matrix* A *is a stability matrix.*

Remark. In order to check Assumption 1, one can use Lyapunov arguments reducing the stability test to finding a feasible solution Y to the linear matrix inequality

$$A'Y + YA + H^*YH \le -\bar{\epsilon}I, \qquad Y' = Y > 0,$$

for a certain constant $\bar{\epsilon} > 0$.

Consider the stochastic differential game defined by (1) and the cost functional

$$\begin{aligned} \Im^{s,h}(u,\xi) &= \int_s^{+\infty} \mathbf{E}\left\{F(x(t),u(t)) - \|\xi(t)\|^2\right\} dt, \\ F(x,u) &:= \langle x, Rx + Qu\rangle + \langle u, Q'x + Gu\rangle, \end{aligned} \tag{15}$$

where $R = R' \in R^{n\times n}$, $Q \in R^{n\times m_1}$, $G = G' \in R^{m_1\times m_1}$, $R \ge 0$, $G > 0$. In (15), $x(\cdot)$ denotes the solution to (1) satisfying the initial condition $x(s) = h$ and driven by the pair of inputs $(u(\cdot), \xi(\cdot))$. In this game, $u(\cdot) \in L_2(s, \mathbf{R}^{m_1})$ is the minimizing strategy, and $\xi(\cdot) \in L_2(s, \mathbf{R}^{m_2})$ is the maximizing strategy. Note that Assumption 1 assures that any pair of inputs $(u(\cdot), \xi(\cdot)) \in L_2(s, \mathbf{R}^{m_1}) \times L_2(s, \mathbf{R}^{m_2})$ result in a square-integrable solution on $[s, +\infty)$ for any initial condition; see, e.g., [3] and also Lemma 4 in Appendix A. This implies that the cost functional (15) is well defined on $L_2(s, \mathbf{R}^{m_1}) \times L_2(s, \mathbf{R}^{m_2})$. Also, in what follows we shall consider the finite horizon version of the cost functional (15) which is defined as follows:

$$\Im_T^{s,h}(u,\xi) = \int_s^T \mathbf{E}\left\{F(x(t),u(t)) - \|\xi(t)\|^2\right\} dt, \qquad s \le T < \infty. \tag{16}$$

Note that for $T = +\infty$, $\Im_\infty^{s,h}(u,\xi) = \Im^{s,h}(u,\xi)$.

In the stochastic differential game, we seek to find

$$V := \inf_{u(\cdot)\in L_2(s,\mathbf{R}^{m_1})} \sup_{\xi\in L_2(s,\mathbf{R}^{m_2})} \Im^{s,h}(u,\xi). \tag{17}$$

Lemma 1. *Suppose that Assumption* 1 *is satisfied. Also*, *suppose there exists a constant* $\varepsilon_2 > 0$ *such that*

$$\Im^{0,0}(0,\xi) \le -\varepsilon_2|||\xi|||^2 \qquad \forall\ \xi \in L_2(0, \mathbf{R}^{m_2}). \tag{18}$$

Then the following conditions hold:

(a) *For each* $s \geq 0$, $h \in L_2^s$, *there exists a unique* $\xi_2^s(\cdot) \in L_2(s; \mathbf{R}^{m_2})$ *such that*

$$\Im^{s,h}(0, \xi_2^s(\cdot)) = \sup_{\xi(\cdot) \in L_2(s, \mathbf{R}^{m_2})} \Im^{s,h}(0, \xi). \tag{19}$$

Also, there exists a unique optimal $\xi_{2T}^s(\cdot) \in L_2(s, T; \mathbf{R}^{m_2})$ *maximizing the cost function* $\Im_T^{s,h}(0, \xi(\cdot))$. *Moreover,*

$$|||\xi_{2T}^s(\cdot)|||^2 \leq c\mathbf{E}\|h\|^2, \qquad \Im_T^{s,h}(0, \xi_{2T}^s(\cdot)) \leq c\mathbf{E}\|h\|^2 \tag{20}$$

for some $c > 0$ *independent of* $s \leq T \leq +\infty$ *and* h.

(b) *There exists a symmetric nonnegative definite matrix* X_2 *such that*

$$\sup_{\xi(\cdot) \in L_2(s, \mathbf{R}^{m_2})} \Im^{s,h}(0, \xi) = \mathbf{E}\langle h, X_2 h\rangle. \tag{21}$$

(c) *For all* $T > 0$, *there exists a unique symmetric nonnegative definite solution* $X_{2T}(\cdot)$ *to the generalized Riccati equation*

$$\begin{aligned} &\frac{dX_{2T}}{dt} + A'X_{2T} + X_{2T}A + H^*X_{2T}H + R \\ &\qquad\qquad + X_{2T}B_2(I - P_2^*X_{2T}P_2)^{-1}B_2'X_{2T} = 0, \\ &X_{2T} = 0. \end{aligned} \tag{22}$$

This solution satisfies the conditions $0 \leq X_{2T}(s) \leq X_2$ *and* $I - P_2^* X_{2T}(s) P_2 > 0$. *The optimal input* $\xi_{2T}^s(\cdot)$ *maximizing the functional* (16) *can be expressed in the form of the feedback law*

$$\xi_{2T}^s(t) = (I - P_2^* X_{2T}(t) P_2)^{-1} B_2' X_{2T} x_{2T}^s(t), \tag{23}$$

where $x_{2T}^s(\cdot)$ *is the optimal trajectory satisfying the corresponding "closed-loop" equation*

$$\begin{aligned} dx(t) = {}& (A + B_2(I - P_2^* X_{2T}(t) P_2)^{-1} B_2' X_{2T}(t)) x dt + H x dw_1(t) \\ &+ P_2 (I - P_2^* X_{2T}(t) P_2)^{-1} B_2' X_{2T}(t) x dw_2(t), \quad x(s) = h. \end{aligned} \tag{24}$$

(d) *The matrix* X_2 *satisfying* (21) *is also the minimal solution to the generalized Riccati equation*

$$A'X_2 + X_2A + H^*X_2H + R + X_2B_2(I - P_2^*X_2P_2)^{-1}B_2'X_2 = 0, \tag{25}$$

such that

$$I - P_2^* X_2 P_2 > 0. \tag{26}$$

Proof. See Appendix B. □

Remark. As in [11], it can be proved that the matrix X_2 satisfying conditions (21), (25), and (26) is such that the system

$$\begin{aligned} dx(t) = {}& (A + B_2(I - P_2^* X_2 P_2)^{-1} B_2' X_2) x dt + H x dw_1(t) \\ &+ P_2 (I - P_2^* X_2 P_2)^{-1} B_2' X_2 x dw_2(t), \qquad x(s) = h \end{aligned} \tag{27}$$

is exponentially mean-square stable. Note that in the particular case where $H = 0$, $P_2 = 0$, (25) becomes a standard Riccati equation, with the stable matrix A. Hence in this case, the equation allows for a nonnegative definite stabilizing solution satisfying condition (26).

The problem of the solvability of a generalized ARE is known in the literature as a challenging problem; see, e.g., reference [18] and the references therein. This reference presents a numerical algorithm based on homotopy methods, which solves a general class of perturbed Riccati equations. The generalized algebraic Riccati equation (25) is virtually the same as those in [18]. Hence a useful idea toward solving (25) may be to apply the method of [18]. Also, it is worth noting that the solution X_2 to (25) and inequality (26) necessarily satisfies the linear matrix inequalities

$$\begin{bmatrix} -A'X_2 - X_2A - R - H^*X_2H & X_2B_2 \\ B_2'X_2 & I - P_2^*X_2P_2 \end{bmatrix} \geq 0, \qquad X_2 \geq 0.$$

Hence, the desired matrix X_2 exists only if the above linear matrix inequalitites (LMIs) are feasible.

THEOREM 1. *Suppose that Assumption* 1 *is satisfied and there exists a constant* $\varepsilon_2 > 0$ *such that condition* (18) *holds. Also, assume that there exists a constant* $\varepsilon_1 > 0$ *such that*

$$F(x,u) > \varepsilon_1\|u\|^2 \qquad \forall\, x \in R^n, u \in R^{m_1}. \tag{28}$$

Then the following conditions hold:

(a) *For each* $s \geq 0$, $h \in L_2^s$, *there exists a unique saddle point (minimax pair) for the cost function* (15) *in* $L_2(s,R^{m_1}) \times L_2(s,R^{m_2})$.

(b) *There exists a unique symmetric nonnegative definite solution* $X \in \mathbf{R}^{n\times n}$ *to the generalized game-type algebraic Riccati equation*

$$\begin{aligned} &A'X + XA + H^*XH + R \\ &-(XB_1 + H^*XP_1 + Q)(G + P_1^*XP_1)^{-1}(XB_1 + H^*XP_1 + Q)' \\ &+XB_2(I - P_2^*XP_2)^{-1}B_2'X = 0 \end{aligned} \tag{29}$$

such that $I - P_2^*XP_2 > 0$ *and*

$$V = \mathbf{E}\langle h, Xh\rangle. \tag{30}$$

(c) *The minimax pair can be expressed in the feedback form*

$$u = F_1x, \qquad \xi = F_2x,$$

where

$$F_1 = -(G + P_1^*XP_1)^{-1}(XB_1 + H^*XP_1 + Q)', \quad F_2 = (I - P_2^*XP_2)^{-1}B_2'X.$$

(d) *The stochastic closed-loop system*,

$$\begin{aligned} &dx = (A + B_1F_1 + B_2F_2)x + (H + P_1F_1)xdw_1(t) + P_2F_2xdw_2(t), \\ &x(s) = h, \end{aligned} \tag{31}$$

is exponentially stable in mean-square sense; *i.e.*,

$$\mathbf{E}\|x(t)\|^2 \leq C\mathbf{E}\|h\|^2e^{-\alpha(t-s)}. \tag{32}$$

As a consequence, *the matrix* $A + B_1F_1 + B_2F_2$ *is Hurwitz.*

Proof. The proof follows using arguments similar to those used in proving the corresponding theorem in [26] but with evident modifications due to the fact that we have $P_1 \neq 0$, $P_2 \neq 0$ in this case. See Appendix C. □

Remark. In the particular case of $H = P_1 = P_2 = 0$, (29) reduces to the Riccati equation known in deterministic H_∞ control. In another particular case in which $B_2 = 0$, $P_2 = 0$, this equation reduces to the generalized Riccati equation known in linear-quadratic stochastic optimization [2, 27]. □

3. Stochastic H_∞control with complete state measurement. In this section we consider a stochastic H_∞ control problem related to system (1). Again, this section adapts corresponding results of [26] to our more general set-up. From now on, we no longer assume that the system (13) is exponentially stable. That is, we no longer assume that Assumption 1 is satisfied, and hence we do not suppose the a priori stability of the matrix A in (1). In what follows, we shall use instead a property which is a stochastic counterpart to the detectability of a pair of matrices.

ASSUMPTION 2. *The system* (1) *is such that* $D'D > 0$, *the matrix* $\tilde{R} = C'(I - D(D'D)^{-1}D')C$ *is nonnegative definite, and there exists a matrix* $N \in R^{n\times m}$ *such that the matrix* $A - B_1(D'D)^{-1}D'C - N\tilde{R}^{1/2}$ *is stable, i.e.,* $\|\exp((A - B_1(D'D)^{-1}D'C - N\tilde{R}^{1/2})t)\| \le ae^{-\alpha t}$, *and*

$$\frac{a^2}{\alpha}\|H - P_1(D'D)^{-1}D'C\|^2_{Q_1} < 1. \tag{33}$$

Given the system (1), the associated stochastic H_∞ control problem is to find a matrix $K \in R^{m_1\times n}$ such that the state feedback controller $u = Kx$ satisfies the following conditions:

(i) The system

$$dx = (A + B_1K)xdt + (H + P_1K)xdw_1(t) \tag{34}$$

is exponentially stable in mean-square sense, and the matrix $A + B_1K$ is stable;

(ii) The closed-loop system, corresponding to system (1) with feedback control $u = Kx$,

$$dx = [(A + B_1K)x(t) + B_2\xi(t)]dt + (H + P_1K)x(t)dw_1(t) + P_2\xi(t)dw_2(t), \tag{35}$$

satisfies the following stochastic H_∞-norm condition: there exists a constant $\varepsilon > 0$ such that

$$\mathbf{E}\int_s^{+\infty} \left(\|(C + DK)x(t)\|^2 - \|\xi(t)\|^2\right)dt \le -\varepsilon\mathbf{E}\int_s^{+\infty} \|\xi(t)\|^2 dt \quad \text{for } x(s) = 0 \tag{36}$$

for each $\xi \in L_2(s; \mathbf{R}^{m_2})$.

Remark. The exponential stability of the nominal closed-loop system (34) is a sufficient condition for the system (35) to have solutions lying in $L_2(s; \mathbf{R}^n)$ for any $\xi(\cdot) \in L_2(s; \mathbf{R}^{m_2})$.

THEOREM 2. *Suppose Assumption* 2 *is satisfied. Then the stochastic* H_∞ *control problem defined above has a solution if and only if there exists a minimal nonnegative definite symmetric solution* X *to the generalized algebraic Riccati equation*

$$\begin{aligned} &A'X + XA + H^*XH + C'C \\ &-(XB_1 + H^*XP_1 + C'D)(D'D + P_1^*XP_1)^{-1}(B_1'X + P_1^*XH + D'C) \\ &+XB_2(I - P_2^*XP_2)^{-1}B_2'X = 0 \end{aligned} \tag{37}$$

such that $I - P_2^* X P_2 > 0$ *and the stochastic system*

$$\begin{aligned} dx &= (A - B_1(D'D + P_1^* X P_1)^{-1}(B_1'X + P_1^* XH + D'C) \\ &\quad + B_2(I - P_2^* X P_2)^{-1} B_2' X) x dt \\ &\quad + (H + P_1(D'D + P_1^* X P_1)^{-1}(B_1'X + P_1^* XH + D'C)) x dw_1 \\ &\quad + P_2(I - P_2^* X P_2)^{-1} B_2' X x dw_2(t) \end{aligned} \tag{38}$$

is exponentially stable in the mean-square sense. If this condition is satisfied, then the corresponding stabilizing feedback controller which solves the stochastic H_∞ *problem is given by*

$$K = -(D'D + P_1^* X P_1)^{-1}(B_1'X + P_1^* XH + D'C). \tag{39}$$

Conversely, if the stochastic H_∞ *control problem defined above has a solution, then the solution to the Riccati equation* (37) *satisfies the condition*

$$\mathbf{E}\langle h, Xh\rangle \le \sup_{\xi \in L_2(s;\mathbf{R}^{m_2})} \mathbf{E} \int_s^{+\infty} \left(\|Cx(t) + Du(t)\|^2 - \|\xi(t)\|^2 \right) dt \tag{40}$$

for any state feedback control $u(\cdot)$ *such that the closed-loop system corresponding to this feedback control has* L_2*-summable solutions for all* $\xi \in L_2(s;\mathbf{R}^{m_2})$, *and the supremum on the right-hand side in* (40) *is finite.*

Proof. The proof follows using arguments of the corresponding result of [26] which are readily extended to the statement above. See Appendix D. □

4. Stochastic absolute stabilization. In this section, we address to the problem of absolute stabilization via a static state feedback controller. We construct a stabilizing controller of the form (8) which leads to a closed-loop uncertain system (1), (2), (8) which is absolutely stable in the following sense.

Definition 3. *A controller of the form* (8) *is said to absolutely stabilize the uncertain system* (1), (2) *if the following conditions hold*:

(i) *The nominal closed-loop system is stable. That is, for any initial condition* $x(s) = h \in L_2^s$, *the system* (34) *is exponentially mean-square stable.*

(ii) *There exists a constant* $c > 0$, *independent of the initial condition and such that for any admissible uncertainty* ϕ, *the corresponding solution to the closed-loop system* (1), (2), (8) *belongs to* $L_2(s, \mathbf{R}^n)$, *the corresponding uncertain input* $\xi_\phi(\cdot)$ *belongs to* $L_2(s, \mathbf{R}^{m_2})$, *and*

$$|||x(\cdot)|||^2 + |||\xi_\phi(\cdot)|||^2 \le c\mathbf{E}\|h\|^2. \tag{41}$$

Given matrices $R \in \mathbf{R}^{n\times n}$, $G \in \mathbf{R}^{m_1 \times m_1}$, $R' = R > 0$, $G' = G > 0$, consider the generalized algebraic Riccati equation of the form of (37),

$$\begin{aligned} &A'X + XA + H^* XH + \bar{C}'\bar{C} \\ &-(XB_1 + H^* X P_1 + \bar{C}'\bar{D})(\bar{D}'\bar{D} + P_1^* X P_1)^{-1}(B_1'X + P_1^* XH + \bar{D}'\bar{C}) \\ &+X\bar{B}_2(I - \bar{P}_2^* X \bar{P}_2)^{-1} \bar{B}_2' X = 0, \end{aligned} \tag{42}$$

where

$$\bar{C} = \begin{bmatrix} R^{1/2} \\ 0 \\ \bar{R}^{1/2} C \end{bmatrix}, \quad \bar{D} = \begin{bmatrix} 0 \\ G^{1/2} \\ \bar{R}^{1/2} D \end{bmatrix}, \quad \bar{B}_2 = B_2 \bar{G}^{-1/2}, \quad \bar{P}_2 = P_2 \bar{G}^{-1/2}. \tag{43}$$

Also, associated with the uncertain system (1), (2), consider the cost functional (7).

THEOREM 3. *Suppose Assumption* 2 *is satisfied and suppose also that there exists a minimal nonnegative definite stabilizing solution X to the generalized Riccati equation* (42), *i.e., such that $I - \bar{P}_2^* X \bar{P}_2 > 0$ and that the system*

$$
\begin{aligned}
dx = {} & \big[A - B_1(\bar{D}'\bar{D} + P_1^* X P_1)^{-1}(B_1'X + P_1^* X H + \bar{D}'\bar{C}) \\
& + \bar{B}_2(I - \bar{P}_2^* X \bar{P}_2)^{-1}\bar{B}_2' X\big]\, xdt \\
& + \big[H - P_1(\bar{D}'\bar{D} + P_1^* X P_1)^{-1}(B_1'X + P_1^* X H + \bar{D}'\bar{C})\big]\, xdw_1(t) \\
& + \bar{B}_2(I - \bar{P}_2^* X \bar{P}_2)^{-1}\bar{B}_2' X x dw_2(t)
\end{aligned} \tag{44}
$$

is exponentially stable in mean-square sense, and as a consequence, the matrix

$$A - B_1(\bar{D}'\bar{D} + P_1^* X P_1)^{-1}(B_1'X + P_1^* X H + \bar{D}'\bar{C}) + \bar{B}_2(I - \bar{P}_2^* X \bar{P}_2)^{-1}\bar{B}_2' X$$

is stable. Then the controller given by

$$u^0(t) = Kx(t), \qquad K := -(\bar{D}'\bar{D} + P_1^* X P_1)^{-1}(B_1'X + P_1^* X H + \bar{D}'\bar{C}), \tag{45}$$

is an absolutely stabilizing controller for uncertain system (1), (2). *Furthermore, the corresponding value of the cost function* (7) *satisfies the bound*

$$\sup_{\phi \in \Phi} J^{s,h}(u^0(\cdot)) \le \mathbf{E}\langle h, (X + W)h\rangle. \tag{46}$$

Proof. Consider a stochastic differential game defined by the system

$$
\begin{aligned}
dx(t) = [Ax(t) + B_1 u(t) + \bar{B}_2 \bar{\xi}(t)]dt + (Hx(t) + P_1 u(t))dw_1(t) \\
+ \bar{P}_2 \bar{\xi}(t) dw_2(t),
\end{aligned} \tag{47}
$$

and the cost function

$$
\begin{aligned}
\Im^{s,h}(u(\cdot), \bar{\xi}(\cdot)) &= J^{s,h}(u(\cdot)) + \mathbf{E}\int_s^\infty \left(\langle z(t), \bar{R} z(t)\rangle - \|\bar{\xi}\|^2\right) dt \\
&= \mathbf{E}\int_s^\infty (\|\bar{C}x + \bar{D}u\|^2 - \|\bar{\xi}\|^2) dt,
\end{aligned} \tag{48}
$$

where matrices $\bar{B}_2$, $\bar{P}_2$, $\bar{C}$, and $\bar{D}$ are defined as in (43). Under Assumption 2, the conditions of the sufficiency part of Theorem 2 are satisfied for the system (47) and cost functional (48). This leads to the conclusion that the controller (45) solves the stochastic H_∞ control problem associated with (47), (48). That is, the system (34), where the matrix K is defined by (45), is exponentially stable in mean-square sense, and as a consequence, the matrix $A + B_1 K$ is stable. Furthermore, the controller (45) satisfies the following stochastic H_∞-norm condition: there exists a constant $\varepsilon > 0$ such that

$$
\begin{aligned}
\mathbf{E}\int_0^{+\infty} \left(\|(\bar{C} + \bar{D}K)x(t)\|^2 - \|\bar{\xi}(t)\|^2\right) dt \le -\varepsilon \mathbf{E}\int_0^{+\infty} \|\bar{\xi}(t)\|^2 dt \\
\text{for } x(0) = 0 \text{ and } \forall\ \bar{\xi}(\cdot) \in L_2(0, \mathbf{R}^{m_2}),
\end{aligned} \tag{49}
$$

where $x(\cdot)$ is a solution to the equation

$$
\begin{aligned}
dx(t) = ((A + B_1 K)x(t) + \bar{B}_2 \bar{\xi}(t))dt + (H + P_1 K)x(t)dw_1(t) \\
+ \bar{P}_2 \bar{\xi}(t) dw_2(t)
\end{aligned} \tag{50}
$$

obtained by the substitution of (45) into (47). This conclusion shows that the system (50) and functional (48) satisfy the conditions of Lemma 1. Using the result of this lemma, we obtain

$$\bar{\Im}^{s,h}(u^0(\cdot),\bar{\xi}(\cdot)) \leq \mathbf{E}\langle h, X_K h\rangle \quad \forall\ \bar{\xi}(\cdot) \in L_2(0,\mathbf{R}^{m_2}), \tag{51}$$

where the matrix X_K is the minimal solution to the following Riccati equation of the form (25):

$$\begin{aligned}&(A+B_1K)'X_K + X_K(A+B_1K) + (H+P_1K)^*X_K(H+P_1K)\\ &+(\bar{C}+\bar{D}K)'(\bar{C}+\bar{D}K) + X_K\bar{B}_2(I-\bar{P}_2^*X_K\bar{P}_2)^{-1}\bar{B}_2'X_K = 0,\\ &I-\bar{P}_2^*X_K\bar{P}_2 > 0.\end{aligned} \tag{52}$$

Also, one may observe that by using elementary transformations, (42) can be transformed into the following equation:

$$\begin{aligned}&(A+B_1K)'X + X(A+B_1K) + (H+P_1K)^*X(H+P_1K)\\ &+(\bar{C}+\bar{D}K)'(\bar{C}+\bar{D}K) + X\bar{B}_2(I-\bar{P}_2^*X\bar{P}_2)^{-1}\bar{B}_2'X = 0,\\ &I-\bar{P}_2^*X\bar{P}_2 > 0,\end{aligned} \tag{53}$$

which is the same as (52). Thus, the minimal solutions to these equations are equal, $X = X_K$, and consequently, one can replace X_K in (51) with the solution to the algebraic Riccati equation (42) to obtain

$$\bar{\Im}^{s,h}(u^0(\cdot),\bar{\xi}(\cdot)) \leq \mathbf{E}\langle h, Xh\rangle. \tag{54}$$

The subsequent proof requires the use of Theorem 3 of [5]. In [5] this result, referred to as the stochastic counterpart of the Kalman–Yakubovich lemma, established the existence of a solution to a certain linear matrix inequality related to the sign-indefinite linear-quadratic stochastic control problem considered in this reference. We wish to apply this result to the "control problem" $\sup_{\bar{\xi}(\cdot)} \Im^{s,h}(u^0,\bar{\xi}(\cdot))$, with the underlying system (50) and the stable system (34), with the matrix K given by (45). The conditions under which one can apply Theorem 3 in [5] to the above control problem are virtually the same as the conditions of Lemma 1. Note that Theorem 3 in [5] requires the cost functional in the above control problem to satisfy a certain condition of coercivity [5, 13]. In our case the fact that this condition is satisfied follows in a straightforward way from the stochastic H_∞ condition (49). Note also that Assumption 3 of [5] holds in our case, since we deal with finite-dimensional equations and operators. Since we have verified the conditions of Lemma 1 for the system (50) and functional (48), the application of Theorem 3 in [5] implies the existence of a symmetric matrix M and a positive constant ϵ satisfying the LMI

$$\begin{aligned}&-2\langle (A+B_1K)x + \bar{B}_2\bar{\xi}, Mx\rangle - \langle x, (H+P_1K)^*M(H+P_1K)x\rangle\\ &-\langle \bar{\xi}, \bar{P}_2^*M\bar{P}_2\bar{\xi}\rangle - \|(\bar{C}+\bar{D}K)x\|^2 + \|\bar{\xi}\|^2 \geq \epsilon(\|x\|^2+\|\bar{\xi}\|^2)\ \forall\ x\in\mathbf{R}^n, \bar{\xi}\in\mathbf{R}^{m_2}.\end{aligned} \tag{55}$$

The standard form for this LMI is the following:

$$\left[\begin{array}{cc} \begin{array}{c}-(A+B_1K)'M - M(A+B_1K)\\ -(H+P_1K)^*M(H+P_1K) - (\bar{C}+\bar{D}K)'(\bar{C}+\bar{D}K)\end{array} & -M\bar{B}_2 \\ -\bar{B}_2'M & I-\bar{P}_2M\bar{P}_2 \end{array}\right] > 0.$$

Recall that the nominal closed-loop system (34) corresponding to the system (50), driven by the uncertainty input $\xi_\phi(t) \equiv 0$, is exponentially stable in mean-square sense. Hence, $\mathbf{E}\|x(t)\|^2 \to 0$ as $t \to \infty$ in this particular case. Therefore, it follows from (55) that

$$\mathbf{E}\langle h, Mh\rangle \geq \epsilon_1 \mathbf{E}\int_s^\infty \|x(t)\|^2 dt,$$

and hence $M > 0$. To establish this fact, we have used the Ito formula along the solution to the system (34) on the interval $[s,t]$, with initial condition $x(s) = h$ and then sent t to ∞.

Now, let $\xi_\phi(\cdot)$ be the uncertainty input corresponding to an admissible uncertainty $\phi(\cdot) \in \Phi$. Then define

$$\bar{\xi}(t) = \begin{cases} \bar{G}^{1/2}\xi_\phi(t) = \bar{G}^{1/2}\phi(t, x|_s^t, Kx|_s^t), & t \in [s, t_j], \\ 0, & t > t_j, \end{cases}$$

where t_j is as defined in Definition 1. Also, let $x(t)$ be the corresponding solution to (50) satisfying the initial condition $x(s) = h$. Then it follows from inequality (55) that

$$\begin{aligned} -\mathbf{E}\langle x(t), Mx(t)\rangle|_s^{t_j} + \mathbf{E}\int_s^{t_j} (-\|(\bar{C}+\bar{D}K)x(t)\|^2 + \|\bar{\xi}(t)\|^2)dt& \\ \geq \epsilon \mathbf{E}\int_s^{t_j} (\|x(t)\|^2 + \langle \xi_\phi(t), \bar{G}\xi_\phi(t)\rangle)dt.& \end{aligned} \tag{56}$$

The Ito formula is used to derive (56) from (55). Hence, using (3) and the nonnegativeness of R, G, we obtain

$$\begin{aligned} &-\mathbf{E}\langle x(t), Mx(t)\rangle|_s^{t_j} + \mathbf{E}\langle x(s), Wx(s)\rangle \\ &\geq -\mathbf{E}\langle x(t), Mx(t)\rangle|_s^{t_j} - \mathbf{E}\int_s^{t_j} (\langle x(t), Rx(t)\rangle + \langle Kx(t), GKx(t)\rangle)dt \\ &-\mathbf{E}\int_s^{t_j} \left(\langle z(t), \bar{R}z(t)\rangle + \langle \xi_\phi(t), \bar{G}\xi_\phi(t)\rangle\right) dt \\ &\geq \epsilon_1 \mathbf{E}\int_s^{t_j} (\|x(t)\|^2 + \langle \xi_\phi(t), \bar{G}\xi_\phi(t)\rangle)dt. \end{aligned} \tag{57}$$

Since we have established that $M > 0$ and $x(s) = h$, (57) implies

$$\epsilon_1 \mathbf{E}\int_s^{t_j} (\|x(t)\|^2 + \langle \xi_\phi(t), \bar{G}\xi_\phi(t)\rangle)dt \leq \mathbf{E}\langle h, (M+W)h\rangle. \tag{58}$$

Thus, we see that the expression on the left in (58) is uniformly bounded with respect to t_j. Therefore, (58) implies that $x(\cdot) \in L_2(s, \mathbf{R}^n)$ for any admissible ϕ. Also, using the fact that $\bar{G} > 0$, it also follows that $\xi_\phi(\cdot) \in L_2(s, \mathbf{R}^{m_2})$ for any admissible ϕ. Therefore, for any admissible uncertainty ϕ, the input $\bar{\xi}(\cdot) = \bar{G}^{1/2}\xi_\phi(\cdot)$ is an admissible uncertainty input in the stochastic H_∞ control problem defined by the system (47) and cost functional (48). For this uncertainty input, we obtain from (48) and (54)

$$\begin{aligned} J^{s,h}(u^0) &\leq -\int_0^\infty \mathbf{E}\left(\langle z(t), \bar{R}z(t)\rangle - \langle \xi_\phi(t), \bar{G}\xi_\phi(t)\rangle\right) dt + \mathbf{E}\langle h, Xh\rangle \\ &\leq \mathbf{E}\langle h, (X+W)h\rangle. \qquad \square \end{aligned}$$

Remark. We have observed that the generalized algebraic Riccati equation (42) can be transformed into an algebraic Riccati equation of the form (25). Hence as in the case of (25), a possible approach to solving the algebraic Riccati equation (42) is to apply homotopy methods [18].

5. Minimax optimal state feedback controller. In this section we assume that the uncertainty in the system is structured and each uncertainty loop satisfies its own stochastic integral quadratic constraint of the form (4). As has been mentioned in the introduction, for any numbers $\tau_1 > 0, \ldots, \tau_k > 0$, we replace all of the constraints by a single stochastic integral quadratic constraint (5). For the uncertain stochastic system (1), with the uncertainty (2) satisfying this structured uncertainty constraint, we solve the corresponding minimax optimal guaranteed cost control problem. The main result of this section, Theorem 4, shows that the problem of finding the minimax optimal guaranteed cost controller can be reduced to a finite-dimensional optimization problem.

We shall assume in this section that the system (1) satisfies Assumption 2.

Let $\mathcal{T}$ denote the set of vectors $\tau \in \mathbf{R}_+^k$ such that the corresponding Riccati equation (42) has a nonnegative stabilizing solution X_τ; i.e.,

$$\mathcal{T} = \left\{ \tau \in \mathbf{R}^k\colon \begin{array}{l} \tau_1 > 0, \ldots, \tau_k > 0 \text{ and Riccati equation (42) with} \\ \bar{R} = \bar{R}_\tau, \bar{G} = \bar{G}_\tau, \text{ has a stabilizing solution} \\ X_\tau \geq 0 \text{ such that } \bar{G}_\tau - P_2^* X_\tau P_2 > 0. \end{array} \right\} \tag{59}$$

THEOREM 4.

(i) *Given $\gamma > 0$ and initial condition h, there exists a guaranteed cost controller for the uncertain system* (1), (2) *if and only if the set $\mathcal{T}$ defined in* (59) *is nonempty.*

(ii) *Suppose the set $\mathcal{T}$ is nonempty. Then*

$$\inf_{u(\cdot)\in\mathcal{K}} \sup_{\phi\in\Phi} J^{s,h}(u(\cdot)) = \inf_{\tau\in\mathcal{T}} \mathbf{E}\left[\langle h, (X_\tau + W_\tau)h\rangle\right]. \tag{60}$$

Furthermore, let $\tau^* \in \mathcal{T}$ attain the infimum on the right-hand side of (60). *Then the corresponding control $u^0_{\tau^*}$ defined by* (45) *with the matrix $X = X_{\tau^*}$ is the state feedback minimax optimal guaranteed cost control which minimizes the worst case of the cost functional* (48) *in the constrained stochastic optimization problem* (12) *subject to the stochastic integral quadratic constraint* (4) *and absolutely stabilizes the uncertain system* (1).

The proof of this theorem follows the same form as the proof of the main result of [20]. As in [20], we use the so-called S-procedure [15, 20, 21, 22, 29] to reduce our constrained optimization problem to a problem without constraints. However, in contrast to these papers, the system (1) is nonautonomous in nature due to the stochastic perturbations. This leads us to consider a special construction of shift operators in order to satisfy conditions of the S-procedure. This construction involves special metrical transitive transformations of stochastic processes. The basic properties of these transformations are given in Appendix A.

As in [20], we begin with a result establishing that a stabilizing guaranteed cost controller for the uncertain system (1), (2) exists if and only if the set $\mathcal{T}$ defined in (59) is nonempty.

LEMMA 2. *Given a positive constant γ and initial condition h, the following statements are equivalent*:

(i) *for the given $\gamma > 0$, there exists a guaranteed cost controller of the form* (8) *which guarantees that the closed-loop uncertain system* (1), (3), (8) *with the cost* *function* (7) *and given initial condition does not exceed the prescribed cost value*

$$\sup_{\phi\in\Phi} J^{s,h}(u(\cdot)) < \gamma; \tag{61}$$

(ii) *for a given initial condition, there exists a* $\tau \in \mathcal{T}$ *such that*

$$\mathbf{E}\langle h, (X_\tau + W_\tau)h\rangle < \gamma. \tag{62}$$

Proof of Lemma 2. (i)⇒(ii). First observe that since (8) corresponds to a guaranteed cost controller, then the nominal closed-loop system (9) is exponentially stable. This implies that for any $\xi(\cdot) \in L_2(s;\mathbf{R}^{m_2})$, the corresponding solution $x(\cdot)$ to (10) belongs to $L_2(s;\mathbf{R}^n)$. As a consequence, the corresponding uncertainty output $z(\cdot)$ lies in $L_2(s;\mathbf{R}^p)$ and the corresponding control input $u_\xi(\cdot)$ lies in $L_2(s;\mathbf{R}^{m_1})$.

Note that (61) implies the existence of a constant $\varepsilon > 0$ such that

$$(1+\varepsilon)J^{s,h}(u_\phi(\cdot)) \le \gamma - \varepsilon \quad \forall\ \phi \in \Phi(\bar{R}_1, \bar{G}_1, W_1, \ldots, \bar{R}_k, \bar{G}_k, W_k), \tag{63}$$

where $u_\phi(\cdot)$ is the control input generated by the closed-loop uncertain system (10), (2).

It follows from the above observation that the following quadratic functionals are well defined on $L_2(s;\mathbf{R}^{m_2})$:

$$\begin{aligned}&\mathcal{G}_0(\xi(\cdot)) := -(1+\varepsilon)J^{s,h}(u_\xi(\cdot)) + \gamma - \varepsilon,\\ &\mathcal{G}_i(\xi(\cdot)) := \mathbf{E}\int_0^\infty \left(\langle z_i(t), \bar{R}_i z_i(t)\rangle - \langle \xi_i(t), \bar{G}_i \xi_i(t)\rangle\right)dt + \mathbf{E}\langle h, W_i h\rangle,\end{aligned} \tag{64}$$

where $x(\cdot)$ and $z_i(\cdot)$ correspond to the solution of the closed-loop system (10) with uncertainty input $\xi(\cdot) \in L_2(s;\mathbf{R}^{m_2})$. Furthermore, since (8) corresponds to a guaranteed cost controller, then for any admissible uncertainty ϕ, the corresponding admissible uncertainty input $\xi_\phi(\cdot) \in L_2(s;\mathbf{R}^{m_2})$ and the corresponding uncertainty output $z(\cdot) \in L_2(s;\mathbf{R}^p)$. Therefore, since $t_j \to \infty$ as $j \to \infty$, it follows from (4) that

$$\mathcal{G}_i(\xi_\phi(\cdot)) \ge 0 \quad \forall\ \phi \in \Phi(\bar{R}_1, \bar{G}_1, W_1, \ldots, \bar{R}_k, \bar{G}_k, W_k) \qquad (i = 1, \ldots, k). \tag{65}$$

We now check that the quadratic functionals $\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_k$ form an S-system in the terminology of [29]; i.e., we check that there exists a sequence of linear bounded operators $\mathbf{T}_j\colon L_2(s;\mathbf{R}^{m_2}) \to L_2(s;\mathbf{R}^{m_2})$, $j = 1, 2, \ldots$, such that this sequence weakly converges to zero in $\mathcal{L}(L_2(s;\mathbf{R}^{m_2}))$ and

$$\begin{aligned}J^{s,0}(u_j(\cdot)) \to J^{s,0}(u(\cdot)), \quad \mathcal{G}_i(\xi_j(\cdot)) \to \mathcal{G}_i(\xi(\cdot)) \quad &\text{as } j \to \infty\\ \forall\ \xi(\cdot) \in L_2(s;\mathbf{R}^{m_2}) \text{ and } i = 1, 2, \ldots, k&,\end{aligned} \tag{66}$$

where $u_j(\cdot)$ corresponds to the solution of (10) generated by the uncertainty input $\mathbf{T}_j\xi(\cdot)$ with zero initial condition. We choose the operators $\mathbf{T}_j$ as follows: for each $\xi(\cdot) \in L_2(s;\mathbf{R}^{m_2})$,

$$\xi_j(t,\omega) = (\mathbf{T}_j\xi)(t,\omega) := \begin{cases} 0, & s \le t < t_j,\\ \xi(t - t_j, \Gamma_{t_j}\omega), & t \ge t_j,\end{cases} \tag{67}$$

where t_j is the sequence from Definition 1, $t_j \to \infty$ as $j \to \infty$, and Γ_{t_j}, $j = 1, 2, \ldots$, are metrically transitive transformations from the translation semigroup generated by the Wiener process (w_1, w_2). (See Appendix A for details.) Given any two functions $\xi_1(\cdot), \xi_2(\cdot) \in L_2(s; \mathbf{R}^{m_2})$, the Cauchy inequality gives

$$\left|\mathbf{E}\int_s^\infty \langle (\mathbf{T}_j\xi_1)(t), \xi_2(t)\rangle dt\right|^2 = \left|\mathbf{E}\int_{t_j}^\infty \langle \xi_1(t-t_j, \Gamma_{t_j}\omega), \xi_2(t)\rangle dt\right|^2$$
$$\leq \mathbf{E}\int_s^\infty \|\xi_1(t)\|^2 dt\, \mathbf{E}\int_{t_j}^\infty \|\xi_2\|^2 dt$$

since Γ_{t_j} preserves probability measures, and $\mathbf{E}f(x(\Gamma_{t_j}\omega)) = \mathbf{E}f(x)$ for any Borel-measurable vector function f. However, the term $\mathbf{E}\int_{t_j}^\infty \|\xi_2\|^2 dt$ in the above inequality tends to 0 as $j \to \infty$. This implies $\mathbf{T}_j \to 0$ weakly.

Next, observe that Lemma 2 of [3] (see also Lemma 3 in Appendix A) states that the solution to the closed-loop system (10) with zero initial condition and corresponding to an uncertainty input $\mathbf{T}_j\xi$ is such that

$$x_j(t,\omega) = \begin{cases} 0, & s \leq t < t_j, \\ x(t-t_j, \Gamma_{t_j}\omega), & t \geq t_j, \end{cases} \tag{68}$$
$$u_j(t,\omega) = \begin{cases} 0, & s \leq t < t_j, \\ u(t-t_j, \Gamma_{t_j}\omega), & t \geq t_j, \end{cases}$$
$$z_j(t,\omega) = \begin{cases} 0, & s \leq t < t_j, \\ z(t-t_j, \Gamma_{t_j}\omega), & t \geq t_j. \end{cases}$$

Then, the substitution of (67) and (68) into (64) gives (66). Thus, we conclude that the family of functionals above forms an S-system. Furthermore, since $W_i > 0$, then for a given nonzero h, zero input $\xi(\cdot) \equiv 0$ gives $\mathcal{G}_i(0) > 0\ \forall\ i = 1, 2, \ldots, k$. This means that constraints are regular in terminology of [29]. Also, it follows from (63) and (65) that the condition $\mathcal{G}_1(\xi(\cdot)) \geq 0, \ldots, \mathcal{G}_k(\xi(\cdot)) \geq 0$ implies that $\mathcal{G}_0(\xi(\cdot)) \geq 0$. We have satisfied all conditions of Theorem 1 of [29]. Therefore, for a given nonzero h, there exist constants $\tau_1 \geq 0, \ldots, \tau_k \geq 0$ such that

$$\mathcal{G}_0(\xi) \geq \sum \tau_i \mathcal{G}_i(\xi) \quad \forall\ \xi(\cdot) \in L_2(s; \mathbf{R}^{m_2}). \tag{69}$$

Now, let us use these τ_i to define the functional

$$J_{\varepsilon,\tau}^{s,h}(\xi(\cdot)) := \varepsilon J^{s,h}(u_\xi(\cdot)) + \bar{\Im}^{s,h}(u_\xi(\cdot), \bar{G}_\tau^{1/2}\xi(\cdot)), \tag{70}$$

where the coefficients of $\bar{\Im}^{s,h}$ in (48) are defined as in (43):

$$\bar{\Im}^{s,h}(u(\cdot), \bar{G}_\tau^{1/2}\xi(\cdot)) = J^{s,h}(u(\cdot)) + \mathbf{E}\int_s^\infty \left(\langle z(t), \bar{R}_\tau z(t)\rangle - \langle \xi(t), \bar{G}_\tau \xi(t)\rangle\right) dt.$$

Then it follows from (69) that

$$J_{\varepsilon,\tau}^{s,h}(\xi(\cdot)) \leq \gamma - \varepsilon - \mathbf{E}\langle h, W_\tau, h\rangle \quad \forall\ \xi \in L_2(s; \mathbf{R}^{m_2}). \tag{71}$$

The same arguments as those in [20] (see Claims 1 and 2 in the proof of Lemma 3.2 of this reference) then give us

$$J_{\varepsilon,\tau}^{s,0}(\xi(\cdot)) \leq 0 \quad \forall\ \xi \in L_2(s; \mathbf{R}^{m_2}), \tag{72}$$

and also $\tau_i > 0\ \forall\ i = 1, \ldots, k$.

Next, let us show that the generalized Riccati equation (42) with $\bar{R} = \bar{R}_\tau$, $\bar{G} = \bar{G}_\tau$ has a nonnegative definite stabilizing solution X_τ such that $\bar{G}_\tau - P_2^* X_\tau P_2 > 0$, where constants $\tau_1, \ldots, \tau_k$ are as in (69). In terms of our notation, this will mean that $\tau \in \mathcal{T}$. Notice that (72) implies

$$\Im^{s,0}(u_\xi, \bar{G}^{1/2}\xi(\cdot)) \leq -\varepsilon J^{s,0}(u_\xi(\cdot)) \leq 0 \qquad \forall\, \xi \in L_2(s; \mathbf{R}^{m_2}). \tag{73}$$

Therefore, since $R > 0$, $G > 0$, $\bar{G}_\tau > 0$, it follows that there exists a constant $\alpha > 0$ such that

$$\mathbf{E}\int_s^\infty \left(\|\bar{C}x(t) + \bar{D}u_\xi(t)\|^2 - \|\bar{\xi}(t)\|^2\right) dt < -\alpha \mathbf{E}\int_s^\infty \|\bar{\xi}(t)\|^2 dt \qquad \forall\, \bar{\xi}(\cdot) \in L_2(s; \mathbf{R}^{m_2}) \tag{74}$$

for $h = 0$. To see this, suppose (74) does not hold; i.e., there exists a sequence $\{\bar{\xi}^l(\cdot)\}_{l=1}^\infty \subset L_2(s; \mathbf{R}^{m_2})$ such that

$$\lim_{l\to\infty} \frac{\mathbf{E}\int_s^\infty \|\bar{C}x^l(t) + \bar{D}u_{\bar{\xi}^l}(t)\|^2 dt}{|||\bar{\xi}^l(t)|||^2} = 1. \tag{75}$$

It is clear that we can choose this sequence such that $|||\bar{\xi}^l||| = 1$, since (10) is linear with respect to $\bar{\xi}$, and (70) is a quadratic functional with respect to $\bar{\xi}$. Then, using the weak compactness of a ball in Hilbert space [31], there exists a subsequence $\{\bar{\xi}^{l_\nu}\}_{\nu=1}^\infty$ such that $\bar{\xi}^{l_\nu} \to \bar{\xi}_*$ weakly as $\nu \to \infty$, $|||\bar{\xi}_*||| = 1$. Also, $x^{l_\nu} \to x_*$ weakly as $\nu \to \infty$, since (10) is linear and (9) is exponentially stable. Furthermore, we observe that (73) and (75) imply that $x^{l_\nu} \to 0$ strongly since $R > 0$. Indeed, from (73) we have

$$\Im^{s,0}(u_{\bar{\xi}^{l_\nu}}, \bar{\xi}^{l_\nu}(\cdot)) \leq -\varepsilon_1 \mathbf{E}\int_0^\infty \|x^{l_\nu}(t)\|^2 dt \leq 0. \tag{76}$$

From (75), the expression on the left-hand side of (76) tends to 0 as $l \to \infty$. Hence $x^{l_\nu} \to 0$ strongly. Then we have $|||\bar{C}x^{l_\nu} + \bar{D}u_{\bar{\xi}^{l_\nu}}||| \to 0$ which contradicts (75). Thus (74) must hold.

The inequality (74) implies that the given guaranteed cost control solves the stochastic H_∞ control problem for the system (1) and the cost function (48). Then, we conclude, using Theorem 2, that there exists a symmetric, nonnegative definite stabilizing solution X_τ to the Riccati equation (42), which is a minimal solution, and

$$\bar{G}_\tau - P_2^* X_\tau P_2 = \bar{G}_\tau^{1/2}(I - \bar{P}_2^* X_\tau \bar{P}_2)\bar{G}_\tau^{1/2} > 0.$$

Thus, $\tau \in \mathcal{T}$. Moreover, inequality (40), established in Theorem 2, and condition (71) imply that

$$\mathbf{E}\langle h, X_\tau h\rangle \leq \sup_{\bar{\xi}(\cdot)} \Im^{s,h}(u_\xi(\cdot), \bar{\xi}(\cdot)) \leq \sup_{\xi(\cdot)} J^{s,h}_{\varepsilon,\tau}(\xi(\cdot)) \leq \gamma - \varepsilon - \mathbf{E}\langle h, W_\tau h\rangle. \tag{77}$$

Hence, condition (62) holds.

(ii)⇒(i). This part of the proof follows immediately from Theorem 3. □

Proof of Theorem 4. The theorem follows immediately from Lemma 2. In particular, note that the absolute stabilizing properties of the minimax optimal guaranteed cost controller were established in the second part of the proof of Lemma 2, where we referred to Theorem 3. □

6. Conclusions.

This paper has been concerned with existence and optimality of a guaranteed cost controller for an uncertain system subject to structured uncertainty. The new class of uncertainty satisfying so-called stochastic integral quadratic constraints has been introduced. It has been shown that such constraints naturally describe some practically important classes of uncertainty.

In this paper, we have shown that for each initial state of the system, the linear static state feedback controller yielding an optimal worst-case performance in the face of stochastic structured uncertainty in the system, is found by parametric optimization of solutions of a parameter-dependent generalized matrix Riccati equation. The generalized Riccati equation is of the type arising in stochastic H_∞ theory and related stochastic differential games.

As in the deterministic case, the S-procedure has been used to convert the constrained stochastic optimization problem into a problem without constraints. However, due to the stochastic character of uncertainty, this has led us to consider a special construction of shift operators in order to satisfy conditions of the S-procedure. This construction has involved special metrical transitive transformations of stochastic processes.

Appendix A.

Let Ω be any space of elements ω. Let $\mathcal{P}$ be a probability measure defined on a Borel field $\mathcal{F}$ of ω sets. A transformation Γ, taking points of Ω into points of Ω, is called a one-to-one *measure-preserving point transformation* if it is one-to-one and has domain and range Ω and if it and its inverse take measurable sets into measurable sets of the same probability.

A set transformation Γ defined on the Borel field $\mathcal{F}$, taking sets of $\mathcal{F}$ into sets of $\mathcal{F}$, is called a *measure-preserving set transformation* if the following conditions are satisfied:

(i) Γ is single-valued, modulo sets of probability 0; i.e., if $\tilde{\Lambda}$ is an image of Λ under Γ, the class of all images of Λ is the class of all measurable sets differing from $\tilde{\Lambda}$ by sets of probability 0.

(ii) $\mathcal{P}(\Gamma\Lambda) = \mathcal{P}(\Lambda)$.

(iii) Neglecting ω sets of probability 0,

$$\Gamma(\Lambda_1 \cup \Lambda_2) = \Gamma\Lambda_1 \cup \Gamma\Lambda_2,$$
$$\Gamma(\bigcup_{i=1}^{\infty} \Lambda_i) = \bigcup_{i=1}^{\infty} \Gamma\Lambda_i,$$
$$\Gamma(\Omega - \Lambda) = \Omega - \Gamma\Lambda.$$

A one-to-one measure-preserving point transformation Γ induces a measure-preserving set transformation.

A measurable set is called *invariant* under a measure-preserving point or set transformation if it differs from its images by sets of probability 0. A measure-preserving point or set transformation is called *metrically* transitive if the only invariant sets are those which have probability 0 or 1.

A family $\{\Gamma_s, s \geq 0\}$ of transformations, taking points of Ω into points of Ω, is called a *translation semigroup of measure-preserving one-to-one point transformations* if each Γ_s is a one-to-one measure-preserving point transformation and if

$$\Gamma_{s+t} = \Gamma_s\Gamma_t \quad \forall\ s, t \geq 0. \tag{A.1}$$

The transformation Γ_0 will necessarily be the identity.

A family $\{\Gamma_s, s \geq 0\}$ of set transformations is called a *translation semigroup of measure-preserving set transformations* if (A.1) is true modulo the sets of probability 0. The transformation Γ_0 will be the identity in the sense that every image of a measurable set Λ under Γ_0 will differ from Λ by at most a set of probability 0. A translation semigroup of measure-preserving point or set transformations is called metrically transitive if the only invariant sets are those which have probability 0 or 1.

The scalar Wiener process is an example of a process generating a metrically transitive translation semigroup of measure-preserving set transformations [6]. These transformations are generated by time shifts of strictly stationary increments of the Wiener process. Also, it is known [6] that the semigroup of shifts of a strictly stationary process is metrically transitive if and only if the semigroup of shifts for the corresponding canonical process in the coordinate space is metrically transitive. For canonical processes, shifts are point transformations. This allows one to pass from considering the Wiener process over an abstract probability space to considering its canonical version for which metrically transitive shifts have a simple construction.

The canonical representation (see, e.g., [14]) for a scalar Wiener process is the probability space $(\overline{\Omega}, \overline{\mathcal{P}}, \overline{\mathcal{F}})$ and the process $\overline{w}(t)$, $t \geq 0$ such that

$\overline{\Omega}$ is the set of continuous functions $\omega(t)$: $[0, \infty) \to \mathbf{R}^1$, starting from zero at $t = 0$;

$\overline{\mathcal{F}}$ is the Borel σ-field generated by the cylindrical subsets in $\overline{\Omega}$;

$\overline{\mathcal{P}}$ is the Wiener probability measure on $(\overline{\Omega}, \overline{\mathcal{F}})$;

$\overline{w}(t)$ is a stochastic process defined as follows:

$$\overline{w}(t, \omega) = \omega(t) \quad (\omega \in \overline{\Omega}, t \geq 0).$$

As in [6], given any $s \geq 0$, each $\omega \in \overline{\Omega}$ has corresponding $\omega_1 \in \overline{\Omega}$ such that

$$\overline{w}(t, \omega_1) = \overline{w}(t + s, \omega) - \overline{w}(s, \omega) \qquad \forall\ t \geq 0.$$

The relation $\omega_1 = \Gamma_s \omega$ generates a metrically transitive translation semigroup of measure-preserving transformations.

Each of the above measure-preserving transformations induces a transformation of random variables, which takes $\bar{\mathcal{F}}_t$-measurable variables into $\bar{\mathcal{F}}_{t+s}$-measurable ones. Here $\{\bar{\mathcal{F}}_t, t \geq 0\}$ denotes the Borel filtration, generated by $\overline{w}(\theta)$, $0 \leq \theta < t$. When applied to diffusion processes satisfying linear Ito equations, this fact gives the following result.

LEMMA 3 (see Brusin [3]). *Let $x(t, \omega)$ and $y(t + s, \omega)$, $t \geq 0$, $s > 0$, be two $\mathbf{R}^n$-valued stochastic processes being the unique solutions to integral equations* (A.2) *and* (A.3) *below, respectively:*

$$\text{(A.2)} \qquad x(t, \omega) = h(\omega) + \int_0^t (Ax(\theta, \omega) + Bu(\theta, \omega))d\theta + \int_0^t (Hx(\theta, \omega) + Pu(\theta, \omega))dw(\theta),$$

$$\text{(A.3)} \qquad y(t + s, \omega) = g(\omega) + \int_s^{t+s} (Ay(\theta, \omega) + Bv(\theta, \omega))d\theta + \int_s^{t+s} (Hy(\theta, \omega) + Pv(\theta, \omega))dw(\theta),$$

$h \in L_2^0$, $g \in L_2^s$, $u \in L_2(0, \mathbf{R}^m)$, $v \in L_2(s, \mathbf{R}^m)$. If $u(t, \Gamma_s \omega) = v(t + s, \omega)$ and $g(\omega) = h(\Gamma_s \omega)$ with probability 1, *then*

$$y(t + s, \omega) = x(t, \Gamma_s \omega) \quad \textit{with probability } 1.$$

In section 2 and in the proof of Theorem 2, we have used another result of reference [3]. For the sake of completeness, we include this result here in the form adapted to the set-up of this paper. Note that in the particular case of a single input system of the form (1) with the matrix A in companion form and state dependent noise, a similar result was established in [23]. Theorem 2 in [23] dealt with an L_2-property of solutions with probability 1 while the result given below concerns the mean-square L_2-property.

Lemma 4 (see Brusin [3, Theorem 5]). *Suppose Assumption* 1 *is satisfied. Then, for any pair of inputs* $(u(\cdot), \xi(\cdot)) \in L_2(s, \mathbf{R}^{m_1}) \times L_2(s, \mathbf{R}^{m_2})$ *and any initial condition* $x(0) = h$, *the corresponding solution to* (1) *satisfies the condition:*

$$|||x(\cdot)|||^2 \leq c_0 \left(\mathbf{E}\|h\|^2 + |||u(\cdot)|||^2 + |||\xi(\cdot)|||^2\right), \tag{A.4}$$

where $c_0 > 0$ *is a constant independent of* h, $u(\cdot)$, *and* $\xi(\cdot)$.

Appendix B. Proof of Lemma 1. The problem (19) is a stochastic control problem with a sign-indefinite integrand in the cost function. Thus, it is natural to refer to the uncertainty inputs $\xi(\cdot)$ as "control" inputs in this control problem. A solution to this class of control problems has been given in [4]; see also [5], where the results of [4] were extended to the infinite-dimensional case. It is readily seen that the conditions of Theorem 1 in [4, 5] are satisfied in the case under consideration. Applying the result of Theorem 1 in [4, 5] to the control problem (19), it follows that the first part of claim (a) of Lemma 1 holds. The existence of a symmetric matrix X_2 satisfying condition (21) is also established by the above-mentioned result of [4, 5]. From the inequality $\Im^{s,h}(0,0) \geq 0$, it follows that $X_2 \geq 0$. Hence, claim (b) also holds.

The control problem

$$\sup_{\zeta \in L_2(s,T;\mathbf{R}^{m_2})} \Im_T^{s,h}(0,\zeta), \tag{B.1}$$

considered in the second part of claim (a) of Lemma 1, can be solved by the same method as that of [2, 11]. Reference [2] presents a solution to the standard finite horizon stochastic optimal control problem. The control problem (B.1) differs from control problems considered in reference [2] in that the integrand in (16) is sign-indefinite. Hence, one needs to extend the results of [2] to the case considered in this paper. This extension can be performed in the same fashion as has been done for similar maximization problems for deterministic time-varying systems [11].

We first note that the control problem (B.1) has a unique solution. Indeed, as in [2], it follows that the functional (16) is continuous. Also, in the same way as in [3, 4, 5], one can obtain using the Riesz representation theorem, that

$$\Im_T^{s,h}(0,\zeta) = -\pi(\zeta,\zeta) + 2\Upsilon(h,\zeta) + \Im_T^{s,h}(0,0),$$

where $\pi(\cdot,\cdot)$, $\Upsilon(\cdot,\cdot)$ are bilinear forms on the Hilbert product-spaces $L_2(s,T;\mathbf{R}^{m_2}) \times L_2(s,T;\mathbf{R}^{m_2})$, $L_2^s \times L_2(s,T;\mathbf{R}^{m_2})$, respectively. We now show that the bilinear form π is coercive [13]. Let $\Phi_{s,T}$ denote a linear bounded operator $L_2(s,T;\mathbf{R}^{m_2}) \to L_2(s,T;\mathbf{R}^n)$ mapping a given control input $\zeta(\cdot)$ into a corresponding solution of the equation

$$dx = (Ax(t) + B_2\zeta(t))dt + Hx(t)dw_1(t) + P_2\zeta(t)dw_2(t), \quad x(s) = 0. \tag{B.2}$$

Then, for any $s_2 \leq s_1 \leq T_1 \leq T_2$ and $\zeta(\cdot) \in L_2(s_1, T_1, \mathbf{R}^{m_2})$,

$$|||R^{1/2}\Phi_{s_1,T_1}\zeta|||^2 = |||R^{1/2}\Phi_{s_2,T_1}\tilde{\zeta}|||^2 \leq |||R^{1/2}\Phi_{s_2,T_2}\tilde{\zeta}|||^2,$$

where $\tilde{\zeta} \in L_2(s_2, T_2; \mathbf{R}^{m_2})$ is the extension of $\zeta(\cdot)$ to $[s_2; T_2]$ by zero. From this inequality and from (18), it follows that

$$-\pi(\zeta,\zeta) = |||R^{1/2}\Phi_{s,T}\zeta|||^2 - |||\zeta|||^2 \leq -\varepsilon_2|||\zeta|||^2 \tag{B.3}$$

for all $0 \leq s < T \leq +\infty$. That is, the form $\pi(\cdot,\cdot)$ is coercive. Then, Theorem 1.1 of [13] implies that the control problem (B.1) has a unique solution which is characterized by the equation

$$(\Phi^*_{s,T}R\Phi_{s,T} - \mathcal{I})\zeta^s_T = -\Phi^*_{s,T}Rx_{s,h},$$

where $x_{s,h}(\cdot)$ is the solution to (13), with the initial condition $x(s) = h$, and $\mathcal{I}$ denotes the identity operator. From (B.3), operators $R^{1/2}\Phi_{s,T}$ and $(\mathcal{I} - \Phi^*_{s,T}R\Phi_{s,T})^{-1}$ are uniformly bounded. Hence, conditions (20) are satisfied. This establishes the second part of claim (a) of Lemma 1.

As in Theorem 6.1 of [2], one can prove that the above facts imply the existence of a unique nonnegative definite solution to the Riccati equation (22). As in [2], we need the following claim.

Claim 1. *Let a symmetric nonnegative matrix $\hat{X}$ be given such that $I - P_2^*\hat{X}P_2 \geq \lambda I$, $\|(I - P_2^*\hat{X}P_2)^{-1}\| \leq 1/\lambda$. For any $\tilde{X}$ such that $\|\tilde{X} - \hat{X}\| \leq \lambda/(2\|P_2\|^2_Q)$, the matrix $I - P_2^*\tilde{X}P_2$ is positive definite and hence is boundedly invertible, and $\|(I - P_2^*\tilde{X}P_2)^{-1}\| \leq 2/\lambda$.*

Proof. The proof of this claim follows from the same arguments as those used in proving the corresponding fact in [2]. □

Let $\alpha > 0$ be a given constant. Consider a constant $\lambda > 0$ and a matrix-valued function $\mathcal{P}(t)$, $t \in [T-\alpha, T]$ such that $0 \leq \mathcal{P}(t) = \mathcal{P}'(t)$, $I - P_2^*\mathcal{P}(t)P_2 \geq \lambda I$ and hence $\sup_t \|(I - P_2^*\mathcal{P}(t)P_2)^{-1}\| \leq 1/\lambda$. As in [2], let us also consider the set of matrix-valued functions

$$\kappa^\alpha_{\mathcal{P}} = \left\{ \tilde{X}(t), t \in [T-\alpha, T]\colon \|\tilde{X} - \mathcal{P}\| \leq \frac{\lambda}{2\|P_2\|^2_Q} \right\}.$$

We define the distance on $\kappa^\alpha_{\mathcal{P}}$ by $\rho(\hat{X}, \tilde{X}) = \sup_{t\in[T-\alpha,T]} \|\tilde{X}(t) - \hat{X}(t)\|$. With this distance, the set $\kappa^\alpha_{\mathcal{P}}$ becomes a closed subset of the Banach space of continuous bounded matrix-valued functions. Let the mapping $\varphi(\cdot)$ be defined as follows:

$$\varphi(Z) \colon = A'Z + ZA + R + H^*ZH + ZB_2(I - P_2^*ZP_2)^{-1}B_2'Z.$$

It follows from Claim 1 that this mapping is well-defined on $\kappa^\alpha_{\mathcal{P}}$. Moreover, since $\mathcal{P}$ has been chosen as described above, it can be proved that the mapping

$$\mathcal{Q}(\tilde{X}(\cdot))(t) \colon = \mathcal{P} + \int_t^T \varphi(\tilde{X}(\theta))d\theta, \qquad \forall\ \tilde{X} \in \kappa^\alpha_{\mathcal{P}},$$

is a contraction on $\kappa^\alpha_{\mathcal{P}}$, provided α is chosen sufficiently small. The reader is referred to [2, Theorem 6.1] for details.

Let $\mathcal{P}(\cdot) = 0$. This choice agrees with the above conditions on the function $\mathcal{P}(\cdot)$. The choice of the constant λ is obvious. From the contraction mapping theorem, the contraction $\mathcal{Q}$ has a unique fixed point in $\kappa^{\alpha}_{\mathcal{P}\equiv 0}$. Denoting this point $X_{2T}(\cdot)$, we observe that $X_{2T}(\cdot)$ satisfies (22) on $[T-\alpha, T]$, and also from Claim 1 for a certain $\lambda_1 > 0$, $I - P_2^* X_{2T} P_2 \geq \lambda_1 I$ and hence is boundedly invertible. As in Lemma 2.2 of [11], this fact implies that

$$\sup_{\zeta \in L_2(s,T,\mathbf{R}^{m_2})} \Im_T^{s,h}(0,\zeta) = E\langle h, X_{2T}(s)h\rangle$$

for any $s \in [T-\alpha, T]$. Also, we see from the above equation that $0 \leq X_{2T}(s) \leq X_2$, and consequently, the operator X_{2T} is uniformly bounded on $[T-\alpha, T]$. Furthermore, it is straightforward to verify, using (22), that the unique optimal control is expressed in the feedback form

$$\xi^s_{2T}(t) = (I - P_2^* X_{2T}(t) P_2)^{-1} B_2' X_{2T} x^s_{2T}(t).$$

Claim (c) of Lemma 1 now follows by iterating the above procedure a finite number of steps, with $\mathcal{P}(t) \equiv X_{2T}(T-(i-1)\alpha)$ at the ith step.

To prove claim (d) of Lemma 1, we note that one can prove in a standard fashion that $X_{2T}(s)$ is monotone increasing in T. Hence, for all $s \geq 0$ there exists a matrix $\bar{X}_2(s) = \lim_{T\to\infty} X_{2T}(s)$. Hence $0 \leq \bar{X}_2(s) \leq X_2 \ \forall \ s \geq 0$. Consequently, $I - P_2^* \bar{X}_2(s) P_2 \geq 0$. Also, letting T approach ∞ in (22), it follows that $\bar{X}_2(s)$ satisfies this equation for all $s \in [0,\infty)$ such that $I - P_2^* \bar{X}_2(s) P_2$ is nonsingular. For all $s \in [0,\infty)$ such that $I - P_2^* \bar{X}_2(s) P_2$ is singular, we must have that either $\frac{d}{ds}\bar{X}_2(s) = \infty$ or $\bar{X}_2(s)$ is not differentiable.

On the other hand, from the second part of claim (a) of Lemma 1, it follows that

$$\Im_T^{s,h}(0, \xi_2^s(\cdot)) \leq \mathbf{E}\langle h, X_{2T}(s)h\rangle.$$

Letting T approach ∞, it follows from this inequality and (19), (21) that

$$\Im^{s,h}(0,\xi_2^s(\cdot)) \leq \mathbf{E}\langle h, \bar{X}_2(s)h\rangle \leq \mathbf{E}\langle h, X_2 h\rangle = \Im^{s,h}(0,\xi_2^s(\cdot)).$$

Hence, $\bar{X}_2(s) = X_2$, and $\frac{d}{ds}\bar{X}_2(s) = 0 \ \forall \ s \in [0,\infty)$. This implies that the matrix $I - P_2^* X_2 P_2$ is nonsingular. That is, condition (26) is satisfied.

Appendix C. Proof of Theorem 1. Note, that condition (28) of Theorem 1 implies that

$$\Im^{0,0}(u,0) \geq \varepsilon_1 \|u(t)\|_2^2 \qquad \forall \ u \in L_2([0,+\infty)\times\Omega; R^{m_1}). \tag{C.1}$$

Conditions (C.1) and (18) are convexity-concavity conditions which guarantee the existence of a unique minimax pair for the cost function (15) on $L_2(s,\mathbf{R}^{m_1}) \times L_2(s,\mathbf{R}^{m_2})$; see [1]. Let $u^s = u^s(h)$, $\xi^s = \xi^s(h)$ denote this saddle point.

The proof of the existence of an operator X satisfying (30) follows the lines of the corresponding result of linear-quadratic stochastic control [4, 5]. As in [4, 5], it can be shown that there exists a self-adjoint operator $\mathcal{X}_s \in \mathcal{L}(L_2^s, L_2^s)$ such that for all $h \in L_2^s$, $V = E\langle h, \mathcal{X}_s h\rangle$. By the Lebesgue–Nikodým theorem [31], it then follows that there exists a weakly measurable mapping $X_s = X_s(\omega)\colon \Omega \to R^{n\times n}$ such that

$$\langle g(\omega), X_s(\omega)h(\omega)\rangle = \langle g(\omega), (\mathcal{X}_s h)(\omega)\rangle \qquad \text{almost surely (a.s.)}$$

for all $h, g \in L_2^s$. Next, we establish that $X_s(\omega) = X_0(\Gamma_s\omega)$ a.s. where $\{\Gamma_s, s \geq 0\}$ is a translation semigroup generated by the Wiener process $(w_1(t), w_2(t))$ (see [6] and Appendix A). Since X_0 is weakly $\mathcal{F}_0$-measurable, then this operator is weakly invariant with respect to Γ_s. This leads to the existence of X satisfying (30).

The proof of the claim that the operator X satisfying (30) also satisfies (29) follows the same arguments as those used in proving Theorem 3.4 of [11]. This proof is based on certain facts of linear-quadratic stochastic control concerning the existence of nonnegative definite solutions to the generalized Riccati equation

$$\text{(C.2)}\quad \begin{aligned} &\frac{dX_{1T}}{dt} + A'X_{1T} + X_{1T}A + H^*X_{1T}H + R - (X_{1T}B_1 + H^*X_{1T}P_1 + Q) \\ &\qquad\qquad \times (G + P_1^*X_{1T}P_1)^{-1}(X_{1T}B_1 + H^*X_{1T}P_1 + Q)' = 0, \\ &X_{1T}(T) = 0, \end{aligned}$$

and the Riccati equation (22). Note that under the conditions of the theorem, the existence of a nonnegative definite solution to (22) is guaranteed by Lemma 1.

Equation (C.2) is a Riccati differential equation corresponding to the following standard stochastic control problem

$$\text{(C.3)}\quad \inf_{u\in L_2([s,T]\times\Omega;R^{m_1})} \Im_T^{s,h}(u,0).$$

For the particular case $Q = 0$, the solution to the control problem (C.3) can be found, e.g., in references [2, 27]. Note that under condition (28) of the theorem, the results of [2] are readily extended to the case $Q \neq 0$. The extension of Theorem 6.1 of [2] to the case $Q \neq 0$ implies that there exists a unique symmetric nonnegative definite bounded solution $X_{1T}(t)$ to (C.2) and that the feedback law $u_{1T} = -(G + P_1^*X_{1T}P_1)^{-1}(B_1'X_{1T} + P_1^*X_{1T}H + Q')x$ solves the control problem (C.3).

Now let us consider the stochastic differential game associated with (1) and the cost functional (16):

$$\text{(C.4)}\quad \inf_{u(\cdot)\in L_2(s,T,\mathbf{R}^{m_1})} \sup_{\xi\in L_2(s,T,\mathbf{R}^{m_2})} \Im_T^{s,h}(u,\xi).$$

Under conditions (28), (18), stochastic counterparts to Theorems 3.1 and 3.2 of [11] can be established.

Claim 2. *If conditions* (28), (18) *are satisfied, then the game problem* (C.4) *has a unique saddle point* (u_T, ξ_T). *Furthermore, there exists a unique nonnegative definite solution to the Riccati equation*

$$\text{(C.5)}\quad \begin{aligned} &\frac{dX_T(s)}{ds} + A'X_T + X_TA + H^*X_TH + R - (X_TB_1 + H^*X_TP_1 + Q) \\ &\qquad\qquad \times (G + P_1^*X_TP_1)^{-1}(X_TB_1 + H^*X_TP_1 + Q)' \\ &\qquad\qquad + X_TB_2(I - P_2^*X_TP_2)^{-1}B_2'X_T = 0, \\ &X_T(T) = 0, \end{aligned}$$

and the saddle point of the game (C.4) *is characterized by the feedback law*

$$\text{(C.6)}\quad u_T = F_{1T}x, \quad F_{1T} := -(G + P_1^*X_TP_1)^{-1}(B_1'X_T + P_1^*X_TH + Q'),$$

$$\text{(C.7)}\quad \xi_T = F_{2T}x, \quad F_{2T} := (I - P_2^*X_TP_2)^{-1}B_2'X_T.$$

Proof of Claim 2. The existence of a unique saddle point follows from the same concavity-convexity arguments as those used above and can be proved using the result of [1].

Suppose that a solution X_T to (C.5) exists on the interval $[T-\alpha, T]$. Let $s \in [T-\alpha, T]$ be given. Then, applying the Ito formula to the quadratic form $x(t)'X_T(t)x(t)$, where $x(t)$ satisfies (1), we obtain

$$\begin{aligned}\Im_T^{s,h}(u,\xi) = \mathbf{E}h'X_T(s)h + \mathbf{E}\int_s^T \|(G+P_1^*X_TP_1)^{1/2}\,(u(t)-F_{1T}x(t))\,\|^2dt \\ -\mathbf{E}\int_s^T \|(I-P_2^*X_TP_2)^{1/2}\,(\xi(t)-F_{2T}x(t))\,\|^2dt,\end{aligned}$$

where the matrices F_{1T} and F_{2T} are defined by (C.6), (C.7). Hence, we conclude that the pair (C.6), (C.7) satisfies the saddle point condition:

$$\Im_T^{s,h}(u_T,\xi) \le \Im_T^{s,h}(u_T,\xi_T) = \mathbf{E}h'X_T(s)h \le \Im_T^{s,h}(u,\xi_T). \tag{C.8}$$

It remains to prove that (C.5) has a solution as required. We can prove this claim in the same fashion as Lemma 1 by choosing α sufficiently small. Indeed for any $\mathcal{P}$ chosen as in the proof of Lemma 1, one can consider the closed set $\kappa_{\mathcal{P}}^{\alpha}$. Note that since $G + P_1^*\mathcal{P}P_1 \ge G \ge \bar{\lambda}I$, then $\|(G+P_1^*\mathcal{P}P_1)^{-1}\| \le 1/\bar{\lambda}$. As in Claim 1 (see also [2, Theorem 6.1]), we have that $\|(G+P_1^*\tilde{X}P_1)^{-1}\| \le 2/\bar{\lambda}$ for any $\tilde{X} \in \kappa_{\mathcal{P}}^{\alpha}$. This observation allows us to conclude that the mapping

$$\begin{aligned}\tilde{\varphi}(Z) := A'Z + ZA + R + H^*ZH + ZB_2(I-P_2^*ZP_2)^{-1}B_2'Z \\ -(ZB_1 + H^*ZP_1 + Q)(G+P_1^*ZP_1)^{-1}(B_1'Z + P_1^*ZH + Q')\end{aligned}$$

is well defined on $\kappa_{\mathcal{P}}^{\alpha}$ and that the mapping

$$\tilde{\mathcal{Q}}(\tilde{X}(\cdot))(t) := \mathcal{P} + \int_t^T \tilde{\varphi}(\tilde{X}(\theta))d\theta$$

is a contraction on $\kappa_{\mathcal{P}}^{\alpha}$ provided α is chosen sufficiently small. Then, fixed point arguments lead us to the conclusion that there exists a bounded solution $X_T(t)$ to (C.5) on $[T-\alpha, T]$, and also $I - P_2^*X_T(t)P_2 > 0$. Using (C.8), with this solution we have that $0 \le X_{1T}(s) \le X_T(s) \le X_{2T}(s) \le X_2 \ \forall \ s \in [T-\alpha, T]$. Thus, by partitioning the interval $[0, T]$ into subintervals not longer than α and iterating the above fixed point procedure, we can show the existence of a unique global solution to the Riccati equation (C.5). Since the considered game (C.4) has the unique saddle point and from (C.8), this saddle point is given by (C.6), (C.7). This completes the proof of Claim 2.

The remainder of the proof of Theorem 1 follows using arguments similar to those used in proving Theorem 3.4 in [11]. Consider a pair of inputs (C.6) and (C.7) extended to $[T, +\infty)$ by zero. This pair is again denoted by (u_T, ξ_T). By optimality we have

$$\Im_T^{s,h}(u_T,\xi^s) \le \Im_T^{s,h}(u_T,\xi_T) \le \Im_T^{s,h}(u^s,\xi_T) \le \Im^{s,h}(u^s,\xi^s). \tag{C.9}$$

Also, we have already shown that $X_T(s) \le X_2$. This bound on $X_T(s)$ holds for any $s < T$. On the other hand, it is readily seen that $X_T(s)$ is monotone increasing in T. Hence, for all $s \ge 0$, there exists $\bar{X}(s) = \lim_{T\to\infty} X_T(s)$. Consequently, $\bar{X}(s) \le X_2 \ \forall \ s \ge 0$ and this inequality and inequality (25) from Lemma 1 imply that $I - P_2^* \bar{X}(s) P_2 > 0$. This in turn implies that there exist bounded limits for $F_{1T}(s)$, $F_{2T}(s)$ as $T \to \infty$. Let $\bar{F}_1(s)$ and $\bar{F}_2(s)$ denote these limits, respectively. In the same fashion as in [11], we obtain that for any $T_0 > s$, $x_T(\cdot) \to \bar{x}(\cdot)$ strongly in $C([s, T_0], L_2(\Omega, P, R^n))$, $u_T(\cdot) \to \bar{u}(\cdot) = \bar{F}_1(\cdot)\bar{x}(\cdot)$ strongly in $C([s, T_0], L_2(\Omega, P, R^{m_1}))$, and $\xi_T(\cdot) \to \bar{\xi}(\cdot) = \bar{F}_2(\cdot)\bar{x}(\cdot)$ strongly in $C([s, T_0], L_2(\Omega, P, R^{m_2}))$, where $x_T(\cdot)$ is a solution to (1), driven by the pair u_T, ξ_T, and $\bar{x}(\cdot)$ is a solution to the equation

$$dx = (A + B_1\bar{F}_1(t) + B_2\bar{F}_2(t))xdt + (H + P_1\bar{F}_1(t))xdw_1(t) + P_2\bar{F}_1(t)xdw_2(t), \quad x(s) = h.$$

Also as in [11], it can be proved that the sequences $\{u_T\}$, $\{\xi_T\}$ are bounded in $L_2(s, \mathbf{R}^{m_1})$, $L_2(s, \mathbf{R}^{m_2})$, respectively. Thus, one can extract subsequences denoted again by $\{u_T\}$, $\{\xi_T\}$ such that $(u_T, \xi_T) \to (\tilde{u}, \tilde{\xi}) \in L_2(s, \mathbf{R}^{m_1}) \times L_2(s, \mathbf{R}^{m_2})$ weakly as $T \to \infty$. This leads us to the conclusion that $\bar{u}(\cdot) = \tilde{u}(\cdot)$, $\bar{\xi}(\cdot) = \tilde{\xi}(\cdot)$, which implies that $(\bar{u}, \bar{\xi}) \in L_2(s, \mathbf{R}^{m_1}) \times L_2(s, \mathbf{R}^{m_2})$.

Letting T approach ∞ in (C.9), we obtain

$$\Im^{s,h}(\bar{u}, \xi^s) \leq E\langle h, \bar{X}(s)h\rangle \leq \Im^{s,h}(u^s, \bar{\xi}) \leq \Im^{s,h}(u^s, \xi^s) \leq \Im^{s,h}(\bar{u}, \xi^s).$$

Thus, we have that $E\langle h, \bar{X}(s)h\rangle = \Im^{s,h}(u^s, \xi^s) = E\langle h, Xh\rangle$ and therefore

$$(u^s, \xi^s) = (\bar{u}, \bar{\xi}).$$

That is, the feedback representation of the saddle point holds. To see that the matrix X satisfies (29) one needs to pass to the limit as $T \to \infty$ in (C.5). Also, we see that $X = \lim_{T\to\infty} X_T$ is a minimal solution to (29) and this solution satisfies condition (30). It is clear that the solution satisfying these conditions is unique.

The claim that the system (31) is exponentially mean-square stable follows from the fact that for linear stochastic systems, stochastic L_2-stability is equivalent to stochastic exponential mean-square stability; see, e.g., [10].

Appendix D. Proof of Theorem 2. With the substitution

$$v = u + (D'D)^{-1}D'Cx \tag{D.1}$$

into (1), the stochastic system under consideration becomes

$$\begin{aligned} dx &= (\tilde{A}x(t) + B_1 v(t) + B_2\xi(t))dt + (\tilde{H}x(t) + P_1 v(t))dw_1(t) \\ &\quad + P_2\xi(t)dw_2(t), \\ \tilde{z} &= \tilde{C}x + \tilde{D}v. \end{aligned} \tag{D.2}$$

Here the following notation is used:

$$\begin{aligned} &\tilde{A} = A - B_1(D'D)^{-1}D'C, \qquad \tilde{H} = H - P_1(D'D)^{-1}D'C, \\ &\tilde{C} = \begin{bmatrix} \tilde{R}^{1/2} \\ 0 \end{bmatrix}, \qquad \tilde{D} = \begin{bmatrix} 0 \\ D \end{bmatrix}, \end{aligned} \tag{D.3}$$

and the matrix $\tilde{R}$ is as defined in Assumption 2. Note that if the matrix K is a solution to the stochastic H_∞ problem defined in section 3, then $\tilde{K} = K + (D'D)^{-1}D'C$ solves a corresponding stochastic H_∞ problem associated with the system (D.2), and vice versa. The stochastic H_∞ control problem associated with the system (D.2) is defined in the same fashion as the original H_∞ control problem; i.e., given the system (D.2), find a matrix $\tilde{K} \in R^{m_1 \times n}$ such that the state feedback controller $v = \tilde{K}x$ satisfies the following conditions:

(i$'$) The system

$$dx = (\tilde{A} + B_1\tilde{K})xdt + (\tilde{H} + P_1\tilde{K})xdw_1(t) \tag{D.4}$$

is exponentially stable in the mean-square sense and the matrix $\tilde{A} + B_1\tilde{K}$ is stable.

(ii$'$) The closed-loop system corresponding to system (D.2) with feedback control $v = \tilde{K}x$,

$$dx = [(\tilde{A} + B_1\tilde{K})x(t) + B_2\xi(t)]dt + (\tilde{H} + P_1\tilde{K})x(t)dw_1(t) + P_2\xi(t)dw_2(t), \tag{D.5}$$

satisfies the following stochastic H_∞-norm condition: there exists a constant $\tilde{\varepsilon} > 0$ such that

$$E\int_s^{+\infty} \left(\|(\tilde{C} + \tilde{D}\tilde{K})x(t)\|^2 - \|\xi(t)\|^2\right)dt \le -\tilde{\varepsilon}\mathbf{E}\int_s^{+\infty} \|\xi(t)\|^2 dt \quad \text{for } x(s) = 0 \tag{D.6}$$

for each $\xi \in L_2(s; \mathbf{R}^{m_2})$.

Note that the problem (i$'$), (ii$'$) is simpler than the original problem because we have $\tilde{C}'\tilde{D} = 0$, $\tilde{D}'\tilde{D} = D'D > 0$ in this case. Also by Assumption 2, there exists a matrix N such that the matrix $\tilde{A} - N\tilde{C}$ is a Hurwitz matrix and

$$\|\exp((\tilde{A} - N\tilde{C})t)\| \le ae^{-\alpha t}, \qquad \frac{a^2}{\alpha}\|\tilde{H}\|^2_{Q_1} < 1. \tag{D.7}$$

This observation implies that the system

$$dy = (\tilde{A} - N\tilde{C})ydt + \tilde{H}ydw_1(t), \qquad y(s) = y_0 \in L_2^s, \tag{D.8}$$

is exponentially mean-square stable. In the particular case where $\tilde{H} = 0$, this fact amounts to the pair $(\tilde{C}, \tilde{A})$ being detectable.

First, we solve the problem (i$'$), (ii$'$).

Lemma 5.

(a) *If the stochastic H_∞ state feedback control problem* (i$'$), (ii$'$) *has a solution, then there exists a symmetric nonnegative definite minimal solution X to the generalized algebraic Riccati equation*

$$\begin{aligned} &\tilde{A}'X + X\tilde{A} + \tilde{H}^*X\tilde{H} + \tilde{C}'\tilde{C} \\ &-(XB_1 + \tilde{H}^*XP_1)(\tilde{D}'\tilde{D} + P_1^*XP_1)^{-1}(B_1'X + P_1^*X\tilde{H}) \\ &+XB_2(I - P_2^*XP_2)^{-1}B_2'X = 0, \end{aligned} \tag{D.9}$$

such that $I - P_2^*XP_2 > 0$ and the stochastic system

$$dx = (\tilde{A} + B_1\tilde{F}_1 + B_2\tilde{F}_2)xdt + (H + P_1\tilde{F}_1)xdw_1(t) + P_2\tilde{F}_2xdw_2(t), \tag{D.10}$$

where

$$\begin{aligned} \tilde{F}_1 &= -(\tilde{D}'\tilde{D} + P_1^*XP_1)^{-1}(B_1'X + P_1^*X\tilde{H}), \\ \tilde{F}_2 &= (I - P_2^*XP_2)^{-1}B_2'Xx, \end{aligned} \tag{D.11}$$

is exponentially stable in the mean-square sense. Also,

$$\mathbf{E}\langle h, Xh\rangle \le \sup_{\xi \in L_2(s;\mathbf{R}^{m_2})} E\int_s^{+\infty} \left(\|\tilde{C}x(t) + \tilde{D}\hat{v}(t)\|^2 - \|\xi(t)\|^2\right)dt \tag{D.12}$$

for any state feedback control $\hat{v}(\cdot)$ such that the closed-loop system corresponding to (D.2) *and this control has L_2-summable solutions for all $\xi \in L_2(s; \mathbf{R}^{m_2})$, and the supremum on the left-hand side of* (D.12) *is finite.*

(b) *Conversely, if there exists a symmetric nonnegative definite solution X to* (D.9) *such that $I - P_2^* X P_2 > 0$ and the stochastic system* (D.10) *is exponentially stable, then the stochastic H_∞ state feedback control problem* (i′), (ii′) *has a solution. The corresponding stabilizing feedback controller which solves this stochastic H_∞ problem is given by*

$$\tilde{K} = -(\tilde{D}'\tilde{D} + P_1^* X P_1)^{-1}(B_1' X + P_1^* X \tilde{H}). \tag{D.13}$$

Proof of Lemma 5. We shall use the following notation: $\tilde{A}_K = \tilde{A} + B_1\tilde{K}$, $\tilde{C}_K = \tilde{C} + \tilde{D}\tilde{K}$, $\tilde{H}_K = \tilde{H} + P_1\tilde{K}$.

(b)⇒(a). Let X be a symmetric nonnegative solution to (D.9) such that (D.10) is exponentially mean-square stable. Let $\tilde{K}$ be given by (D.13). We wish to establish that X and $\tilde{K}$ satisfy conditions (i′), (ii′).

Let us prove the stability of the system (D.4). Note that using (D.13), (D.9) can be transformed as follows:

$$\begin{aligned}\tilde{A}_K' X + X\tilde{A}_K + \tilde{H}_K^* X \tilde{H}_K + \tilde{C}'\tilde{C} + \tilde{K}'(\tilde{D}'\tilde{D})\tilde{K}& \\ + X B_2 (I - P_2^* X P_2)^{-1} B_2' X = 0.&\end{aligned} \tag{D.14}$$

This implies that

$$\tilde{A}_K' X + X\tilde{A}_K + \tilde{H}_K^* X \tilde{H}_K + \tilde{C}'\tilde{C} + \tilde{K}'(\tilde{D}'\tilde{D})\tilde{K} \le 0. \tag{D.15}$$

We now proceed as in [10, Lemma 4.6]. Letting $x(t)$ be the solution to (D.4), first we note that from (D.15),

$$\mathbf{E}\int_0^\infty \left(\|\tilde{C}x(t)\|^2 + \|\tilde{D}\tilde{K}x(t)\|^2\right) dt \le \mathbf{E} h' X h < \infty,$$

and hence $\tilde{C}x(\cdot)$ and $\tilde{K}x(\cdot)$ are square integrable (the latter holds since $\tilde{D}'\tilde{D} > 0$). Next, note that $\|\tilde{H} + P_1\tilde{K}\|_{Q_1}^2 \le (1+\nu)\|\tilde{H}\|_{Q_1}^2 + (1+\frac{1}{\nu})\|P_1\tilde{K}\|_{Q_1}^2$ for any $\nu > 0$. From (D.7), ν can be chosen sufficiently small in order to guarantee that

$$\frac{(1+\nu)a^2}{\alpha}\|\tilde{H}\|_{Q_1}^2 < 1. \tag{D.16}$$

This observation allows us to obtain in the same way as in [10, Lemma 4.6] that $\mathbf{E}\|x(t)\|^2 \in L_1[0,\infty)$. This implies that (i′) holds.

We will now prove that condition (ii′) is also satisfied. From (D.14), we have that

$$\begin{aligned}&\mathbf{E}\langle x(t), X x(t)\rangle + \mathbf{E}\int_0^t \{\|\tilde{C}_K x(t)\|^2 - \|\xi\|^2\} dt \\ &\quad = -\mathbf{E}\int_0^t \|(I - P_2^* X P_2)^{1/2}(\xi - (I - P_2^* X P_2)^{-1} B_2' X x)\|^2 dt,\end{aligned} \tag{D.17}$$

where $x(\cdot)$ is the solution to (D.5) corresponding to the initial condition $x(0) = 0$.

Note, that the substitution $\xi = \zeta + (I - P_2^* X P_2)^{-1} B_2' X x$ into (D.5) leads to the following equation:

$$\begin{aligned} dx &= ((\tilde{A}_K + B_2(I - P_2^* X P_2)^{-1} B_2' X)x + B_2\zeta(t))dt + \tilde{H}_K x dw_1(t) \\ &\qquad + P_2(\zeta + (I - P_2^* X P_2)^{-1} B_2' X x) dw_2, \\ x(0) &= 0. \end{aligned} \tag{D.18}$$

In particular, the input $\zeta(t) = 0$ corresponds to the stable system (D.10). This implies (see, e.g., [3] and also Lemma 4 in Appendix A) that solutions of (D.18) satisfy the condition $|||x||| \le c_0|||\zeta|||$, where $c_0 > 0$ is a constant independent of ζ. That is, the mapping $\zeta(\cdot) \to x(\cdot)$ and, consequently, the mapping $\zeta(\cdot) \to \xi(\cdot) = \zeta(\cdot) + (I - P_2XP_2)^{-1}B_2'Xx(\cdot)$ generated by (D.18) are bounded mappings $L_2(0;\mathbf{R}^{m_2}) \to L_2(0;\mathbf{R}^n)$, $L_2(0;\mathbf{R}^{m_2}) \to L_2(0;\mathbf{R}^{m_2})$, respectively. Thus, there exists a constant $c > 0$ such that

$$|||\xi(\cdot)||| \le c|||\zeta(\cdot)||| \quad \forall\, \zeta \in L_2(0;\mathbf{R}^{m_2}).$$

Note that the restriction of any solution of (D.18) to an interval $[0,t]$ is equal to the restriction of a corresponding solution of (D.5). Since $X \ge 0$ and $I - P_2^*XP_2 > 0$, then (D.17) implies that

$$\mathbf{E}\int_0^\infty \{\|(\tilde{C} + \tilde{D}\tilde{K})x(t)\|^2 - \|\xi\|^2\}dt \le -\epsilon\mathbf{E}\int_0^\infty \|\zeta\|^2dt \le -\frac{\epsilon}{c^2}\mathbf{E}\int_0^\infty \|\xi\|^2dt.$$

(a)⇒(b). Given a matrix $\tilde{K}$ satisfying conditions (i′) and (ii′), we wish to prove that there exists a solution to (D.9) stabilizing the system (D.10).

Let $\delta \in (0, \bar{\delta}]$ be a given constant. Consider the stochastic differential game of the form (17) associated with the system

$$\text{(D.19)} \qquad \begin{aligned} &dx = (\tilde{A}_Kx + B_1v + B_2\xi)dt + (\tilde{H}_Kx + P_1v)dw_1(t) + P_2\xi dw_2(t),\\ &\tilde{z} = \tilde{C}_Kx + \tilde{D}v \end{aligned}$$

and cost functional

$$\text{(D.20)} \qquad \Im_\delta^{s,h}(v,\xi) = \int_s^{+\infty} \mathbf{E}\left\{\|\tilde{C}_Kx(t) + \tilde{D}v(t)\|^2 + \delta\|v(t)\|^2 - \|\xi(t)\|^2\right\}dt.$$

By the condition (i′), (D.4) is stable, and hence the system (D.19) satisfies Assumption 1. Also, it follows from (ii′) that the system (D.5) satisfies condition (18) of Lemma 1 and Theorem 1 with $\varepsilon_2 = \tilde{\varepsilon}$. In particular, Lemma 1 defines a matrix X_2 such that $I - P_2^*X_2P_2 > 0$. Note that this matrix is independent of δ.

Condition (28) of Theorem 1 is also satisfied with $\varepsilon_1 = \delta$. It follows from this theorem that for each $\delta \in (0, \bar{\delta}]$, the equation

$$\text{(D.21)} \qquad \begin{aligned} \tilde{A}_K'X + X\tilde{A}_K + \tilde{H}_K^*X\tilde{H}_K + \tilde{C}_K\tilde{C}_K &- (XB_1 + \tilde{H}_K^*XP_1 + \tilde{C}_K'\tilde{D})\\ &\times (\tilde{D}'\tilde{D} + \delta I + P_1^*XP_1)^{-1}(B_1'X + P_1^*X\tilde{H}_K + \tilde{D}'\tilde{C}_K)\\ &+ XB_2(I - P_2^*XP_2)^{-1}B_2'X = 0 \end{aligned}$$

has a symmetric nonnegative definite minimal solution X^δ such that the system

$$\text{(D.22)} \qquad \begin{aligned} dx = (\tilde{A}_K + B_1F_{1,\delta} + B_2F_{2,\delta})xdt &+ (\tilde{H}_K + P_1F_{1,\delta})xdw_1(t)\\ &+P_2F_{2,\delta}xdw_2(t) \end{aligned}$$

is exponentially mean-square stable. In (D.22),

$$\begin{aligned} &F_{1,\delta} = -(\tilde{D}'\tilde{D} + \delta I + P_1^*X^\delta P_1)^{-1}(B_1'X^\delta + P_1^*X^\delta\tilde{H}_K + \tilde{C}_K'\tilde{D}),\\ &F_{2,\delta} = (I - P_2^*X^\delta P_2)^{-1}B_2'X^\delta. \end{aligned}$$

As we have shown when proving Theorem 1, $X^\delta \le X_2$ and hence $I - P_2^*X^\delta P_2 > 0$.

As in reference [11, Theorem 4.2], it follows that there exists a matrix $X{:}=\lim_{\delta\downarrow 0} X^\delta$, $X=X'\geq 0$ which satisfies (D.9). To verify this fact, one must take into account the fact that $\tilde{C}'\tilde{D}=0$. Also, we note that since X_2 is independent of δ, $X\leq X_2$ and hence $I-P_2^*XP_2>0$. Next, letting $\delta\downarrow 0$ in (D.21), we obtain that the above defined matrix X satisfies (D.9). Also, there exist the limits

$$\begin{aligned}F_{1,K}&=\lim_{\delta\downarrow 0}F_{1,\delta}=-\tilde{K}-(\tilde{D}'\tilde{D}+P_1^*XP_1)^{-1}(B_1'X+P_1'X\tilde{H})=\tilde{F}_1-\tilde{K},\\ F_{2,K}&=\lim_{\delta\downarrow 0}F_{2,\delta}=(I-P_2^*XP_2)^{-1}B_2'X=\tilde{F}_2.\end{aligned}$$

Note that X is defined as the limit of a sequence of minimal solutions, hence it represents the minimal solution to (D.9).

We now show that the system (D.10) is exponentially mean-square stable. Consider solutions $x_\delta(\cdot)$ and $x(\cdot)$ to (D.22) and (D.10), respectively, that correspond to the initial condition $x_\delta(s)=x(s)=h$. Note that

$$\text{(D.23)}\qquad \begin{aligned}0\leq \Im_\delta^{s,h}(F_{1,\delta}x_\delta,0)\leq \mathbf{E}h'X^\delta h&=\Im_\delta^{s,h}(F_{1,\delta}x_\delta,F_{2,\delta}x_\delta)\\ &\leq \Im^{s,h}(0,F_{2,\delta}x_\delta)\leq \mathbf{E}h'X_2h,\end{aligned}$$

where the matrix X_2 has been defined above. Hence, $F_{2,\delta}x_\delta(\cdot)$ is bounded. This fact is established in the same fashion as that in the proof of Theorem 1. Furthermore, from (D.23), $\Im_\delta^{s,h}(F_{1,\delta}x_\delta,F_{2,\delta}x_\delta)$ is bounded. Since $F_{2,\delta}x_\delta(\cdot)$ is bounded, this then implies that $\tilde{C}x_\delta(\cdot)$ and $(\tilde{K}+F_{1,\delta})x_\delta(\cdot)$ are bounded in the corresponding space $L_2[s,\infty)$. Again, we have used the fact that $\tilde{C}'\tilde{D}=0$ and $\tilde{D}'\tilde{D}>0$ to reach this conclusion. Hence, one can extract subsequences which have weak limit points in corresponding L_2-spaces. These limits are L_2-summable functions on $[s,+\infty)\times\Omega$. Also, as in the proof of Theorem 1, for any $T_0>s$, $x_\delta(\cdot)\to x(\cdot)$ in $C([s,T_0];L_2(\Omega,P,R^n))$. Therefore, for any $T_0>s$, the restrictions of the functions $\tilde{C}x(\cdot)$, $(\tilde{K}+F_{1,K})x(\cdot)=\tilde{F}_1x(\cdot)$, and $\tilde{F}_2x(\cdot)$ to the interval $[s,T_0]$ are equal to restrictions of the corresponding weak limit points. Hence $\tilde{C}x(\cdot)$, $\tilde{F}_1x(\cdot)$, $\tilde{F}_2x(\cdot)$ are square-integrable functions on $[s,+\infty)\times\Omega$.

Let us rewrite (D.10) in the following form:

$$\begin{aligned}dx=((\tilde{A}-N\tilde{C})x+B_1\tilde{F}_1x(t)+B_2\tilde{F}_2x(t)+N\tilde{C}x(t))dt+(\tilde{H}+P_1\tilde{F}_1)xdw_1(t)\\ +P_2\tilde{F}_2xdw_2(t),\end{aligned}$$

where N is the matrix defined in Assumption 2. Since (D.8) is stable and inputs $\tilde{F}_1x(\cdot)$, $\tilde{F}_2x(\cdot)$, and $\tilde{C}x(\cdot)$ are square integrable on $[s,\infty)\times\Omega$, then $x(\cdot)\in L_2(s;\mathbf{R}^n)$. Hence, (D.10) is exponentially mean-square stable.

To see that (D.12) holds, let us consider the finite horizon version of the functional (D.20):

$$\Im_{\delta,T}^{s,h}(v,\xi)=\int_s^T\mathbf{E}\left\{\|\tilde{C}_Kx(t)+\tilde{D}v(t)\|^2+\delta\|v(t)\|^2-\|\xi(t)\|^2\right\}dt,$$

with $v\in L_2(s,T;\mathbf{R}^{m_1})$, $\xi\in L_2(s,T;\mathbf{R}^{m_2})$. As in the infinite horizon case, we obtain from Claim 2 (cf. [11, Theorem 4.1]), that there exists a matrix $X_T=\lim_{\delta\downarrow 0}X_T^\delta$, satisfying the matrix Riccati equation

$$\text{(D.24)}\qquad \frac{dX_T(s)}{ds}+\tilde{A}_K'X_T+X_T\tilde{A}_K+\tilde{H}_K^*X_T\tilde{H}_K+\tilde{C}_K'\tilde{C}_K$$

$$\begin{aligned}
&\quad - (X_T B_1 + \tilde{H}_K^* X_T P_1 + \tilde{C}_K' \tilde{D})(\tilde{D}'\tilde{D} + P_1^* X_T P_1)^{-1} \\
&\quad \times (X_T B_1 + \tilde{H}_K^* X_T P_1 + \tilde{C}_K' \tilde{D})' \\
&\quad + X_T B_2 (I - P_2^* X_T P_2)^{-1} B_2' X_T = 0, \\
&X_T(T) = 0.
\end{aligned}$$

Let $\hat{v}(\cdot)$ be any state feedback controller such that the closed-loop system obtained from (D.2) has L_2-summable trajectories for all L_2-summable uncertainty inputs. Then, the corresponding closed-loop system obtained from (D.19), with state feedback $v(\cdot) = \hat{v}(\cdot) - \tilde{K}x$ has also L_2-summable trajectories for all L_2-summable uncertainty inputs. Hence, if $\xi \in L_2(s, T; \mathbf{R}^{m_2})$ and $\tilde{\xi}(\cdot)$ denotes its extension to the interval $[T, \infty)$ by zero, then (D.24) implies

$$\begin{aligned}
\mathbf{E}\langle h, X_T(s)h\rangle &\leq \sup_{\xi \in L_2(s,T;\mathbf{R}^{m_2})} \int_s^T \mathbf{E}\left\{\|\tilde{C}_K x(t) + \tilde{D}v(t)\|^2 - \|\xi(t)\|^2\right\} dt \\
&\leq \sup_{\xi \in L_2(s,T;\mathbf{R}^{m_2})} \int_s^\infty \mathbf{E}\left\{\|\tilde{C}_K x(t) + \tilde{D}v(t)\|^2 - \|\tilde{\xi}(t)\|^2\right\} dt \\
&\leq \sup_{\xi \in L_2(s;\mathbf{R}^{m_2})} \int_s^\infty \mathbf{E}\left\{\|\tilde{C}_K x(t) + \tilde{D}v(t)\|^2 - \|\xi(t)\|^2\right\} dt \\
&= \sup_{\xi \in L_2(s;\mathbf{R}^{m_2})} \int_s^\infty \mathbf{E}\left\{\|\tilde{C} x(t) + \tilde{D}\hat{v}(t)\|^2 - \|\xi(t)\|^2\right\} dt.
\end{aligned}$$

Also, note that $X_T \to X$ as $T \to \infty$, since X is a minimal solution. Thus, (D.12) holds. □

We now are in a position to finish the proof of Theorem 2. One can observe that letting

$$K = \tilde{K} - (D'D)^{-1} D'C$$

and using the notation (D.3), (D.9) and (D.10) are transformed into (37) and (38), respectively, and the feedback matrix (D.13) is transformed into the matrix (39). Hence the theorem follows from Lemma 5. □

REFERENCES

[1] A. Bensoussan, *Saddle point of convex concave functionals with applications to linear quadratic differential games*, in Differential Games and Related Topics, H. W. Kuhn and G. P. Szegö, eds., North-Holland, Amsterdam, 1971, pp. 177–199.

[2] J. M. Bismut, *Linear quadratic optimal stochastic control with random coefficients*, SIAM J. Control Optim., 14 (1976), pp. 419–444.

[3] V. A. Brusin, *Global stability and dichotomy of a class of nonlinear systems with stochastic parameters*, Sibirsk. Math J., 22 (1981), pp. 57–73 (in Russian).

[4] V. A. Brusin and V. A. Ugrinovskii, *Stochastic stability of a class of nonlinear differential equations of Ito type*, Siberian Math. J, 28 (1987), pp. 381–393.

[5] V. A. Brusin and V. A. Ugrinovskii, *Absolute stability approach to stochastic stability of infinite dimensional nonlinear systems*, Automatica J. IFAC, 31 (1995), pp. 381–393.

[6] J. L. Doob, *Stochastic Processes*, John Wiley, New York, 1953.

[7] A. El Bouhtouri and A. J. Pritchard, *Stability radii of linear systems with respect to stochastic perturbations*, Systems Control Lett., 19 (1992), pp. 29–33.

[8] L. El Ghaoui, *State-feedback control of systems with multiplicative noise via linear matrix inequalities*, Systems Control Lett., 24 (1995), pp. 223–228.

[9] D. Hinrichsen and A. J. Pritchard, *Stability radii of systems with stochastic uncertainty and their optimization via output feedback*, SIAM J. Control. Optim., 34 (1996), pp. 1972–1998.

[10] A. Ichikawa, *Dynamic programming approach to stochastic evolution equations*, SIAM J. Control Optim., 17 (1979), pp. 152–174.

[11] A. Ichikawa, *Quadratic games and H_∞-type problems for time varying systems*, Internat. J. Control, 54 (1991), pp. 1249–1271.

[12] P. P. Khargonekar, I. R. Petersen, and K. Zhou, *Robust stabilization of uncertain linear systems: Quadratic stabilizability and H_∞ control theory*, IEEE Trans. Automat. Control, 35 (1990), pp. 356–361.

[13] J. L. Lions, *Contrôlle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod Gauthier-Villars, Paris, 1968.

[14] H. P. McKean, *Stochastic Integrals*, Academic Press, New York, 1969.

[15] A. Megretsky and S. Treil, *Power distribution inequalities in optimization and robustness of uncertain systems*, J. Math. Systems Estim. Control, 3 (1993), pp. 301–319.

[16] I. R. Petersen and M. R. James, *Performance analysis and controller synthesis for nonlinear systems with stochastic uncertainty constraints*, Automatica J. IFAC, 32 (1996), pp. 959–972.

[17] I. R. Petersen and D. C. McFarlane, *Optimal guaranteed cost control and filtering for uncertain linear systems*, IEEE Trans. Autom. Control, 39 (1994), pp. 1971–1977.

[18] S. Richter, A. S. Hodel, and P. Pruett, *Homotopy methods for the solution of general modified algebraic Riccati equations*, IEE Proc. Part D Control Theory Appl., 160 (1993), pp. 449–454.

[19] A. V. Savkin and I. R. Petersen, *A method for robust stabilization related to the Popov stability criterion*, Internat. J. Control, 62 (1995), pp. 1105–1115.

[20] A. V. Savkin and I. R. Petersen, *Minimax optimal control of uncertain systems with structured uncertainty*, Internat. J. Robust Nonlinear Control, 5 (1995), pp. 119–137.

[21] A. V. Savkin and I. R. Petersen, *Nonlinear versus linear control in the absolute stabilizability of uncertain systems with structured uncertainty*, IEEE Trans. Autom. Control, 40 (1995), pp. 122–127.

[22] A. V. Savkin and I. R. Petersen, *Robust H^∞ control of uncertain systems with structured uncertainty*, J. Math. Systems Estim. Control, 6 (1996), pp. 339–342.

[23] H. Shibata and S. Hata, *L_2-bounded stability for stochastic systems*, Internat. J. Control, 18 (1973), pp. 1275–1280.

[24] H. Sussmann, *On the gap between deterministic and stochastic differential equations*, Ann. Probab., 6 (1978), pp. 19–41.

[25] V. A. Ugrinovskii, *Exponential stabilization of non-linear stochastic systems*, J. Appl. Math. Mech., 52 (1988), pp. 11–17.

[26] V. A. Ugrinovskii, *Robust H_∞ control in the presence of stochastic uncertainty.*, Internat. J. Control, 71 (1998), pp. 219–237.

[27] W. M. Wonham, *On matrix Riccati equation of stochastic control*, SIAM J. Control Optim., 6 (1968), pp. 681–697.

[28] L. Xie and C. D. de Souza, *Robust H_∞ control for linear systems with norm-bounded time-varying uncertainty*, IEEE Trans. Autom. Control, 37 (1992), pp. 1188–1191.

[29] V. A. Yakubovich, *Nonconvex optimization problem: The infinite-horizon linear-quadratic problem with quadratic constraints*, Systems Control Lett., 19 (1992), pp. 13–22.

[30] E. Yaz and N. Yildizbayrak, *Robustness of feedback-stabilized systems in the presence of nonlinear and random perturbation*, Internat. J. Control, 41 (1985), pp. 345–353.

[31] K. Yosida, *Functional Analysis*, Springer-Verlag, Berlin, Göttingen, Heidelberg, 1965.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 4, pp. 1123–1149

CONTROL OF THE STOCHASTIC BURGERS MODEL OF TURBULENCE*

GIUSEPPE DA PRATO[†] AND ARNAUD DEBUSSCHE[‡]

Abstract. We consider a control problem for a stochastic Burgers equation. This problem is motivated by a model from the control of Turbulence (see [Choi et al., *J. Fluid Mech.*, 253 (1993), pp. 509–543]). We study a sequence of approximated Hamilton–Jacobi equations by using dynamic programming.

Key words. stochastic Burgers equations, turbulence, Hamilton–Jacobi equations, dynamic programming

AMS subject classifications. 93C20, 93C90, 93E20, 76F99

PII. S0363012996311307

1. Introduction.

It has been shown in [3] that the stochastic Burgers equation is a good and simple model with which to study turbulence phenomena. The mathematical study of this equation has been the object of several papers [2], [9], [10], [17], [21].

This model also has been used in [6] to test a numerical algorithm for reducing the cost function in the very important problem of the control of turbulence.

In this paper we consider the stochastic Burgers equation with distributed parameter controls. The cost function has the same form as in [6] and contains the analogue of the kinetic energy. The problem is as follows: minimize

$$J(z) = \mathbb{E}\left(\int_0^T \left[\left|\frac{\partial X}{\partial \xi}\right|^2_{L^2(0,1)} + \frac{1}{2}\ |z(s)|^2_{L^2(0,1)}\right] ds + \frac{1}{2}\ |X(T)|^2_{L^2(0,1)}\right),$$

where the control z is in $L^2(\Omega \times [0,T] \times [0,1])$, and $X(t,\xi),\ \xi \in [0,1],\ t \in [0,T]$, is the solution of the controlled Burgers equation

$$\left\{\begin{array}{l} dX = \left(\dfrac{\partial^2 X}{\partial \xi^2} + \dfrac{\partial}{\partial \xi}(X^2)\right) dt + \sqrt{Q} z dt + \sqrt{Q} dW,\ \xi \in [0,1],\ t \geq 0, \\ X(t,0) = X(t,1) = 0,\ t \geq 0, \\ X(0,\xi) = x(\xi),\ \xi \in [0,1], \end{array}\right. \tag{1.1}$$

where $x \in L^2(0,1)$.

Here W is a cylindrical Wiener process on $L^2(0,1)$ (in other words $\frac{dW}{dt}$ is the "space–time white noise") and is adapted to a stochastic basis $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ (of course the control z has to be adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$). Moreover Q is a symmetric linear operator on $L^2(0,1)$. In (1.1) the operator $\sqrt{Q}$ acts both on the noise and on the control. This is essential in our work: it enables us to use a Hopf

*Received by the editors October 28, 1996; accepted for publication (in revised form) June 10, 1998; published electronically May 13, 1999.

http://www.siam.org/journals/sicon/37-4/31130.html

[†] Scuola Normale Superiore di Pisa, 56126 Pisa, Italy. The research of this author was partially supported by the Italian National Project MURST "Equazioni di Evoluzione e Applicazioni Fisico-Matematiche."

[‡] CNRS et Université de Paris Sud, 91405 Orsay, France. This research was done while this author was visiting the Scuola Normale Superiore di Pisa.

transform on the Hamilton–Jacobi equation (see below). This might be a restriction in the applications. However, this assumption is not artificial. It can be interpreted as a control acting on the solution in the same way as the noise or as a noise acting on the control.

It is easy to see that the cost functional J cannot have finite values unless Q is a nuclear operator. This is a simple consequence of the Ito formula.

In this paper we study this control problem following the dynamic programming approach. We solve the associated Hamilton–Jacobi equation and prove that it has a solution that coincides with the value function. More precisely, let A be the unbounded operator on $L^2(0,1)$ defined by

$$Ax = \frac{\partial^2 x}{\partial \xi^2},\ D(A) = H^2(0,1) \cap H_0^1(0,1),$$

and F, g are the nonlinear functions:

$$F(x) = \frac{\partial (x^2)}{\partial \xi},\quad g(x) = \left| \frac{\partial x}{\partial \xi} \right|^2_{L^2(0,1)}.$$

Then we can associate our control problem with the Hamilton–Jacobi equation

$$\left\{ \begin{aligned} u_t(t,x) &= \frac{1}{2}\ \mathrm{Tr}\ [Qu_{xx}(t,x)] + (Ax + F(x), u_x(t,x)) \\ &\quad - \frac{1}{2}\ |\sqrt{Q}\ u_x(t,x)|^2 + g(x), \\ u(0,x) &= \frac{1}{2}\ |x|^2_{L^2(0,1)} \end{aligned} \right. \tag{1.2}$$

for $x \in L^2(0,1)$, $t \in [0,T]$.

We prove below that there exists a solution u and that

$$u(T,x) = \inf_z J(z). \tag{1.3}$$

Moreover, for each control z and its associated solution X of (1.1), the fundamental identity holds:

$$u(T,x) + \frac{1}{2}\mathbb{E} \int_0^T |\sqrt{Q}\ u_x(T-s, X(s,x)) + z(s)|^2_{L^2(0,1)} ds = J(z). \tag{1.4}$$

We prove that the *closed loop equation*

$$\left\{ \begin{array}{l} dX^* = \left(\dfrac{\partial^2 X^*}{\partial \xi^2} + \dfrac{\partial}{\partial \xi}(X^{*2}) - Qu_x(T-t, X^*(t)) \right) dt + \sqrt{Q}\ dW, \\ X^*(t,0) = X^*(t,1) = 0,\ t \geq 0, \\ X^*(0,\xi) = x(\xi),\ \xi \in [0,1] \end{array} \right. \tag{1.5}$$

has a unique solution. It follows that there exists a unique optimal control given by

$$z^*(t) = -\sqrt{Q}\ u_x(T-t, X^*(t)). \tag{1.6}$$

To prove the existence of a solution to the Hamilton–Jacobi equation (1.2) we use a Hopf transform

$$u = -\ln v.$$

The function v satisfies

$$v_t(t,x) = \frac{1}{2}\,\mathrm{Tr}\,[Qv_{xx}(t,x)] + (Ax + F(x), v_x(t,x)) - g(x)v \tag{1.7}$$

so that using the Feynmann–Kac formula we have an explicit representation for u,

$$u(t,x) = -\ln \mathbb{E}\left(\exp\left[-\frac{1}{2}\,|Y(t)|^2_{L^2(0,1)} - \int_0^t g(Y(s))ds\right]\right), \tag{1.8}$$

where Y is the solution to the uncontrolled equation

$$\begin{cases} dY = \left(\dfrac{\partial^2 Y}{\partial \xi^2} + \dfrac{\partial}{\partial \xi}(Y^2)\right)dt + \sqrt{Q}dW,\ \xi \in [0,1],\ t \geq 0, \\ Y(t,0) = Y(t,1) = 0,\ t \geq 0, \\ Y(0,\xi) = x(\xi),\ \xi \in [0,1]. \end{cases} \tag{1.9}$$

The study of second-order Hamilton–Jacobi equations has been the object of several articles. Existence and uniqueness in finite and infinite dimensions have been obtained using semigroups methods (see [1], [8], [4], [5], [13], [14], [15]) and also using the concept of viscosity solution (see [7], [12], [19], [20], [16]). However, these results do not cover our case. Indeed here we simultaneousy have a non-Lipschitz Hamiltonian $H(u) = \frac{1}{2}\,|\sqrt{Q}\,u_x|^2$, a singular term in the cost functional $g(x) = \frac{1}{2}\,|\frac{\partial}{\partial\xi}X|^2_{L^2(0,1)}$, and the nonlinear term $f(x) = \frac{\partial}{\partial\xi}(X^2)$ coming from the Burgers equation.

All the formula described above can be derived formally; we use an approximation technique to justify them. We consider an approximate problem which is finite dimensional by using a Galerkin approximation and in which g and f are replaced by bounded and Lipschitz functions. We obtain a control problem which we can solve easily and a sequence $\{u^m\}$ of approximations of the solution to (1.2). We derive several a priori estimates and prove convergence of the approximation. The main difficulty is that we are not able to obtain an a priori estimate in the space of C^1 bounded functions on u^m. We have only the estimates

$$|u^m(t,x)| \leq \frac{1}{2}\,\left(|x|^2_{L^2(0,1)} + \ \mathrm{Tr}\ tQ\right),$$
$$|u^m_x(t,x)| \leq Ce^{\frac{1}{2}\,(|x|^2_{L^2(0,1)} + \ \mathrm{Tr}\ tQ)}$$

and a similar estimate on $u^m_{xx}(t)$.

However, we are able to prove that u^m converges to a C^2 function u which is a solution of (1.2), that the formulas (1.1) and (1.3) hold, and that the closed loop equation (1.5) has a unique solution. Thus the original control problem is completely solved.

2. Preliminaries and main results.

Let $H = L^2(0,1)$ endowed with the usual norm and inner product denoted by $|\cdot|$ and $(\cdot,\cdot)$. We define a linear operator A in H by setting

$$Ax = \frac{\partial^2 x}{\partial \xi^2},\ x \in D(A) = H^2(0,1) \cap H^1_0(0,1).$$

As usual, $H^k(0,1),\ k \in \mathbb{N}$, is the Sobolev space of all functions in H whose derivatives up to the order k belong to H, and $H_0^1(0,1)$ is the subspace of $H^1(0,1)$ of all functions whose traces at 0 and 1 vanish.

The operator A is self-adjoint and strictly negative and has a compact inverse. We can define $(-A)^s$ and $D((-A)^s)$ for any $s \in \mathbb{R}$. For $s = \frac{1}{2}$, we have $D((-A)^{1/2}) = H_0^1(0,1)$ and its norm and inner product are denoted by

$$\|x\| = |(-A)^{1/2}x|,\ ((x,y)) = \left((-A)^{1/2}x, (-A)^{1/2}y\right),\ x,y \in H_0^1(0,1).$$

The sequence of eigenvalues of A is

$$\lambda_k = -k^2\pi^2,\ k \in \mathbb{N},$$

and it is associated with the orthonormal basis of eigenvectors $\{e_k\}_{k\in\mathbb{N}}$,

$$e_k = \sqrt{2/\pi}\ \sin k\xi,\ k \in \mathbb{N},\ \xi \in [0,1].$$

For any positive integer m we denote by P_m the orthogonal projector on the space spanned by $e_1, \dots, e_m$.

We also consider a linear operator Q which is assumed to be symmetric, nonnegative, and of trace class; and a cylindrical Wiener process W on H associated with a stochastic basis $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$. (The reader is referred to [11] for precise definitions.) Let $L_W^2(\Omega \times [0,T]; H)$ be the space of all square integrable and adapted processes with values in $L^2(0,T;H)$. For $x \in H_0^1(0,1)$ we set

$$F(x) = \frac{\partial}{\partial \xi}(x^2).$$

The control problem we want to study is

$$\tag{2.1} \left\{ \begin{array}{l} \text{Minimize} \\ J(z) = \mathbb{E}\left(\displaystyle\int_0^T \left(\|X(t)\|^2 + \frac{1}{2}\,|z(t)|^2\right) dt + \frac{1}{2}\,|X(T)|^2\right) \\ \text{over all}\ \ z \in L_W^2(\Omega \times [0,T]; H), \end{array} \right.$$

where X is the solution of the controlled Burgers equation

$$\tag{2.2} \left\{ \begin{array}{l} dX = (AX + F(X) + \sqrt{Q}\ z)dt + \sqrt{Q}\ dW, \\ X(0) = x \end{array} \right.$$

and the initial datum x is in H.

For any $z \in L_W^2(\Omega \times [0,T]; H)$, (2.2) has a unique solution. More precisely, its solution can be constructed as the limit of Galerkin approximations. For $m \in \mathbb{N}$, we define F_m by

$$F_m(x) = \frac{\partial}{\partial \xi}\left(f_m(x)\right),\ x \in H_0^1(0,1),$$

where

$$f_m(\alpha) = \frac{m\alpha^2}{m + \alpha^2},\ \alpha \in \mathbb{R},$$

and we consider the following Galerkin approximation of (2.2):

$$\begin{cases} dX_m = (AX_m + P_m F_m(X_m) + P_m \sqrt{Q}\, z_m)dt + P_m \sqrt{Q}\, dW, \\ X_m(0) = x_m, \end{cases} \tag{2.3}$$

where $z_m \in L^2_W(\Omega \times [0,T]; P_m H)$ and $x_m \in P_m H$. The existence and uniqueness of X_m follow from the classical theory of finite dimensional stochastic differential equations.

We will need a lemma, whose proof is given in the appendix.

LEMMA 2.1. *Let $\{x_m\}_{m\in\mathbb{N}}, \{z_m\}_{m\in\mathbb{N}}$ be such that $x_m \to x$ in H and $z_m \to z$ in $L^2_W(\Omega \times [0,T]; H)$ and almost surely in $L^2([0,T]\times H)$. Let X_m be the solution of* (2.3). *Then $\{X_m\}_{m\in\mathbb{N}}$ is convergent to the unique solution X of* (2.2) *in*

$$L^2(\Omega \times [0,T]; H^1_0(0,1)) \cap L^2(\Omega; C([0,T]; H))$$

and almost surely in $L^2([0,T]\times H)$.

As mentioned in the introduction, we can formally associate the following Hamilton–Jacobi equation with the control problem (2.1)–(2.2):

$$\begin{cases} u_t(t,x) = \dfrac{1}{2}\,\mathrm{Tr}\,[Qu_{xx}(t,x)] + (Ax + F(x), u_x(t,x)) \\ \qquad\qquad - \dfrac{1}{2}\,|\sqrt{Q}\, u_x(t,x)|^2 + \|x\|^2, \\ u(0,x) = \dfrac{1}{2}\,|x|^2, \end{cases} \tag{2.4}$$

for $t \in [0,T], x \in H$. Using the Hopf transformation

$$u = -\ln v,$$

v formally satisfies

$$v_t(t,x) = \frac{1}{2}\,\mathrm{Tr}\,[Qv_{xx}(t,x)] + (Ax + F(x), v_x(t,x)) - \|x\|^2 v, \tag{2.5}$$

and so, by the Feynmann–Kac formula,

$$v(t,x) = \mathbb{E}\left(\exp\left[-\frac{1}{2}\,|Y(t)|^2 - \int_0^t \|Y(s)\|^2 ds\right]\right), \tag{2.6}$$

where Y is the solution to the uncontrolled Burgers equation

$$\begin{cases} dY = (AY + F(Y))dt + \sqrt{Q}\, dW, \\ Y(0) = x. \end{cases} \tag{2.7}$$

It is classical that Y is two times differentiable with respect to x. More precisely, we have the following lemma, whose proof is given in the appendix.

LEMMA 2.2. *The function v defined by* (2.6)–(2.7) *is two times differentiable with respect to $x \in H$. For any $(x,h) \in H \times H$, its derivative at x in the direction of h is given by*

$$v_x(t,x)h = -\mathbb{E}\left[\left((Y(t), \eta^h(t)) + 2\int_0^t ((Y(s), \eta^h(s)))ds\right) e^{-\frac{1}{2}|Y(t)|^2 - \int_0^t \|Y(s)\|^2 ds}\right], \tag{2.8}$$

where η^h is the solution of

$$\begin{cases} \dfrac{d\eta^h}{dt} = A\eta^h + 2\dfrac{\partial}{\partial \xi}\left(Y\eta^h\right), \\ \eta^h(0) = h. \end{cases} \tag{2.9}$$

Moreover, its second derivative is given by

$$\begin{aligned} v_{xx}(t,x)(h,h) = -\mathbb{E}\Bigg[\Bigg(&|\eta^h(t)|^2 + 2\int_0^t \|\eta^h(s)\|^2 ds + (Y(t),\zeta^h(t)) \\ &+ 2\int_0^t ((Y(s),\zeta^h(s)))ds\Bigg) e^{-\frac{1}{2}|Y(t)|^2 - \int_0^t \|Y(s)\|^2 ds}\Bigg] \\ +\mathbb{E}\Bigg[\Bigg((Y(t),\eta^h(t)) + 2\int_0^t ((Y(s),\eta^h(s)))ds\Bigg)^2 & e^{-\frac{1}{2}|Y(t)|^2 - \int_0^t \|Y(s)\|^2 ds}\Bigg], \end{aligned} \tag{2.10}$$

where ζ^h is the solution of

$$\begin{cases} \dfrac{d\zeta^h}{dt} = A\zeta^h + 2\dfrac{\partial}{\partial \xi}\left(Y\zeta^h + (\eta_h)^2\right), \\ \zeta^h(0) = 0. \end{cases} \tag{2.11}$$

We will also consider the Galerkin approximation of (2.9),

$$\begin{cases} \dfrac{d\eta_m^h}{dt} = A\eta_m^h + P_m\dfrac{\partial}{\partial \xi}\left(f_m'(Y_m)\eta_m^h\right), \\ \eta_m^h(0) = P_m h, \end{cases} \tag{2.12}$$

and of (2.11),

$$\begin{cases} \dfrac{d\zeta_m^h}{dt} = A\zeta_m^h + P_m\dfrac{\partial}{\partial \xi}\left(f_m'(Y_m)\zeta_m^h + f_m''(Y_m)(\eta_m^h)^2\right), \\ \zeta_m^h(0) = 0, \end{cases} \tag{2.13}$$

where Y_m is the solution to

$$\begin{cases} dY_m = (AY_m + F_m(Y_m))dt + P_m\sqrt{Q}\, dW, \\ Y_m(0) = x_m \end{cases} \tag{2.14}$$

and $x_m \in P_m H,\ h \in P_m h$.

LEMMA 2.3. *Let $\{x_m\}_{m\in\mathbb{N}}$ be such that $x_m \to x$ in H; then*

$$\sup_{h\in H, |h|=1} \left|\eta_m^{P_m h} - \eta^h\right|^2_{L^2(0,T;H_0^1(0,1))} \to 0,$$

$$\sup_{h\in H, |h|=1} \left|\eta_m^{P_m h} - \eta^h\right|_{C([0,T];H)} \to 0,$$

$$\sup_{h\in H, |h|=1} \left|\zeta_m^{P_m h} - \zeta^h\right|^2_{L^2(0,T;H_0^1(0,1))} \to 0,$$

$$\sup_{h\in H, |h|=1} \left|\zeta_m^{P_m h} - \zeta^h\right|_{C([0,T];H)} \to 0$$

almost surely when $m \to \infty$. The proof of this lemma is given in the appendix.

In section 4 we will prove, by an approximation technique, that v given by (2.6) is a strict solution of (2.5). By strict solution we mean that v is a C^2 function with respect to x; that for any $x \in D(A)$, $t \to v(t,x)$ is a C^1 function; and that (2.5) holds for any $(t,x) \in D(A) \times [0,T]$. We will also obtain that $u = \ln v$ is a strict solution of (2.4).

Then, again by approximation, we show that the fundamental identity (1.4) holds.

It remains to be proved that the closed loop equation (1.5) has a unique solution X^*. The difficulty here is that we have only a rather bad estimate on u_x. We will consider this problem in section 5.

The main result of this paper, whose proof is presented in sections 4 and 5, is the following.

THEOREM 2.4. *Let v be defined by* (2.6)–(2.7) *and $u = -\ln v$; then u is a strict solution to the Hamilton–Jacobi equation* (2.4). *Moreover for any $z \in L^2_W(\Omega \times [0,T]; H)$, we have*

$$u(T,x) + \frac{1}{2}\mathbb{E}\int_0^T |\sqrt{Q}\, u_x(T-s, X(s,x)) + z(s)|^2 ds = J(z),$$

where X is the solution of (2.2) *and J is defined by* (2.1).

The control problem (2.1) *has a unique solution given by*

$$z^*(t) = -\sqrt{Q}\, u_x(T-t, X^*(t)),$$

where X^* is the unique solution to the closed loop equation

$$\begin{cases} dX^* = (AX^* + F(X^*)dt - Qu_x(T-t, X^*(t)))\, dt + \sqrt{Q}\, dW, \\ X^*(0) = x. \end{cases}$$

Remark 2.5. In fact we prove a little bit more. Indeed, we show that the optimal control z^* and the optimal state X^* are the limits of an approximated finite dimensional problem.

3. Approximations.

We already have introduced the Galerkin approximation (2.3) of (2.2). We also need to approximate the terms $\|\cdot\|^2$ and $\frac{1}{2}|\cdot|^2$ in the functional J. If $l \in \mathbb{N}$, we set

$$\varphi_l(x) = \frac{1}{2}\,\frac{l|x|^2}{l + |x|^2},\ x \in H, \quad g_l(x) = \frac{l\|x\|^2}{l + \|x\|^2},\ x \in H_0^1(0,1).$$

The approximated control problem is

$$(3.1) \qquad \begin{cases} \text{Minimize} \\ J_{l,m}(z_m) = \mathbb{E}\left(\int_0^T (g_l(X_m(t)) + \frac{1}{2}\,|z_m(t)|^2)dt + \varphi_l(x_m(T))\right) \\ \text{over all } \ z_m \in L^2_W(\Omega \times [0,T]; P_mH), \end{cases}$$

where X_m is the solution of (2.3).

We define for $l, m \in \mathbb{N}$, $x_m \in P_mH$, $t \in [0,T]$

$$v^{l,m}(t, x_m) = \mathbb{E}\left(e^{-\varphi_l(Y_m(t)) - \int_0^t g_l(Y_m(s))ds}\right),$$

where Y_m is the solution of (2.14). It defines a two times continuously differentiable function with respect to $x_m \in P_mH$, and for $h \in P_mH$ we have

$$\begin{aligned} v^{l,m}_{x_m}(t,x)h &= -\mathbb{E}\Bigg[\bigg(\big(D_x\varphi_l(Y_m(t)),\eta^h_m(t)\big) + \int_0^t \big(D_x g_l(Y_m(s)),\eta^h_m(s)\big)\,ds\bigg) \\ &\qquad e^{-\varphi_l(Y_m(t))-\int_0^t g_l(Y_m(s))ds}\Bigg], \end{aligned} \tag{3.2}$$

where η^h_m is the solution of (2.12) and

$$\begin{aligned} v^{l,m}_{x_mx_m}(t,x)(h,h) &= -\mathbb{E}\Bigg[\bigg(\big(D_x\varphi_l(Y_m(t)),\zeta^h_m(t)\big) + \int_0^t \big(D_x g_l(Y_m(s)),\zeta^h_m(s)\big)\,ds \\ &+D^2_x\varphi_l(Y_m(t))(\eta^h_m(t),\eta^h_m(t)) + \int_0^t D^2_x g_l(Y_m(s))(\eta^h_m(s),\eta^h_m(s))ds\bigg) \\ &\qquad \times e^{-\varphi_l(Y_m(t))-\int_0^t g_l(Y_m(s))ds}\Bigg] \\ &+\mathbb{E}\Bigg[\bigg(\big(D_x\varphi_l(Y_m(t)),\eta^h_m(t)\big) + \int_0^t \big(D_x g_l(Y_m(s)),\eta^h_m(s)\big)\,ds\bigg)^2 \\ &\qquad \times e^{-\varphi_l(Y_m(t))-\int_0^t g_l(Y_m(s))ds}\Bigg], \end{aligned} \tag{3.3}$$

where ζ^h_m is the solution of (2.13). By the Feynman–Kac formula we know that $v^{l,m}$ satisfies the equation

$$\begin{cases} v^{l,m}_t = \dfrac{1}{2}\,\mathrm{Tr}\,[P_mQv^{l,m}_{x_m,x_m}] + \big(Ax_m + P_mF_m(x_m), v^{l,m}_{x_m}\big) - g_l(x_m)v^{l,m}, \\ v^{l,m}(0,x_m) = e^{-\varphi_l(x_m)} \end{cases} \tag{3.4}$$

on $P_mH \times [0,T]$. Also, clearly

$$v^{l,m}(t,x_m) \geq e^{-l(\frac{1}{2}+T)}. \tag{3.5}$$

Therefore the function

$$u^{l,m} = -\ln v^{l,m}$$

is two times continuously differentiable and it can be checked that it is a solution of the Hamilton–Jacobi equation associated with (3.1):

$$\begin{cases} u^{l,m}_t(t,x) &= \dfrac{1}{2}\,\mathrm{Tr}\,[P_mQu^{l,m}_{x_m,x_m}] + \big(Ax_m + P_mF_m(x_m), u^{l,m}_{x_m}\big) \\ &- \dfrac{1}{2}\,|P_m\sqrt{Q}\,u^{l,m}_{x_m}|^2 + g_l(x_m), \\ u^{l,m}(0,x_m) &= \varphi_l(x_m). \end{cases} \tag{3.6}$$

A standard computation using Ito's formula shows that

$$\begin{aligned} u^{l,m}(T,x_m) &+ \frac{1}{2}\int_0^T \left|\sqrt{Q}\, u^{l,m}_{x_m}(T-t,X_m(t)) + z_m(t)\right|^2 dt \\ &= \varphi_l(X_m(T)) + \int_0^T \left(g_l(X_m(t)) + \frac{1}{2}\,|z_m(t)|^2\right) dt \\ &+ \int_0^T \left(u^{l,m}_{x_m}(T-t,X_m(t)), P_m\sqrt{Q}\, dW(t)\right). \end{aligned} \tag{3.7}$$

Taking the expectation, we obtain the fundamental identity

$$u^{l,m}(T,x_m) + \frac{1}{2}\mathbb{E}\int_0^T \left|\sqrt{Q}\, u^{l,m}_{x_m}(T-t,X_m(t)) + z_m(t)\right|^2 dt = J_{l,m}(z_m). \tag{3.8}$$

We deduce that if $X^*_{l,m}$ is the solution to the closed loop equation

$$\left\{\begin{array}{l} dX^*_{l,m} = \left(AX^*_{l,m} + P_m F_m(X^*_{l,m}) - P_m Q u^{l,m}_{x_m}(T-t,X^*_{l,m})\right) dt + P_m\sqrt{Q}\, dW, \\ X^*_{l,m}(0) = x_m, \end{array}\right. \tag{3.9}$$

then there exists a unique optimal control $z^*_{l,m}$ for (3.1) which is given by the feedback formula

$$z^*_{l,m}(t) = -\sqrt{Q}\, u^{l,m}_{x_m}(T-t,X^*_{l,m}). \tag{3.10}$$

We will see below (see Lemma 4.1) that $v^{l,m}_{x_m}(T-t,X^*_{l,m})$ is a globally Lipschitz and bounded function so that by (3.5) the same holds for $u^{l,m}_{x_m}(T-t,X^*_{l,m})$ and we know that $X^*_{l,m}$ exists and is unique. We also have

$$J_{l,m}(z^*_{l,m}) = u^{l,m}(T,x_m) = \inf_{z_m} J_{l,m}(z_m) = -\ln\mathbb{E}\left(e^{-\varphi_l(y_m(T)) - \int_0^T g_l(Y_m(s))ds}\right), \tag{3.11}$$

where Y_m satisfies (2.14).

We will also use the function

$$v^m(t,x_m) = \mathbb{E}\left(e^{-\frac{1}{2}|Y_m(t)|^2 - \int_0^t \|Y_m(s)\|^2 ds}\right), \tag{3.12}$$

where Y_m is the solution of (2.14), with first and second derivatives given by

$$\begin{aligned} &v^m_{x_m}(t,x_m)h \\ &= -\mathbb{E}\left[\left((Y_m(t),\eta^h_m(t)) + 2\int_0^t ((Y_m(s),\eta^h_m(s)))ds\right) e^{-\frac{1}{2}|Y_m(t)|^2 - \int_0^t \|Y_m(s)\|^2 ds}\right], \end{aligned} \tag{3.13}$$

$$\begin{aligned} v^m_{x_m x_m}(t,x)(h,h) &= -\mathbb{E}\left[\left((Y_m(t),\zeta^h_m(t)) + 2\int_0^t ((Y_m(s),\zeta^h_m(s)))ds\right.\right. \\ &\left.\left. + |\eta^h_m(t)|^2 + 2\int_0^t \|\eta^h_m(s)\|^2 ds\right) e^{-\frac{1}{2}|Y_m(t)|^2 - \int_0^t \|Y_m(s)\|^2 ds}\right] \\ &+ \mathbb{E}\left[\left((Y_m(t),\eta^h_m(t)) + 2\int_0^t ((Y_m(s),\eta^h_m(s)))ds\right)^2 e^{-\frac{1}{2}|Y_m(t)|^2 - \int_0^t \|Y_m(s)\|^2 ds}\right], \end{aligned} \tag{3.14}$$

where η_m^h and ζ_m^h are defined by (2.12), (2.13).

In the next two sections, c denotes any constant depending only on the data A, Q, T. We always use the same symbol c although the constants have different values. Sometimes, we use a constant depending on $\omega \in \Omega$, or $m \in \mathbb{N}, \dots$, in which case we will write $C(\omega)$ or $k_m, \dots$.

Also, when f is a C^1 (resp., C^2) function from H or P_mH to $\mathbb{R}$, we will identify its first (resp., second) differential f_x (resp., f_{xx}) with the gradient (resp., the Hessian) of f; i.e., we use the two notations

$$f_x(x)h = (f_x(x), h),\ x, h \in H$$

and

$$f_{xx}(x)(h,h) = (f_{xx}(x)h, h),\ x, h \in H,$$

respectively.

4. Passing to the limit. We take the limit in our approximation in two steps. We first proceed to the limit $l \to \infty$, then, using a priori estimates on the Galerkin approximation, we take the limit $m \to \infty$. We first bound $v^{l,m}$ uniformly in l.

LEMMA 4.1. *For any $m \in \mathbb{N}$ there exists a constant $k_m$depending on m and on A, Q, T such that for any $x_m \in P_mH, t \in [0,T]$*

$$\text{(i)}\quad |v^{l,m}_{x_m}(t, x_m)| \le k_m,$$
$$\text{(ii)}\quad |v^{l,m}_{x_m x_m}(t, x_m)|_{\mathcal{L}(P_mH)} \le k_m.$$

Proof. We have the following inequalities:

$$|(-A)^{1/2} D_x g_l(y)|^2 \le 4 g_l(y),\quad y \in H_0^1(0,1), \tag{4.1}$$

$$|D_x \varphi_l(y)|^2 \le 4 \varphi_l(y),\quad y \in H, \tag{4.2}$$

$$|D_x^2 g_l(y)(\eta,\eta)| \le 6\|\eta\|^2,\quad y, \eta \in H_0^1(0,1), \tag{4.3}$$

$$|D_x^2 \varphi_l(y)(\eta,\eta)| \le 6|\eta|^2,\quad y, \eta \in H. \tag{4.4}$$

Since f_m' is bounded by $\sqrt{m}$ and (2.12) is a linear system of ordinary differential equations, there exists a constant $c(m,T)$ such that

$$|\eta_m^h(t)| \le c(m,T)|h|,\ h \in H. \tag{4.5}$$

Similarly we have for the solution of (2.13)

$$|\zeta_m^h(t)| \le c(m,T)|h|^2,\ h \in H. \tag{4.6}$$

By (3.2), (4.1), (4.2), and the Cauchy–Schwarz inequality, for $x_m, h \in P_mH, t \in [0,T]$,

$$\begin{aligned}
&|v^m_{x_m}(t, x_m)h| \\
&\le c(m,T)\mathbb{E}\left[\left(\varphi_l(Y_m(t)) + \int_0^t g_l(Y_m(s))ds\right)^{1/2} e^{-\varphi_l(Y_m(t)) - \int_0^t g_l(Y_m(s))ds}\right]|h| \\
&\le C(m,T)|h|,
\end{aligned}$$

since $\sqrt{x}e^{-x}$ is bounded. This proves (i). Similarly (ii) follows from (3.3), (4.1)–(4.4), and elementary inequalities. □

Using (4.1)–(4.4) and the dominated convergence theorem it can be seen that for any $x_m \in P_mH,\ t \in [0,T]$,

$$\begin{aligned} &v^{l,m}(t,x_m) \to v^m(t,x_m), \\ &v^{l,m}_{x_m}(t,x_m) \to v^m_{x_m}(t,x_m) \text{ in } P_mH, \\ &v^{l,m}_{x_mx_m}(t,x_m) \to v^m_{x_mx_m}(t,x_m) \text{ in } \mathcal{L}(P_mH) \end{aligned} \tag{4.7}$$

when $l \to \infty$. Also, using Lemma 4.1 and with another application of the dominated convergence theorem, it follows that v_m is a solution of

$$\begin{cases} v^m_t = \dfrac{1}{2}\operatorname{Tr}\left[P_mQv^m_{x_m,x_m}\right] + \left(Ax_m + P_mF_m(x_m), v^m_{x_m}\right) - \|x_m\|^2v^m, \\ \\ v^m(0,x_m) = e^{-\frac{1}{2}|x_m|^2}. \end{cases} \tag{4.8}$$

From Lemma 4.1 we deduce the following estimates on

$$u^{l,m} = -\ln v^{l,m}.$$

Lemma 4.2. *For any $m \in \mathbb{N}$, there exists a constant $k_m$depending on m and on A, Q, T such that for any $x_m \in P_mH, t \in [0,T]$*

(i) $|u^{l,m}(t,x_m)| \le \frac{1}{2}\ \left(|x_m|^2 + T\ Tr\,Q\right)$,

(ii) $|u^{l,m}_{x_m}(t,x_m)| \le k_m\ e^{\frac{1}{2}\ \left(|x_m|^2+T\ Tr\ Q\right)}$,

(iii) $|u^{l,m}_{x_mx_m}(t,x_m)|_{\mathcal{L}(P_mH)} \le k_me^{\frac{1}{2}\ \left(|x_m|^2+T\ Tr\ Q\right)} + k_m^2e^{|x_m|^2+T\ Tr\ Q}$.

Proof. By Jensen's inequality we have

$$v^{l,m}(t,x_m) \ge e^{-\mathbb{E}(\varphi_l(Y_m(t))+\int_0^t g_l(Y_m(s))ds)} \ge e^{-\mathbb{E}(\frac{1}{2}|Y_m(t)|^2+\int_0^t \|Y_m(s)\|^2ds)}.$$

By Ito's formula we have

$$\begin{aligned} \frac{1}{2}|Y_m(t)|^2 + \int_0^t \|Y_m(s)\|^2ds = &\int_0^t \left(Y_m(s), \sqrt{Q}\ dW(s)\right) \\ &+ \frac{1}{2}\ \left(|x_m|^2 +\ t\operatorname{Tr}\left(P_mQ\right)\right), \end{aligned}$$

since $(F_m(Y_m), Y_m) = 0$. Thus

$$v^{l,m}(t,x_m) \ge e^{-\frac{1}{2}\left(|x_m|^2+t\operatorname{Tr}\,(P_mQ)\right)} \ge e^{-\frac{1}{2}\left(|x_m|^2+T\operatorname{Tr}\,(P_mQ)\right)}.$$

Now (i) follows from the definition of $u^{l,m}$, and (ii), (iii) from the chain rule. □

Let us define

$$u^m = -\ln v^m.$$

Then by (4.7) for any $x_m \in P_mH,\ t \in [0,T]$,

$$\begin{aligned} &u^{l,m}(t,x_m) \to u^m(t,x_m), \\ &u^{l,m}_{x_m}(t,x_m) \to u^m_{x_m}(t,x_m) \text{ in } P_mH, \\ &u^{l,m}_{x_mx_m}(t,x_m) \to u^m_{x_mx_m}(t,x_m) \text{ in } \mathcal{L}(P_mH), \end{aligned} \tag{4.9}$$

and by (4.8) u_m is a solution of

$$\left\{\begin{aligned} u_t^m(t,x) &= \frac{1}{2}\,\mathrm{Tr}\,[P_mQu^m_{x_m,x_m}] + \big(Ax_m + P_mF_m(x_m), u^m_{x_m}\big)\\ &\quad -\frac{1}{2}\,|P_m\sqrt{Q}u^m_{x_m}|^2 + \|x_m\|^2,\\ u^m(0,x_m) &= \frac{1}{2}|x_m|^2. \end{aligned}\right. \tag{4.10}$$

Using Ito's formula we have for any $z_m \in L^2(\Omega\times[0,T];P_mH)$, $x_m \in P_mH$,

$$\begin{aligned} u^m(T,x_m) &+ \frac{1}{2}\int_0^T \left|\sqrt{Q}\;u^m_{x_m}(T-t,X_m(t)) + z_m(t)\right|^2 dt\\ &= \frac{1}{2}\,|X_m(T)|^2 + \int_0^T \left(\|X_m(t)\|^2 + \frac{1}{2}\,|z_m(t)|^2\right) dt\\ &+ \int_0^T \left(u^m_{x_m}(T-t,X_m(t)), P_m\sqrt{Q}\;dW(t)\right). \end{aligned} \tag{4.11}$$

We now derive some a priori estimates uniform in m in order to take the limit $m\to\infty$.

LEMMA 4.3. *There exists a constant k_1 depending on A,Q,T such that for any $x_m \in P_mH, t\in[0,T]$*

$$\text{(i)}\quad |v^m_{x_m}(t,x_m)| \le k_1,$$
$$\text{(ii)}\quad |v^m_{x_mx_m}(t,x_m)|_{\mathcal{L}(P_mH)} \le k_1.$$

Remark 4.4.

- We are not able to give an a priori estimate on $v^{l,m}$ independently of m. This explains why we take the limit in two steps.
- We do not have a lower bound on v^m such as in (3.5) for $v^{l,m}$. Thus we do not know whether u^m has a bounded derivative. Formally u^m and v^m are associated to the control problem in which the cost functional $J_{l,m}$ is replaced by

$$J_m(z_m) = \mathbb{E}\left(\int_0^T \left(\|X_m(t)\|^2 + \frac{1}{2}\,|z_m(t)|^2\right)dt + \frac{1}{2}\,|X_m(t)|^2\right). \tag{4.12}$$

We shall prove in section 5 that the corresponding closed loop equation has a unique solution.

Proof of Lemma 4.3. Let us first note that

$$f''_m(\alpha) \le 2,\ \alpha\in\mathbb{R}. \tag{4.13}$$

Let $h\in P_mH$. We take the scalar product of (2.12) with η^h_m and obtain

$$\begin{aligned} \frac{1}{2}\,\frac{d}{dt}|\eta^h_m|^2 + \|\eta^h_m\|^2 &= \left(P_m\frac{\partial}{\partial\xi}(f'_m(Y_m)\eta^h_m), \eta^h_m\right)\\ &= \frac{1}{2}\int_0^1 f''_m(Y_m)\left(\frac{\partial}{\partial\xi}Y_m\right)(\eta^h_m)^2 d\xi\\ &\le \|Y_m\|\;|\eta^h_m|^2_{L^4(0,1)} \end{aligned} \tag{4.14}$$

by integration by parts and Hölder's inequality. Using interpolation and the Sobolev embedding theorem, we have

$$|\eta_m^h|^2_{L^4(0,1)} \leq c|\eta_m^h|^{3/2}\ \|\eta_m^h\|^{1/2}. \tag{4.15}$$

Hence, using Young's inequality,

$$\frac{1}{2}\ \frac{d}{dt}\ |\eta_m^h|^2 + \|\eta_m^h\|^2 \leq c\|Y_m\|^{4/3}|\eta_m^h|^2 + \frac{1}{2}\ \|\eta_m^h\|^2,$$

and, by Gronwall's lemma,

$$\begin{cases} |\eta_m^h(t)|^2 \leq e^{c\int_0^t \|Y_m(s)\|^{4/3}ds}|h|^2, \\ \displaystyle\int_0^t \|\eta_m^h(s)\|^2 ds \leq e^{c\int_0^t \|Y_m(s)\|^{4/3}ds}|h|^2. \end{cases} \tag{4.16}$$

We infer from (3.13) and the Cauchy–Schwarz inequality that

$$|v_{x_m}^m(t,x_m)h| \leq \mathbb{E}\left[\left(|Y_m(t)|^2 + 2\int_0^t \|Y_m(s)\|^2 ds\right)^{1/2} \right.$$
$$\left. e^{-\frac{1}{2}|Y_m(t)|^2 - \int_0^t \|Y_m(s)\|^2 ds + c\int_0^t \|Y_m(s)\|^{4/3}ds}\right]|h|$$

and (i) follows from elementary inequalities.

For the second estimate we take the scalar product of (2.13) with ζ_m^h and obtain

$$\frac{1}{2}\ \frac{d}{dt}|\zeta_m^h|^2 + \|\zeta_m^h\|^2 = \left(P_m\frac{\partial}{\partial\xi}(f_m'(Y_m)\zeta_m^h), \zeta_m^h\right) + \left(P_m\frac{\partial}{\partial\xi}(f_m''(Y_m)(\eta_m^h)^2), \zeta_m^h\right)$$

and use

$$\left|\left(P_m\frac{\partial}{\partial\xi}(f_m'(Y_m)\zeta_m^h), \zeta_m^h\right)\right| \leq C\|Y_m\|^{4/3}|\zeta_m^h|^2 + \frac{1}{4}\ \|\zeta_m^h\|^2$$

and

$$\left|\left(P_m\frac{\partial}{\partial\xi}(f_m''(Y_m)(\eta_m^h)^2), \zeta_m^h\right)\right| = \left|\int_0^1 f_m''(Y_m)(\eta_m^h)^2\frac{\partial}{\partial\xi}\zeta_m^h d\xi\right|$$
$$\leq 2|\eta_m^h|^2_{L^4(0,1)}\|\zeta_m^h\| \leq C|\eta_m^h|^4_{L^4(0,1)} + \frac{1}{4}\ \|\zeta_m^h\|^2.$$

We deduce

$$\frac{d}{dt}\ |\zeta_m^h|^2 + \|\zeta_m^h\|^2 \leq c\|Y_m\|^{4/3}|\zeta_m^h|^2 + c\|\eta_m^h\|^4_{L^4(0,1)}$$

and by (4.15), (4.16), and Gronwall's lemma

$$\begin{cases} |\zeta_m^h(t)|^2 \leq e^{c\int_0^t \|Y_m(s)\|^{4/3}ds}|h|^4, \\ \displaystyle\int_0^t \|\zeta_m^h(s)\|^2 ds \leq e^{c\int_0^t \|Y_m(s)\|^{4/3}ds}\ |h|^4. \end{cases} \tag{4.17}$$

By the Cauchy–Schwarz inequality, (3.14), (4.16), (4.17), we obtain

$$|v^m_{x_m x_m}(t, x_m)(h, h)| \le c\mathbb{E}\left[\left(1 + |Y_m(t)|^2 + \int_0^t \|Y_m(s)\|^2 ds\right)\right.$$

$$\left. e^{-\frac{1}{2}|Y_m(t)|^2 - \int_0^t \|Y_m(s)\|^2 ds + c\int_0^t \|Y_m(s)\|^{4/3} ds}\right] |h|^2,$$

and (ii) follows. □

Applying Lemma 2.1 with $z_m = 0$ and $x_m = P_m x$, we easily prove that for each $x \in H, t \in [0, T]$

$$v^m(t, P_m x) \to v(t, x) \tag{4.18}$$

when $m \to \infty$. Also, we have for any $x \in H,\ t \in [0, T]$,

$$\begin{aligned} &|v_{x_m}(t, P_m x) - v_x(t, x)| \\ &= \sup_{|h|=1} (v_{x_m}(t, P_m x), P_m h) - (v_x(t, x), h) \\ &= \sup_{|h|=1} \mathbb{E}\Bigg[\left((Y_m(t), \eta_m^{P_m h}(t)) + 2\int_0^t ((Y_m(s), \eta_m^{P_m h}(s)))ds\right) e^{-\frac{1}{2}|Y_m(t)|^2 - \int_0^t ||Y_m(s)||^2 ds} \\ &\qquad -\left((Y(t), \eta(t)) + 2\int_0^t ((Y(s), \eta(s)))ds\right) e^{-\frac{1}{2}|Y(t)|^2 - \int_0^t ||Y(s)||^2 ds}\Bigg] \\ &\le \mathbb{E}\Bigg[\sup_{|h|=1} \Bigg|\left((Y_m(t), \eta_m^{P_m h}(t)) + 2\int_0^t ((Y_m(s), \eta_m^{P_m h}(s)))ds\right) e^{-\frac{1}{2}|Y_m(t)|^2 - \int_0^t ||Y_m(s)||^2 ds} \\ &\qquad -\left((Y(t), \eta(t)) + 2\int_0^t ((Y(s), \eta(s)))ds\right) e^{-\frac{1}{2}|Y(t)|^2 - \int_0^t ||Y(s)||^2 ds}\Bigg|\Bigg]. \end{aligned}$$

It follows from Lemma 2.2 and Lemma 2.3 that the quantity inside of the expectation of the right-hand side above almost surely goes to zero. We infer from the dominated convergence theorem and estimate (4.16) that

$$v^m_{x_m}(t, P_m x) \to v_x(t, x) \text{ in } H. \tag{4.19}$$

By a similar argument, we prove that for any $x \in H,\ t \in [0, T]$,

$$v^m_{x_m x_m}(t, P_m x) \to v_{xx}(t, x) \text{ in } \mathcal{L}(H) \tag{4.20}$$

when $m \to \infty$. (The expressions of v, v_x, v_{xx} are given in (2.6), (2.8), (2.10).) Integrating (4.8), we have for $x \in H, t \in [0, T]$,

$$\begin{aligned} &v^m(t, P_m x) = e^{-\frac{1}{2}|P_m x|^2} \\ &+ \int_0^t \Big[\frac{1}{2} \text{ Tr } \big(P_m Q v^m_{x_m x_m}(s, P_m x)\big) + \big(P_m A x + P_m F_m(P_m x), v^m_{x_m}(s, P_m x)\big) \\ &\qquad -\|P_m x\|^2 v^m(s, P_m x)\Big] ds. \end{aligned}$$

We choose $x \in D(A)$. Using Lemma 4.3, we have for any $s \in [0,T]$

$$\Bigg|\frac{1}{2}\ \mathrm{Tr}\ \left(P_m Q v^m_{x_m x_m}(s, P_m x)\right) + \left(P_m A x + P_m F_m(P_m x), v^m_{x_m}(s, P_m x)\right)$$

$$-\|P_m x\|^2 v^m(s, P_m x)\Bigg|$$

$$\leq \frac{1}{2}\ k_1\ \mathrm{Tr}\ Q + k_1 \left(|Ax|^2 + c|x|^{1/2}\|x\|^{3/2}\right) + \|x\|^2.$$

We have used inequalities

$$0 \leq v^m(s, P_m x) \leq 1$$

and the following consequence of Agmon's inequality:

$$|P_m F_m(P_m x)| \leq |F_m(P_m x)| \leq |f'_m(P_m x)\frac{\partial}{\partial \xi} P_m x|$$

$$\leq 2|P_m x|_{L^\infty(0,1)}\|P_m x\| \leq c|P_m x|^{1/2}\|P_m x\|^{3/2} \leq c|x|^{1/2}\|x\|^{3/2}.$$

We deduce from (4.18), (4.19), (4.20), and the dominated convergence theorem that for $x \in D(A), t \in [0,T]$,

$$v(t,x) = e^{-\frac{1}{2}|x|^2}$$

$$+ \int_0^t \left[\frac{1}{2}\ \mathrm{Tr}\ (Q v^m_{xx}(s,x)) + (Ax + F(x), v_x(s,x)) - \|x\|^2 v(s,x)\right] ds.$$

Since v, v_x, v_{xx} are continuous with respect to t, it follows that $t \mapsto v(t,x)$ is a C^1 function for $x \in D(A)$, and for $x \in D(A), t \in [0,T]$,

$$v_t(t,x) = \frac{1}{2}\ \mathrm{Tr}\ [Q v_{xx}(t,x)] + (Ax + F(x), v_x(t,x)) - \|x\|^2 v(t,x),$$

so that v is a strict solution of (2.5). The following lemma is an easy consequence of Lemma 4.2, Lemma 4.3, and (4.9).

LEMMA 4.5. *For any $x_m \in P_m H, t \in [0,T]$ we have*

(i) $0 \leq |u^m(t, x_m)| \leq \frac{1}{2}\ \left(|x_m|^2 + T\ \mathrm{Tr}\ Q\right)$,

(ii) $|u^m_{x_m}(t, x_m)| \leq k_1\ e^{\frac{1}{2}\ \left(|x_m|^2 + T\ \mathrm{Tr}\ Q\right)}$,

(iii) $|u^m_{x_m x_m}(t, x_m)|_{\mathcal{L}(P_m H)} \leq k_1 e^{\frac{1}{2}\ \left(|x_m|^2 + T\ \mathrm{Tr}\ Q\right)} + k_1^2 e^{|x_m|^2 + T\ \mathrm{Tr}\ Q}$.

From (4.18), (4.19), and (4.20) we have for $x \in H, t \in [0,T]$,

$$\begin{aligned} & u^m(t, P_m x) \to u(t,x), \\ (4.21) \qquad & u^m_{x_m}(t, P_m x) \to u_x(t,x) \text{ in } H, \\ & u^m_{x_m x_m}(t, P_m x) P_m \to u_{xx}(t,x) \text{ in } \mathcal{L}(H). \end{aligned}$$

Arguing as above we see that $t \mapsto u(t,x)$ is a C^1 function for $x \in D(A)$, and for $x \in D(A), t \in [0,T]$,

$$u_t(t,x) = \frac{1}{2}\ \mathrm{Tr}\ [Q u_{xx}(t,x)] + (Ax + F(x), u_x(t,x)) - \frac{1}{2}\ |\sqrt{Q} u_x(t,x)|^2 + \|x\|^2,$$

and u is a strict solution of the Hamilton–Jacobi equation (2.4).

We now want to take the limit $m \to \infty$ in (4.11).

LEMMA 4.6. *Let $x \in H$ and $z \in L^2_W(\Omega \times [0,T]; H)$, let X be the solution of* (2.2) *and X_m the solution of* (2.3), *with $x_m = P_m x, z_m = P_m z$. Then*

$$u^m_{x_m}(T-t, X_m(t)) \to u_x(T-t, X(t))$$

in $L^2(0,T;H)$ $\mathbb{P}$–almost surely.

Proof. It is shown in the proof of Lemma 2.1 that X_m is almost surely bounded in $L^\infty(0,T;H)$ and it converges almost surely in $L^2(0,T;H)$ to X. By the mean value theorem and Lemma 4.5(iii) it follows

$$u^m_{x_m}(T-t, X_m(t)) - u^m_{x_m}(T-t, P_m X(t)) \to 0$$

in $L^2(0,T;H)$ almost surely. Also by (4.21)

$$u^m_{x_m}(T-t, P_m X(t)) - u_x(T-t, X(t)) \to 0$$

in H for any $t \in [0,T]$, and by Lemma 4.5(ii) and the dominated convergence theorem we deduce that this convergence holds in $L^2(0,T;H)$. □

Let $x \in H$ and $z \in L^2_W(\Omega \times [0,T]; H)$. We take $x_m = P_m x, z_m = P_m z$ in (4.11). Then thanks to Lemma 2.1, (4.21), and Lemma 4.6 we can take the limit and deduce that

$$\begin{aligned} u(T,x) &+ \frac{1}{2}\int_0^T \left|\sqrt{Q}\, u_x(T-t,X(t)) + z(t)\right|^2 dt \\ &= \frac{1}{2}\,|X(T)|^2 + \int_0^T \left(\|X(t)\|^2 + \frac{1}{2}\,|z(t)|^2\right) dt \\ &+ \int_0^T \left(u_x(T-t,X(t)), \sqrt{Q}\, dW(t)\right). \end{aligned} \tag{4.22}$$

From (3.8) we have

$$\begin{aligned} &\mathbb{E}\left(\int_0^T |\sqrt{Q}\, u^{l,m}_{x_m}(T-t, X_m(t))|^2 dt\right) \\ &\le 4J_{l,m}(z_m) + 2\mathbb{E}\int_0^T |z_m|^2 dt. \end{aligned}$$

By Ito's formula, for $t \in [0,T]$

$$\begin{aligned} &\frac{1}{2}\mathbb{E}|X_m(t)|^2 + \int_0^t ||X_m(s)||^2 ds \\ &= \mathbb{E}\int_0^t (X_m(s), z_m(s))ds + \frac{1}{2}|x_m|^2 + \frac{1}{2}t \text{ Tr } P_m Q \\ &\le \frac{1}{2}\int_0^t |X_m(s)|^2 ds + \frac{1}{2}s\int_0^t |z(s)|^2 ds + \frac{1}{2}|x|^2 + \frac{1}{2}T \text{ Tr } Q. \end{aligned}$$

By the Gronwall lemma it follows easily that

$$J_{l,m}(z_m) \le c\left(|x|^2 + T \text{ Tr } (Q) + \mathbb{E}\int_0^T |z|^2 dt\right)$$

and

$$\mathbb{E}\left(\int_0^T |\sqrt{Q}\, u^{l;m}_{x_m}(T-t,X_m(t))|^2 dt\right) \le c\left(|x|^2 + T \text{ Tr }(Q) + \mathbb{E}\int_0^T |z|^2 dt\right).$$

It follows that $\sqrt{Q}\, u^{l;m}_{x_m}(T-t,X_m(t))$ is bounded in $L^2(\Omega\times[0,T])$. Since it converges pointwise to $\sqrt{Q}\, u_x(T-t,X(t))$ we have by Fatou's lemma

$$\mathbb{E}\left(\int_0^T |\sqrt{Q}\, u_x(T-t,X(t))|^2 dt\right) \le c\left(|x|^2 + T \text{ Tr }(Q) + 2\mathbb{E}\int_0^T |z|^2 dt,\right)$$

and $\sqrt{Q}\, u_x(T-t,X(t))$ belongs to $L^2(\Omega\times[0,T])$. Therefore we can take the expectation in (4.22) and obtain the fundamental identity

$$u(T,x) + \frac{1}{2}\,\mathbb{E}\int_0^T |\sqrt{Q}\, u_x(T-t,X(t)) + z(t)|^2 dt = J(z). \tag{4.23}$$

5. Existence of a solution to the closed loop equation.

We now consider the closed loop equation

$$\begin{cases} dX^* = (AX^* + F(X^*)dt - Qu_x(T-t,X^*(t)))\,dt + \sqrt{Q}\,dW, \\ X^*(0) = x. \end{cases} \tag{5.1}$$

We first note that thanks to Lemma 4.5 and (4.21)

$$|u_x(t,x)| \le k_1\, e^{\frac{1}{2}\,\left(|x|^2 + T \text{ Tr } Q\right)},\ x\in H, \tag{5.2}$$

$$|u_{xx}(t,x)|_{\mathcal{L}(H)} \le 2k_1^2\, e^{\left(|x|^2 + T \text{ Tr } Q\right)},\ x\in H. \tag{5.3}$$

Hence u_x is locally Lipschitz in x. This is the main ingredient in the proof of the following result.

Lemma 5.1. *There exists at most one solution of* (5.1) *with trajectories in*

$$L^\infty(0,T;H)\cap L^2(0,T;H^1_0(0,1)).$$

Proof. Let X_1, X_2 be two solutions of (5.1) and $X = X_1 - X_2$. We have

$$\frac{dX}{dt} = AX + F(X_1) - F(X_2) - Q\left(u_x(T-t,X_1(t)) - u_x(T-t,X_2(t))\right).$$

It follows that $\frac{dX}{dt}\in L^2(0,T;H^{-1}(0,1))$ and

$$\frac{1}{2}\,\frac{d}{dt}|X|^2 + \|X\|^2 = (F(X_1)-F(X_2),X) - (Qu_x(T-t,X_1) - Qu_x(T-t,X_2),X)$$

$$\le \frac{1}{2}\,\|X\|^2 + c\left(|X_1|^2_{L^\infty(0,1)} + |X_2|^2_{L^\infty(0,1)}\right)|X|^2 + 4k_1^4 e^{2\left(M_1^2 + T \text{ Tr } Q\right)}|X|^2,$$

where

$$M_1 = \max\{|X_1|_{L^\infty(0,T;H)}, |X_2|_{L^\infty(0,T;H)}\}.$$

By the Sobolev embedding theorem

$$|X_i|_{L^\infty(0,1)} \le c\|X_i\|, \quad i = 1, 2.$$

Thus by Gronwall's lemma

$$|X(t)|^2 \le e^{c\int_0^T (\|X_1\|^2+\|X_2\|^2)ds+M_2T}|X(0)|^2,$$

with

$$M_2 = 4k_1^4 e^{2\left(M_1^2+T \text{ Tr } Q\right)}.$$

The result follows since $X(0) = 0$. □

We prove the existence of X^* by approximation. Let $X^*_{l,m}$ be the solution of (3.9) with $x_m = P_m x$.

Lemma 5.2. *There exists a constant k_2 depending only on A, Q, T such that for any $l, m \in \mathbb{N}$*

$$\mathbb{E}\left(\sup_{t\in[0,T]} |X^*_{l,m}(t)|^2\right) \le k_2\left(|x|^2 + \text{ Tr } Q\right).$$

Proof. First we have

$$J_{l,m}(z^*_{l,m}) \le J_{l,m}(0) \le \frac{1}{2}\left(|x_m|^2 + t \text{ Tr } Q\right)$$

by Ito's formula, where $z^*_{l,m}$ is defined in (3.10). It follows that

$$\mathbb{E}\left(\int_0^T |z^*_{l,m}(t)|^2 dt\right) \le |x_m|^2 + t \text{ Tr } Q \le |x|^2 + t \text{ Tr } Q. \tag{5.4}$$

By Ito's formula

$$\begin{aligned}
&\frac{1}{2}|X^*_{l,m}(t)|^2 + \int_0^t \|X^*_{l,m}(s)\|^2 ds \\
&= \int_0^t \left(\sqrt{Q}\, z^*_{l,m}(s), X^*_{l,m}(s)\right) ds + \int_0^t \left(X^*_{l,m}(s), \sqrt{Q}\, dW(s)\right) \\
&+\frac{1}{2}\left(|x_m|^2 + t \text{ Tr } (P_m Q)\right) \\
&\le c\int_0^T |z^*_{l,m}(s)|^2 ds + \frac{1}{2}\int_0^T \|X^*_{l,m}(s)\|^2 ds \\
&+ \sup_{t\in[0,T]} \int_0^t \left(X^*_{l,m}(s), \sqrt{Q}\, dW(s)\right) + \frac{1}{2}\left(|x|^2 + t \text{ Tr } Q\right).
\end{aligned}$$

The result follows by the application of a Martingale inequality. □

We deduce that there exists $\overline{X}_m$ in $L^2(\Omega; L^\infty(0,T;H))$ such that

$$\mathbb{E}\left(\sup_{t\in[0,T]} |\overline{X}_m(t)|\right) \le k_2\left(|x|^2 + \text{ Tr } Q\right)$$

and

$$X^*_{l,m} \rightharpoonup \overline{X}_m \text{in } L^2(\Omega; L^\infty(0,T;H)), \text{ weak star.}$$

We now derive a pathwise estimate for solutions of (3.9).

LEMMA 5.3. *Let $k(\omega)$ be a random variable. For any $m \in \mathbb{N}$, there exist random times t^m_k and constants c^m_k such that if $\widetilde{X}_{l,m}$ is a solution of* (3.9) *satisfying*

$$|\widetilde{X}_{l,m}(0)| \le k, \ \mathbb{P}\text{–almost surely,}$$

then

$$\sup_{t\in[0,t^m_k]} |\widetilde{X}_{l,m}(t)|^2 + \int_0^{t_k} \|\widetilde{X}_{l,m}(s)\|^2 ds \le c^m_k, \ \mathbb{P}\text{–almost surely}$$

Proof. Let

$$W^m_A(t) = \int_0^t e^{(t-s)A} P_m \sqrt{Q}\, dW(s)$$

and

$$\overline{X}_{l,m} = \widetilde{X}_{l,m} - W^m_A.$$

Then

$$\frac{d}{dt}\overline{X}_{l,m} = A\overline{X}_{l,m} + P_m F_m(\overline{X}_{l,m} + W^m_A) - Qu^{l,m}_{x_m}(T-t, \overline{X}_{l,m} + W^m_A).$$

Using similar arguments as in the proofs of Lemmas 2.1 and 4.2 we can prove

$$\begin{aligned} \frac{d}{dt}|\overline{X}_{l,m}|^2 &+ \|\overline{X}_{l,m}\|^2 \le c|W^m_A|^{4/3}_{L^4(0,1)}|\overline{X}_{l,m}|^2 + ck_m e^{2|\overline{X}_{l,m}|^2+2|W^m_A|^2+T\ \mathrm{Tr}\ Q} \\ &+ c|W^m_A|^4_{L^4(0,1)}. \end{aligned}$$

We set

$$\begin{aligned} F_m(t) &= e^{-c\int_0^t |W^m_A(s)|^{4/3}_{L^4(0,1)}ds}|\overline{X}_{l,m}|^2, \\ g_m &= 2e^{c\int_0^t |W^m_A(s)|^{4/3}_{L^4(0,1)}ds}, \\ h_m &= ck_m e^{\sup_{t\in[0,T]}|W^m_A(t)|^2+T\ \mathrm{Tr}\ Q}, \\ k_m &= c\sup_{t\in[0,T]}|W^m_A(t)|^4. \end{aligned}$$

It is easy to obtain

$$\frac{d}{dt}\,F_m \le h_m e^{g_m F_m} + k_m$$

so that

$$e^{-g_m F_m(t)} \ge -\left(h_m + k_m\right)t + e^{-g_m k^2},$$

and if we take

$$t_k^m = \frac{1}{2(h_m + k_m)}\, e^{-g_m k^2}$$

we have

$$F_m(t) \le \frac{1}{g_m}\, \ln 2 + k^2$$

for $t \in [0, t_k^m]$. Now the proof can be completed easily. □

It is not difficult to use the estimate in Lemma 5.3 and to prove that, for almost all $\omega \in \Omega$, a subsequence $\{X_{l,m}^*\}$ converges to X_m^* a solution of

$$(5.5)\quad \begin{cases} dX_m^* = (AX_m^* + P_m F_m(X_m^*) - P_m Q u_{x_m}^m (T-t, X_m^*))dt + P_m\sqrt{Q}\, dW, \\ \\ X_m^*(0) = x_m \end{cases}$$

on the interval $[0, t_m^k]$ whenever $|x_m| \le k$. Arguing as in Lemma 5.1, (5.5) has at most one solution so that the whole sequence converges.

We take

$$k = |\overline{X}_m|_{L^\infty(0,T;H)}.$$

Since $\{X_{l,m}^*\}$ converges pointwise to X_m^* and in $L^2(\Omega; L^\infty(0,T;H))$ weak star to $\overline{X}_m$, we have $X_m^* = \overline{X}_m$ $\mathbb{P}$−almost surely on $[0, t_k^m]$. It follows $|X_m^*(t_k^m)| \le k$, so that our construction can be reiterated and X_m^* can be prolonged to a solution of (5.5) on the interval $[0,T]$. Moreover, by Lemma 5.2 if $x_m = P_m x$

$$\mathbb{E}\left(\sup_{t\in[0,T]} |X_m^*|^2\right) \le k_2\left(|x|^2 + T\ \mathrm{Tr}\ Q\right).$$

Arguing as in the proof of Lemma 5.3 and using Lemma 4.5 and the uniform boundedness of W_A^m in $L^\infty(0,T;L^4(0,1))$, we prove the following pathwise estimate on X_m^*.

Lemma 5.4. *Let $k(\omega)$ be a random variable; there exists a random time t_k and a constant c_k such that if $\widetilde{X}_m$ is a solution of* (5.5) *satisfying*

$$|\widetilde{X}_m(0)| \le k, \ \textit{almost surely,}$$

then

$$\sup_{t\in[0,t_k]} |\widetilde{X}_m(t)|^2 + \int_0^{t_k} \|\widetilde{X}_m(s)\|^2 ds \le c_k, \ \textit{almost surely.}$$

Now we can repeat the argument that we have used to construct X_m^* and obtain X^*, a solution of (5.1) on $[0,T]$ such that

$$\mathbb{E}\left(\sup_{t\in[0,T]} |X^*(t)|^2\right) \le k_2\left(|x|^2 + T\ \mathrm{Tr}\ Q\right).$$

It remains to prove that

$$z^*(t) = -\sqrt{Q}\, u_x(T-t, X^*(t))$$

is an admissible control, i.e., that $z^* \in L^2_W(\Omega \times [0,T]; H)$.

Arguing as in Lemma 4.6 we have

$$u^{l,m}_{x_m}(T-t, X^*_{l,m}(t)) \overset{l\to\infty}{\to} u^m_{x_m}(T-t, X^*_m(t)) \overset{m\to\infty}{\to} u_x(T-t, X^*(t))$$

in $L^2(0,T;H)$ $\mathbb{P}$–almost surely. Thus by (5.4) we have $u_x(T-t, X^*(t)) \in L^2_W(\Omega \times [0,T]; H)$ and

$$\mathbb{E}\left(\int_0^T |z^*(t)|^2 dt\right) \le |x|^2 + T \operatorname{Tr} Q.$$

This ends the proof of Theorem 2.4. □

Appendix A.

A.1. Proof of Lemma 2.1. For any $m \in N$ we set

$$W^m_A(t) = \int_0^t e^{(t-s)A} P_m \sqrt{Q}\, dW(s);$$

it is the unique solution of

$$\begin{cases} dW^m_A = AW^m_A dt + P_m \sqrt{Q}\, dW, \\ W^m_A(0) = 0. \end{cases}$$

Also

$$W_A(t) = \int_0^t e^{(t-s)A} \sqrt{Q}\, dW(s)$$

is the unique solution (see [11]) of

$$\begin{cases} dW_A = AW_A dt + \sqrt{Q}\, dW, \\ W_A(0) = 0. \end{cases}$$

It is not difficult to see that W^m_A converges to W_A in $L^4([0,T] \times [0,1])$ almost surely.

Let X_m be the solution to (2.3). We set

$$\overline{X}_m = X_m - W^m_A;$$

thus

$$\text{(A.1)} \qquad \begin{cases} \dfrac{d\overline{X}_m}{dt} = A\overline{X}_m + P_m F_m(\overline{X}_m + W^m_A) + P_m \sqrt{Q}\, z_m, \\ \overline{X}_m(0) = x_m. \end{cases}$$

To derive an a priori estimate, we take the scalar product of (A.1) by $\overline{X}_m$. Using integration by parts, interpolation inequality, Sobolev embedding theorem, and Young's inequality, we have

$$\begin{aligned}&\left(P_mF_m(\overline{X}_m+W_A^m),\overline{X}_m\right)=-\int_0^1\left(f_m(\overline{X}_m+W_A^m)-f_m(\overline{X}_m)\right)\frac{\partial}{\partial\xi}\overline{X}_m d\xi\\&\leq 2\int_0^1|2\overline{X}_m+W_A^m|\,|W_A^m|\left|\frac{\partial}{\partial\xi}\overline{X}_m\right|d\xi\\&\leq 2\left(2|\overline{X}_m|_{L^4(0,1)}+|W_A^m|_{L^4(0,1)}\right)|W_A^m|_{L^4(0,1)}\|\overline{X}_m\|\\&\leq c|W_A^m|_{L^4(0,1)}|\overline{X}_m|^{3/4}\|\overline{X}_m\|^{5/4}+\frac{1}{8}\,\|\overline{X}_m\|^2+8|W_A^m|^4_{L^4(0,1)}\\&\leq c|W_A^m|^{8/3}_{L^4(0,1)}|\overline{X}_m|^2+\frac{1}{4}\,\|\overline{X}_m\|^2+8|W_A^m|^4_{L^4(0,1)}.\end{aligned}$$

We deduce

$$\frac{d}{dt}\,|\overline{X}_m|^2+\|\overline{X}_m\|^2\leq c|W_A^m|^{8/3}_{L^4(0,1)}|\overline{X}_m|^2+16|W_A^m|^4_{L^4(0,1)}+c|\sqrt{Q}\,z_m|^2$$

and

$$\begin{aligned}&|\overline{X}_m(t)|^2+\int_0^t\|\overline{X}_m(s)\|^2ds\\&\leq e^{c\int_0^t|W_A^m(s)|^{8/3}_{L^4(0,1)}ds}|x_m|^2\\&\quad+\int_0^t e^{c\int_s^t|W_A^m(r)|^{8/3}_{L^4(0,1)}dr}\left(16|W_A^m(s)|^4_{L^4(0,1)}+c+|\sqrt{Q}\,z_m|^2\right)\,ds.\end{aligned}$$

This proves that for fixed $\omega\in\Omega$, $\{\overline{X_m}\}$ is a bounded sequence in $L^\infty(0,T;L^2(0,1))$ and $L^2(0,T;H_0^1(0,1))$. By standard arguments based on compactness and the uniqueness of the limit (see [18]), we deduce that $\{\overline{X}_m\}$ converges almost surely to $\overline{X}$ in $L^2([0,T]\times[0,1])$, the unique solution of

$$\begin{cases}\dfrac{d\overline{X}}{dt}=A\overline{X}+F(\overline{X}+W_A)+\sqrt{Q}\,z,\\ \overline{X}(0)=x.\end{cases}$$

We set $X=\overline{X}+W_A$ and have

$$\begin{cases}dX=(AX+F(X)+\sqrt{Q}\,z)dt+\sqrt{Q}\,dW,\\ X(0)=x.\end{cases}$$

We apply Ito's formula to $|X^m|^2$ and take the expectation

$$\begin{aligned}&\frac{1}{2}\,\mathbb{E}|X^m(t)|^2+\mathbb{E}\int_0^t\|X^m(s)\|^2ds=\frac{1}{2}\,|x_m|^2\\&\qquad+\mathbb{E}\left(\int_0^t\left(\sqrt{Q}\,z_m,X^m\right)ds+\frac{1}{2}\,t\mathrm{Tr}\,[P_mQ]\right).\end{aligned}\tag{A.2}$$

Hence

$$\mathbb{E}|X^m(t)|^2+\mathbb{E}\int_0^t\|X^m\|^2ds\leq\ |x_m|^2+c\mathbb{E}\left(\int_0^t|\sqrt{Q}\,z_m|^2ds+t\mathrm{Tr}\ Q\right),$$

which proves that X^m is bounded in $L^2(\Omega, L^2(0,T; H_0^1(0,1)))$ and $X^m(t)$ in $L^2(\Omega, L^2(0,1))$. It is classical that this implies

$$\begin{aligned} &X^m \rightharpoonup X \text{ in } L^2(\Omega, L^2(0,T; H_0^1(0,1)) \text{ weak}, \\ &X^m(t) \rightharpoonup X(t) \text{ in } L^2(\Omega, L^2(0,1)) \text{ weak}. \end{aligned} \tag{A.3}$$

Since z_m converges to z in $L^2(\Omega, L^2(0,T; L^2(0,1)))$ strongly, we also have

$$\mathbb{E}\int_0^t \left(\sqrt{Q}\, z_m, X^m\right) ds \to \mathbb{E}\int_0^t \left(\sqrt{Q}\, z, X\right) ds. \tag{A.4}$$

By Ito's formula for $|X|^2$, we also have

$$\frac{1}{2}\,\mathbb{E}|X(t)|^2 + \mathbb{E}\int_0^t \|X(s)\|^2 ds = \frac{1}{2}\,|x|^2 + \mathbb{E}\int_0^t \left(\sqrt{Q}\, z, X\right) ds + \frac{1}{2}\, t \operatorname{Tr} Q,$$

and by $(A.2)$, $(A.4)$

$$\frac{1}{2}\,\mathbb{E}|X^m(t)|^2 + \mathbb{E}\int_0^t \|X^m(s)\|^2 ds \to \frac{1}{2}\,\mathbb{E}|X(t)|^2 + \mathbb{E}\int_0^t \|X(s)\|^2 ds$$

so that convergences in $(A.3)$ hold in the strong topology.

Let us write Ito's formula for $\frac{1}{2}\,|X^m - P_m X|^2$:

$$\begin{aligned} &\frac{1}{2}\,|X^m - P_m X|^2 + \int_0^t \|X^m - P_m X\|^2 ds \\ &= \int_0^t \left(\sqrt{Q}\,(z^m - P_m z), X^m - P_m X\right) ds \\ &\quad + \int_0^t (P_m F_m(X^m) - P_m F(X), X^m - P_m X)\, ds \\ &\le c\int_0^t |\sqrt{Q}\,(z^m - P_m z)|^2 ds + c\int_0^t |f_m(X^m) - X^2|^2 ds + \frac{1}{2}\int_0^t \|X^m - P_m X\|^2 ds. \end{aligned}$$

We deduce

$$\begin{aligned} \mathbb{E}\left(\sup_{t\in[0,T]} |X^m - P_m X|\right) &\le c\mathbb{E}\left(\int_0^T |\sqrt{Q}\,(z_m - P_m z)|^2 ds\right)^{1/2} \\ &+ c\mathbb{E}\left(\int_0^T |f_m(X^m) - X^2|^2 ds\right)^{1/2}. \end{aligned}$$

By standard estimates based on Ito's formula it can be seen that $\{X^m\}$ is bounded in $L^p(\Omega, C([0,T]; L^2(0,1)))$ for any $p \ge 1$. By Sobolev's embedding theorem and the strong convergence of X^m to X in $L^2(\Omega, L^2(0,T; H_0^1(0,1)))$, we can prove

$$\mathbb{E}\left[\left(\int_0^T |f_m(X^m) - X^2|^2 ds\right)^{1/2}\right] \to 0,$$

implying

$$\mathbb{E}\left(\sup_{t\in[0,T]} |X^m - P_m X|\right) \to 0.$$

Since X^m is bounded in any $L^p(\Omega, C([0,T]; L^2(0,1)))$, the conclusion follows. □

A.2. Proof of Lemma 2.2. The existence of η^h and ζ^h solutions of (2.9) and (2.11) is classical. Let Y^x (resp., Y^{x+h}) be the solution of (2.7) with initial datum $x \in H$ (resp., $x+h \in H$). We set

$$r = Y^{x+h} - Y^x - \eta^h.$$

r satisfies the equation

$$\begin{aligned}\frac{dr}{dt} &= Ar + \frac{\partial}{\partial \xi}\left((Y^{x+h})^2 - (Y^x)^2 - 2Y^x\eta^h\right)\\ &= Ar + \frac{\partial}{\partial \xi}\left((Y^{x+h} - Y^x)^2 + 2Y^x r\right).\end{aligned}$$

By similar arguments as in the proof of Lemma 4.3, we have

$$|Y^{x+h}(t) - Y^x(t)|^2 + \int_0^t \|Y^{x+h}(s) - Y^x(s)\|^2 ds \le ce^{c\int_0^t \|Y^x(s)\|^{4/3}ds}\,|h|^2 \tag{A.5}$$

and

$$\begin{aligned}&|r(t)|^2 + \int_0^t \|r(s)\|^2 ds\\ &\le ce^{c\int_0^t \|Y^x(s)\|^{4/3}ds}\int_0^t |Y^{x+h}(s) - Y^x(s)|^3 \|Y^{x+h}(s) - Y^x(s)\| ds.\end{aligned}$$

It follows that

$$|r(t)|^2 + \int_0^t \|r(s)\|^2 ds \le ce^{c\int_0^t \|Y^x(s)\|^{4/3}ds}|h|^4. \tag{A.6}$$

We have

$$\begin{aligned}&|v(t,x+h) - v(t,x) - v_x(t,x)h|\\ &= \mathbb{E}\Bigg(e^{-\frac{1}{2}|Y^{x+h}(t)|^2 - \int_0^t \|Y^{x+h}(s)\|^2 ds} - e^{-\frac{1}{2}|Y^x(t)|^2 - \int_0^t \|Y^x(s)\|^2 ds}\\ &\qquad + \left((Y^x(t), \eta^h(t)) + 2\int_0^t (Y^x(s), \eta^h(s))ds\right) e^{-\frac{1}{2}|Y^x(t)|^2 - \int_0^t \|Y^x(s)\|^2 ds}\Bigg)\\ &= \mathbb{E}\Bigg(\Big(e^{-\frac{1}{2}(|Y^{x+h}(t)|^2 - |Y^x(t)|^2) - \int_0^t (\|Y^{x+h}(s)\|^2 - \|Y^x(s)\|^2)ds}\\ &\qquad -1 + \left(Y^x(t), Y^{x+h}(t) - Y^x(t)\right) + 2\int_0^t \left(Y^x(s), Y^{x+h}(s) - Y^x(s)\right) ds\\ &\qquad -(Y^x(t), r(t)) - 2\int_0^t (Y^x(s), r(s))\, ds\Big) e^{-\frac{1}{2}|Y^x(t)|^2 - \int_0^t \|Y^x(s)\|^2 ds}\Bigg).\end{aligned}$$

By (A.5), (A.6), and elementary inequalities we obtain

$$\begin{aligned}&|v(t,x+h) - v(t,x) - v_x(t,x)h|\\ &\le c\mathbb{E}\Bigg[\left(1 + |Y^x(t)|^2 + \int_0^t \|Y^x(s)\|^2 ds\right)\\ &\qquad\qquad e^{-\frac{1}{2}|Y^x(t)|^2 - \int_0^t \|Y^x(s)\|^2 ds + c\int_0^t \|Y^x(s)\|^{4/3}ds}\Bigg]\,|h|^2.\end{aligned}$$

This proves the differentiability of v. The proof that v is twice differentiable is similar. □

A.3. Proof of Lemma 2.3. We define

$$e_m^h = \eta_m^{P_m h} - P_m \eta^h.$$

By integration by parts, Hölder's inequality, and Agmon's inequality

$$|x|_{L^\infty(0,1)} \le c|x|^{1/2}\|x\|^{1/2},\ x \in H_0^1(0,1),$$

we have

$$\begin{aligned}
&\frac{1}{2}\ \frac{d}{dt}\ |e_m^h|^2 + \|e_m^h\|^2 = -\int_0^1 \Bigg[\big((f_m'(Y_m) - f_m'(Y))\eta_m^h \\
&+f_m'(Y)e_m^h + f_m'(Y)\left((P_m - I)\right)\eta^h + (f_m'(Y) - 2Y)\eta^h\big)\ \frac{\partial}{\partial\xi}e_m^h\Bigg] d\xi \\
&\le \frac{1}{2}\ \|e_m^h\|^2 + c|\eta_m^{P_m h}|\ \|\eta_m^{P_m h}\|\ |Y_m - Y|^2 + c|Y|\|Y\|(|e_m^h|^2 + |(I - P_m)\eta^h|^2) \\
&+c|\eta^h|\ \|\eta^h\|\ |f_m'(Y) - 2Y|^2.
\end{aligned}$$

Thus by Gronwall's lemma

$$\begin{aligned}
&|e_m^h|^2 + \int_0^t \|e_m^h(s)\|^2 ds \le ce^{c\int_0^t |Y(s)|\ \|Y(s)\| ds} \\
&\left(\int_0^t |\eta_m^{P_m h}|\ \|\eta_m^{P_m h}\|\ |Y_m - Y|^2 ds + \int_0^t |Y(s)|\ \|Y(s)\|\ |(I - P_m)\eta^h|^2 ds\right. \\
&\qquad \left. + \int_0^t |\eta^h|\ \|\eta^h\|\ |f_m'(Y) - 2Y|^2 ds\right).
\end{aligned}$$

We write

$$|(I - P_m)\eta^h|^2 \le c\|(I - P_m)\|^2_{\mathcal{L}(D((-A)^{1/4}),H)}|\eta^h|\ \|\eta^h\|,$$

and since Y is almost surely in

$$L^\infty(0,T;L^2(0,1)) \cap L^2(0,T;H_0^1(0,1)),$$

by (4.16) and a similar estimate on η we have

$$\begin{aligned}
&|e_m^h(t)|^2 + \int_0^t \|e_m^h(s)\|^2 ds \le c(\omega)\Bigg(\sup_{t\in[0,T]} |Y_m - Y|^2 \\
&+|I - P_m|_{\mathcal{L}(D((-A)^{1/4}),H)} + c\left(\int_0^T |f_m'(Y) - 2Y|^4 ds\right)^{1/2}\Bigg)|h|^2.
\end{aligned}$$

We have

$$\begin{aligned}
&|f_m'(Y) - 2Y|^4 = \left|\frac{2}{m}\ \frac{Y^3}{(1 + \frac{1}{m}Y^2)^2} - \frac{2}{m^2}\ \frac{Y^4}{(1 + \frac{1}{m}Y^2)^2}\right|^4 \\
&\le \frac{c}{m^2}\ |Y|^8_{L^4(0,1)} \le \frac{c}{m^2}\ |Y|^6\|Y\|^2,
\end{aligned}$$

so that

$$|e_m^h|^2 + \int_0^t \|e_m^h(s)\|^2 ds \le c(\omega)\Bigg(\sup_{t\in[0,T]} |Y_m - Y|^2$$
$$+|I - P_m|_{\mathcal{L}(H)} + \frac{c}{m}\Bigg) |h|^2.$$

The first part of the lemma follows by Lemma 2.1. The proof of the second part goes along the same line. □

REFERENCES

[1] V. Barbu and G. Da Prato, *Hamilton-Jacobi Equations in Hilbert Spaces*, Pitman Res. Notes Math. Ser. 86, Longman, Harlow, UK, 1982.

[2] L. Bertini, N. Cancrini, and G. Jona-Lasinio, *The stochastic Burgers equation*, Comm. Math. Phys., 165 (1994), pp. 211–232.

[3] D.H. Chambers, R.J. Adrian, P. Moin, D.S. Stewart, and H.J. Sung, *Karhunuen-Loeve expansion of Burgers model of turbulence,* Phys. Fluids, 31 (1988), p. 2573.

[4] P. Cannarsa and G. Da Prato, *Some results on nonlinear optimal control problems and Hamilton-Jacobi equations in infinite dimensions*, J. Funct. Anal., 90 (1990), pp. 27–47.

[5] P. Cannarsa and G. Da Prato, *Direct solution of a second order Hamilton-Jacobi equation in Hilbert spaces*, in Stochastic Partial Differential Equations and Applications, Pitman Res. Notes Math. Ser. 268, G. Da Prato and L. Tubaro, eds., Longman, Harlow, UK, 1992, pp. 72–85.

[6] H. Choi, R. Temam, P. Moin, and J. Kim, *Feedback control for unsteady flow and its application to the stochastic Burgers equation*, J. Fluid Mech., 253 (1993), pp. 509–543.

[7] M.G. Crandall, H. Ishi, and P.L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[8] G. Da Prato, *Some results on Bellman equation in Hilbert spaces*, SIAM J. Control Optim., 23 (1985), pp. 61–71.

[9] G. Da Prato, A. Debussche, and R. Temam, *Stochastic Burgers equation*, NoDEA Nonlinear Differential Equations Appl., 1 (1994), pp. 389–402.

[10] G. Da Prato and D. Gątarek, *Stochastic Burgers equation with correlated noise*, Stochastics Stochastics Rep., 52 (1995), pp. 29–41.

[11] G. Da Prato and J. Zabczyk, *Stochastic Evolution Equations in Infinite Dimensions*, Cambridge University Press, London, UK, 1992.

[12] W.H. Fleming and H.M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, Berlin, New York, 1993.

[13] F. Gozzi, *Regularity of solutions of a second order Hamilton–Jacobi equation and application to a control problem*, Comm. Partial Differential Equations, 20 (1995), pp. 775–826.

[14] F. Gozzi, *Global regular solutions of second order Hamilton-Jacobi equations in Hilbert spaces with locally Lipschitz nonlinearities,* J. Math. Anal. Appl., 198 (1996), pp. 399–443.

[15] F. Gozzi and E. Rouy, *Regular solutions of second order stationary Hamilton–Jacobi equations*, J. Differential Equations, 130 (1996), pp. 201–234.

[16] F. Gozzi, E. Rouy, and A. Swiech, *Second Order Hamilton–Jacobi Equations in Hilbert Spaces and Stochastic Boundary Control*, preprint 46, Scuola Normale Superiore di Pisa, Pisa, Italy, 1996.

[17] Y. Kifer, *The Burgers Equation with a Random Force and a General Model for Directed Polymers in Random Environments*, preprint, 1995.

[18] J.L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Gauthier–Villars, Paris, 1969.

[19] P.L. Lions, *Viscosity solutions of fully nonlinear second-order equations and optimal stochastic control in infinite dimensions, Part* I: *The case of bounded stochastic evolution*, Acta Math., 161 (1988), pp. 243–278, *Part* II: *Optimal control of Zakai's equation*, in Stochastic Partial Differential Equations and Applications, Lecture Notes in Math. 1390, G. Da Prato

and L. Tubaro, eds., Springer-Verlag, New York, 1989, *Part* III: *Uniqueness of viscosity solutions for general second order equations*, J. Funct. Anal., 86 (1989), pp. 1–18.

[20] A. Swiech, *Viscosity Solutions of Fully Nonlinear Partial Differential Equations with "Unbounded" Terms in Infinite Dimensions*, Ph.D. thesis, University of California, Santa Barbara, 1993 .

[21] M. Tessitore and J. Zabczyk, *Invariant Measures for Stochastic Heat Equation*, preprint 16, Scuola Normale Superiore di Pisa, Pisa, Italy.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 4, pp. 1150–1161

BEYOND MONOTONICITY IN REGULARIZATION METHODS FOR NONLINEAR COMPLEMENTARITY PROBLEMS[*]

FRANCISCO FACCHINEI[†] AND CHRISTIAN KANZOW[‡]

Abstract. Regularization methods for the solution of nonlinear complementarity problems are standard methods for the solution of monotone complementarity problems and possess strong convergence properties. In this paper, we replace the monotonicity assumption by a P_0-function condition. We show that many properties of regularization methods still hold for this larger class of problems. However, we also provide some counterexamples which indicate that not all results carry over from monotone to P_0-function complementarity problems.

Key words. nonlinear complementarity problem, regularization method, P_0-function, mountain pass theorem

AMS subject classifications. 90C33, 90C31

PII. S0363012997322935

1. Introduction.

We consider the *nonlinear complementarity problem* which is to find a vector in $\mathbb{R}^n$ satisfying the conditions

$$x \geq 0, \;\; F(x) \geq 0, \;\; x^T F(x) = 0;$$

here all inequalities are taken componentwise and $F : \mathbb{R}^n \to \mathbb{R}^n$ is any given function which we assume to be continuously differentiable throughout this paper.

There exist several methods for the solution of the complementarity problem NCP(F); see, e.g., the recent paper [12]. The particular class of methods to be considered in this paper are the so-called *regularization methods*, which are designed to handle *ill-posed* problems. In fact, regularization-type methods have recently been used very successfully in order to improve the robustness of several complementarity solvers on difficult test problems; see [1, 2]. For a detailed discussion of ill-posedness in mathematical programming, we refer the reader to [8]. Very roughly speaking, an ill-posed problem may be difficult to solve since small errors in the computations can lead to a totally wrong solution.

Regularization methods try to circumvent this difficulty by substituting the solution of the original problem with the solution of a sequence of well-posed (i.e., nicely behaved) problems whose solutions form a trajectory converging to the solution of the original problem. In the context of complementarity problems, if we consider the so-called *Tikhonov-regularization*, this scheme consists of solving a sequence of complementarity problems NCP(F_ε)

$$x \geq 0, \;\; F_\varepsilon(x) \geq 0, \;\; x^T F_\varepsilon(x) = 0,$$

where $F_\varepsilon(x) := F(x) + \varepsilon x$ and ε is a positive parameter converging to 0.

[*]Received by the editors June 16, 1997; accepted for publication (in revised form) April 21, 1998; published electronically May 18, 1999. This research was partially supported by NATO grant CRG 960137.

http://www.siam.org/journals/sicon/37-4/32293.html

[†]Dipartimento di Informatica e Sistemistica, Università di Roma "La Sapienza," Via Buonarroti 12, 00185 Roma, Italy (soler@dis.uniroma1.it).

[‡]Institute of Applied Mathematics, University of Hamburg, Bundesstrasse 55, 20146 Hamburg, Germany (kanzow@math.uni-hamburg.de).

Regularization methods for complementarity problems have already been considered in the literature; see, e.g., [28] and [6, Theorem 5.6.2 (b)]. The basic results that can be established in the monotone case and that parallel the classical results for regularization methods for convex optimization problems (see [8] or [25]) are as follows:

(a) The regularized problem NCP(F_ε) has a unique solution $x(\varepsilon)$ for every $\varepsilon > 0$.

(b) The trajectory $x(\varepsilon)$ is continuous for $\varepsilon > 0$.

(c) For $\varepsilon \to 0$, the trajectory $x(\varepsilon)$ converges to the least l_2-norm solution of NCP(F) if NCP(F) has a nonempty solution set; otherwise it diverges.

In this paper, we try to generalize as much as possible the above results to the larger class of P_0 nonlinear complementarity problems. Actually, such a scheme has already been considered in the case of P_0 linear complementarity problems in [32] (see also [6]) and in [9]. These results will be discussed in section 2, where we also show, by an example, the rather counterintuitive fact that if F is a nonlinear P_0-function, then F_ε is not necessarily a uniform P-function. This fact makes the extension of some known results for linear problems to nonlinear problems more difficult than one would expect. In this paper we accomplish this task by an approach based on Fischer's function and the mountain pass theorem. In section 3, we then extend item (a) to the class of P_0-function complementarity problems, whereas section 4 is devoted to the (partial) generalization of items (b) and (c). In section 5 we investigate an algorithm which requires only an approximate solution of the perturbed problems; as far as we are aware of, this is the first implementable algorithm which guarantees that a solution of a P_0 complementarity problem can be computed under the mere assumption that the solution set is nonempty and bounded. We conclude with some final remarks in section 6.

2. Preliminaries.

We first restate some basic definitions.

DEFINITION 2.1. *A matrix $M \in \mathbb{R}^{n\times n}$ is called*

(a) *a P_0-matrix if, for every $x \in \mathbb{R}^n$ with $x \neq 0$, there is an index $i_0 = i_0(x)$ with*

$$x_{i_0} \neq 0 \quad \textit{and} \quad x_{i_0}[Mx]_{i_0} \geq 0;$$

(b) *a P-matrix if, for every $x \in \mathbb{R}^n$ with $x \neq 0$, it holds that*

$$\max_i x_i[Mx]_i > 0;$$

(c) *an R_0-matrix if $x = 0$ is the only solution of NCP(F) for $F(x) := Mx$.*

We refer the reader to the excellent book [6] by Cottle, Pang, and Stone for a discussion of several properties of these classes of matrices. Some nonlinear generalizations of these classes are defined in the following.

DEFINITION 2.2. *The function $F : \mathbb{R}^n \to \mathbb{R}^n$ is called a*

(a) *P_0-function if, for all $x, y \in \mathbb{R}^n$ with $x \neq y$, there is an index $i_0 = i_0(x, y)$ with*

$$x_{i_0} \neq y_{i_0} \quad \textit{and} \quad (x_{i_0} - y_{i_0})[F_{i_0}(x) - F_{i_0}(y)] \geq 0;$$

(b) *P-function if, for all $x, y \in \mathbb{R}^n$ with $x \neq y$, it holds that*

$$\max_i (x_i - y_i)[F_i(x) - F_i(y)] > 0;$$

(c) *uniform P-function if there is a constant $\mu > 0$ such that*

$$\max_i (x_i - y_i)[F_i(x) - F_i(y)] \geq \mu \|x - y\|^2$$

holds for all $x, y \in \mathbb{R}^n$.

Obviously, every uniform P-function is a P-function and every P-function is a P_0-function. Moreover, an affine mapping $F(x) := Mx+q$ is a P_0-function (P-function) if and only if M is a P_0-matrix (P-matrix). Moreover, the class of P_0-functions includes the class of monotone functions. For further discussion, we refer the reader to Moré and Rheinboldt [23].

In the affine case, there are some known results for regularization methods which partially generalize the properties (a) and (c) illustrated in the introduction from monotone to P_0 problems. We summarize these results in the following theorem.

Theorem 2.3. *Assume that $F(x) = Mx + q$ with $M \in \mathbb{R}^{n \times n}$ being a P_0-matrix and $q \in \mathbb{R}^n$. Then*

(a) *the regularized problem NCP(F_ε) has a unique solution $x(\varepsilon)$ for every $\varepsilon > 0$.*

(b) *if M is also an R_0-matrix, then the sequence $x(\varepsilon)$ is bounded for $\varepsilon \to 0$, and every limit point is a solution of NCP(F).*

A proof of these results can be found in [6, Theorem 5.6.2 (a)]. Note also that, in [32], item (b) is proved under an assumption which implies that the original problem has a unique solution. A relaxation of the R_0-condition is also discussed in the recent paper [9] by Ebiefung. In the linear case the proof of statement (a) is quite simple because if M is a P_0-matrix, then $M + \varepsilon I$ is a P-matrix by Theorem 3.4.2 in [6], so that NCP(F_ε) has a unique solution by Theorem 3.3.7 in [6].

Therefore, in an attempt to extend the previous results from the linear to the nonlinear case, the following question seems very natural: Is F_ε a uniform P-function for every fixed $\varepsilon > 0$ if F itself is a P_0-function? If the answer to this question were in the affirmative, then item (a) above could readily be extended, since a complementarity problem with a uniform P-function has a unique solution ([21, Corollary 3.2]). Unfortunately, the following example shows that F_ε is not necessarily a uniform P-function over $\mathbb{R}^n_+$ when F is nonlinear.

Example 2.4. *Consider the function F: $\mathbb{R}^2 \to \mathbb{R}^2$ defined by*

$$F(x) := F(x_1, x_2) := \begin{pmatrix} 0 \\ -e^{x_1} \end{pmatrix}.$$

Since the Jacobian

$$F'(x) = \begin{pmatrix} 0 & 0 \\ -e^{x_1} & 0 \end{pmatrix}$$

is obviously a P_0-matrix for all $x \in \mathbb{R}^2$, the function F itself is a P_0-function by Corollary 5.3 *in* [23]. *Now let $\varepsilon > 0$ and define*

$$F_\varepsilon(x) = F(x) + \varepsilon x = \begin{pmatrix} \varepsilon x_1 \\ \varepsilon x_2 - e^{x_1} \end{pmatrix}.$$

We want to show that F_ε is not a uniform P-function on $\mathbb{R}^n_+$. This means that we want to show that, given a fixed value ε, we can find, for every fixed value μ, two points in $\mathbb{R}^n_+$ (possibly depending on μ) for which the definition of a uniform P-function is not satisfied with that μ.

We will actually show that F_ε is not a uniform P-function for every positive ε. So suppose that $\varepsilon > 0$ is fixed. Choose a positive μ. Consider the following point $x = (x_1, x_2)$:

$$x_1 = 1, \quad x_2 = \sqrt{\frac{\varepsilon}{\mu}}(c-1), \tag{2.1}$$

where c is a constant such that

$$c \geq 2, \tag{2.2}$$

$$\frac{\varepsilon^2}{\mu}(c-1)^2 - \sqrt{\frac{\varepsilon}{\mu}}(e^c - e^1) \leq \varepsilon. \tag{2.3}$$

Note that it is always possible to choose c large enough so that (2.3) *is satisfied; in fact the second term on the left-hand side of* (2.3) *is negative and decreases exponentially with c and hence dominates the first term. Multiplying* (2.3) *by $(c-1)^2$, we also obtain*

$$\frac{\varepsilon^2}{\mu}(c-1)^4 - \sqrt{\frac{\varepsilon}{\mu}}(c-1)^2(e^c - e^1) \leq \varepsilon(c-1)^2. \tag{2.4}$$

We also have, by (2.1),

$$\varepsilon(c-1)^2 < \mu + \varepsilon(c-1)^2 = \mu\left(1 + \frac{\varepsilon}{\mu}(c-1)^2\right) = \mu(x_1^2 + x_2^2). \tag{2.5}$$

Set $y = cx$. Then

$$\begin{aligned}
&\max_{i\in\{1,2\}} (x_i - y_i)[F_{\varepsilon,i}(x) - F_{\varepsilon,i}(y)] \\
&= \max\left\{\varepsilon(x_1-y_1)^2, \varepsilon(x_2-y_2)^2 + (x_2-y_2)(e^{y_1} - e^{x_1})\right\} \\
&= \max\left\{\varepsilon(c-1)^2 x_1^2, \varepsilon(c-1)^2 x_2^2 - (c-1)x_2(e^{cx_1} - e^{x_1})\right\} \\
&\overset{(2.1)}{=} \max\left\{\varepsilon(c-1)^2, \frac{\varepsilon^2}{\mu}(c-1)^4 - \sqrt{\frac{\varepsilon}{\mu}}(c-1)^2(e^c - e^1)\right\} \\
&\overset{(2.4)}{=} \varepsilon(c-1)^2 \\
&\overset{(2.5)}{<} \mu(x_1^2 + x_2^2) \\
&= \frac{\mu}{(c-1)^2}\|x-y\|_2^2 \\
&\overset{(2.2)}{\leq} \mu\|x-y\|_2^2.
\end{aligned}$$

Hence F_ε is not a uniform P-function.

In the next section we shall show that, in spite of the fact that F_ε is not necessarily a uniform P-function, the regularized problems NCP(F_ε) have a unique solution $x(\varepsilon)$ for every $\varepsilon > 0$. However, due to Example 2.4, the analysis is more complicated than one would expect.

3. Existence of regularized solutions. In this section, we want to prove that the regularized problem NCP(F_ε) has a unique solution $x(\varepsilon)$ for every $\varepsilon > 0$. The main tool for proving this result is the (nonsmooth) function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$\varphi(a,b) := \sqrt{a^2 + b^2} - a - b.$$

This function was introduced by Fischer [13] and plays a central role in the design of several nonsmooth Newton-type methods for the solution of NCP(F); see, e.g., [11, 7, 5]. Here, however, we use this function as a theoretical tool. To this end, let us introduce the operator $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ by

$$\Phi(x) := \begin{pmatrix} \varphi(x_1, F_1(x)) \\ \vdots \\ \varphi(x_n, F_n(x)) \end{pmatrix}$$

as well as the corresponding merit function $\Psi : \mathbb{R}^n \to \mathbb{R}$ by

$$\Psi(x) := \frac{1}{2}\Phi(x)^T\Phi(x) = \frac{1}{2}\|\Phi(x)\|^2.$$

We summarize some of the elementary properties of these functions in the following result (see, e.g., [14, 11, 7]).

PROPOSITION 3.1. *The following statements hold:*

(a) *$x^* \in \mathbb{R}^n$ solves NCP(F) if and only if x^* solves the nonlinear system of equations $\Phi(x) = 0$.*

(b) *The merit function Ψ is continuously differentiable on the whole space $\mathbb{R}^n$.*

(c) *If F is a P_0-function, then every stationary point of Ψ is a solution of NCP(F).*

For the regularized problem, we define the corresponding operator and the corresponding merit function similarly by

$$\Phi_\varepsilon(x) := \begin{pmatrix} \varphi(x_1, F_{\varepsilon,1}(x)) \\ \vdots \\ \varphi(x_n, F_{\varepsilon,n}(x)) \end{pmatrix}$$

and

$$\Psi_\varepsilon(x) := \frac{1}{2}\Phi_\varepsilon(x)^T\Phi_\varepsilon(x),$$

where $F_{\varepsilon,i}$ denotes the ith component function of F_ε. The main result of this section is based on the following three preliminary results.

LEMMA 3.2. *Let $\varepsilon > 0$ be arbitrary. Then the Jacobian matrices $F'_\varepsilon(x)$ are P-matrices for all $x \in \mathbb{R}^n$. In particular, the function $F_\varepsilon : \mathbb{R}^n \to \mathbb{R}^n$ is a P-function.*

Proof. Since F is a P_0-function, the Jacobian matrices $F'(x)$ are P_0-matrices for all $x \in \mathbb{R}^n$ by Theorem 5.8 in [23]. In view of Theorem 3.4.2 in [6], the Jacobian matrices $F'_\varepsilon(x) = F'(x) + \varepsilon I$ are therefore P-matrices for all $x \in \mathbb{R}^n$. Hence F_ε is a P-function by Theorem 5.2 in [23]. □

A proof of the following simple result can be found in [18].

LEMMA 3.3. *Let $\{a^k\}, \{b^k\} \subseteq \mathbb{R}$ be any two sequences such that $a^k, b^k \to +\infty$ or $a^k \to -\infty$ or $b^k \to -\infty$. Then $|\varphi(a^k, b^k)| \to \infty$.*

The following proposition contains the main step in order to prove the existence of a solution of the regularized problems NCP(F_ε).

PROPOSITION 3.4. *Suppose that F is a P_0-function and $\varepsilon > 0$. Then the merit function Ψ_ε is coercive, i.e.,*

$$\lim_{\|x\|\to\infty} \Psi_\varepsilon(x) = +\infty.$$

Proof. Suppose by contradiction that the theorem is false. Then we can find an unbounded sequence $\{x^k\}$ such that $\{\Psi_\varepsilon(x^k)\}$ is bounded. Since the sequence $\{x^k\}$ is unbounded, the index set $J := \{i \in \{1,\dots,n\} | \{x_i^k\} \text{ is unbounded}\}$ is nonempty. Subsequencing if necessary, we can assume without loss of generality that $\{|x_j^k|\} \to +\infty$ for all $j \in J$. Let $\{y^k\}$ denote the bounded sequence defined in the following way:

$$y_i^k := \begin{cases} 0 & \text{if } i \in J, \\ x_i^k & \text{if } i \notin J. \end{cases}$$

From the definition of $\{y^k\}$ and the assumption that F is a P_0-function, we get

$$\begin{array}{rcl} 0 & \le & \max_{1\le i\le n}(x_i^k - y_i^k)[F_i(x^k) - F_i(y^k)] \\ & = & \max_{i\in J} x_i^k[F_i(x^k) - F_i(y^k)] \\ & = & x_j^k[F_j(x^k) - F_j(y^k)], \end{array} \tag{3.1}$$

where j is one of the indices for which the max is attained, which we have, without loss of generality, assumed to be independent of k. Since $j \in J$, we have that

$$\{|x_j^k|\} \to \infty. \tag{3.2}$$

We now consider two cases.

Case 1. $x_j^k \to +\infty$. In this case, since $F_j(y^k)$ is bounded by the continuity of F_j, (3.1) implies that $F_j(x^k)$ does not tend to $-\infty$. This in turn implies

$$\left\{\sqrt{(x_j^k)^2 + (F_j(x^k) + \varepsilon(x_j^k))^2} - x_j^k - (F_j(x^k) + \varepsilon x_j^k)\right\} \to +\infty$$

by Lemma 3.3 since $F_j(x^k) + \varepsilon x_j^k$ tends to $+\infty$.

Case 2. $x_j^k \to -\infty$. In this case it follows immediately from Lemma 3.3 that

$$\left\{\sqrt{(x_j^k)^2 + (F_j(x^k) + \varepsilon(x_j^k))^2} - x_j^k - (F_j(x^k) + \varepsilon x_j^k)\right\} \to +\infty$$

(both if $F_j(x^k) + \varepsilon x_j^k$ is unbounded or not).

In either case we get $\Psi_\varepsilon(x^k) \to +\infty$, thus contradicting the boundedness of the sequence $\{\Psi_\varepsilon(x^k)\}$. □

Note that Proposition 3.4 can also be stated in an equivalent way by saying that the level sets $\mathcal{L}_\varepsilon(c) := \{x \in \mathbb{R}^n |\, \Psi_\varepsilon(x) \le c\}$ are compact for every $c \in \mathbb{R}^n$. We are now in a position to prove the following existence and uniqueness result.

THEOREM 3.5. *Assume that F is a P_0-function. Then the regularized complementarity problem NCP(F_ε) has a unique solution $x(\varepsilon)$ for every $\varepsilon > 0$.*

Proof. Let $\varepsilon > 0$. Then F_ε is a P-function by Lemma 3.2. Therefore NCP(F_ε) has at most one solution by Theorem 2.3 in [22].

In order to prove the existence of a solution, let $x^0 \in \mathbb{R}^n$ be arbitrary and define $c := \Psi_\varepsilon(x^0)$. Because of Proposition 3.4, the corresponding level set $\mathcal{L}_\varepsilon(c)$ is nonempty and compact. Hence the continuous function Ψ_ε attains a global minimum x_ε on $\mathcal{L}(c)$ which, in view of the definition of the level set, is also a global minimum of Ψ_ε on $\mathbb{R}^n$. Therefore x_ε is a stationary point of Ψ_ε. However, F_ε is a P-function; in particular, F_ε

itself is a P_0-function, so that x_ε must be a solution of NCP(F_ε) because of Proposition 3.1 (c). ◻

Remark 3.6. It was pointed out to us by Gowda [15] (see also [27]) that the existence of the regularized solutions $x(\varepsilon)$ can also be deduced, with a possibly marginally shorter proof, by Theorem 3.4 in [20]. This offers an interesting alternative point of view. The approach we use here is more algorithmically oriented and Proposition 3.4 has a great practical significance, as will become clear in section 5.

4. Behavior of the solution path. The aim of this section is to study the properties of the solution path $\mathcal{P} := \{x(\varepsilon) | \, \varepsilon > 0\}$ and, in particular, conditions under which $x(\varepsilon)$ remains bounded when $\varepsilon \to 0$. We are interested in the boundedness of $x(\varepsilon)$ because the following easily verifiable result holds.

THEOREM 4.1. *Let $\{\varepsilon_k\}$ be a sequence of positive values converging to* 0. *If $\{x(\varepsilon_k)\}$ converges to a point $\bar{x}$, then $\bar{x}$ solves NCP(F).*

The first noteworthy property we can establish is the continuity of $x(\varepsilon)$.

LEMMA 4.2. *Assume that F is a P_0-function. Then the mapping $\varepsilon \mapsto x(\varepsilon)$ is continuous at any $\varepsilon > 0$.*

Proof. By Lemma 3.2, the Jacobian matrix $F'_\varepsilon(x)$ is a P-matrix for every $\varepsilon > 0$ and every $x \in \mathbb{R}^n$; in particular, $M := F'_\varepsilon(x(\varepsilon))$ is a P-matrix. This immediately implies that every principal submatrix of M is again a P-matrix. Moreover, using the same technique of proof as for Lemma 2.3 in [3], it is easy to see that any Schur-complement of a P-matrix is also a P-matrix. Hence the assertion follows from Theorem 3.1 in Kyparisis [19]. ◻

Note that Lemma 4.2 does not say anything about the continuity of the mapping $\varepsilon \mapsto x(\varepsilon)$ at $\varepsilon = 0$. Continuity at 0 is equivalent to convergence of the solution path $x(\varepsilon)$ when ε goes to 0. As discussed in the introduction, this result holds if F is monotone and the complementarity problem admits a solution. In the more general setting we are considering, we are no longer able to prove such a strong result. However, we can state the following result.

THEOREM 4.3. *Let F be a P_0-function and assume that the solution set $\mathcal{S}$ of NCP(F) is nonempty and bounded. Then the path $\mathcal{P}_{\bar{\varepsilon}} = \{x(\varepsilon) | \, \varepsilon \in (0, \bar{\varepsilon}]\}$ is bounded for any positive $\bar{\varepsilon}$ and*

$$\lim_{\varepsilon \downarrow 0} \operatorname{dist}(x(\varepsilon) | \mathcal{S}) = 0.$$

We postpone the proof of this theorem until the next section, where it will follow from a more general result.

We next state two immediate consequences of Theorem 4.3.

COROLLARY 4.4. *Let F be a P_0-function and assume that NCP(F) has a unique solution $\bar{x}$. Then $\lim_{\varepsilon \downarrow 0} x(\varepsilon) = \bar{x}$.*

Due to a recent result in [10], the uniqueness of a solution of NCP(F) is, for P_0 complementarity problems, equivalent to the existence of an isolated solution of NCP(F). Hence, alternatively, we could have stated Corollary 4.4 under the assumption that NCP(F) has a locally isolated solution.

COROLLARY 4.5. *Let $F(x) = Mx + q$ be an affine mapping with $M \in \mathbb{R}^{n \times n}$ being a P_0- and R_0-matrix. Then the path $\mathcal{P}_{\bar{\varepsilon}}$ is bounded for any positive $\bar{\varepsilon}$ and*

$$\lim_{\varepsilon \downarrow 0} \operatorname{dist}(x(\varepsilon) | \mathcal{S}) = 0.$$

Proof. Since the solution set of NCP(F) is known to be nonempty and bounded under the stated assumptions (see [6]), the result follows immediately from Theorem 4.3. □

Note that Corollary 4.5 is already known (see Theorem 5.6.2 (a) in [6], restated in Theorem 2.3 of this paper); however, our proof is completely different from the one given in [6]. Moreover, it is easy to see that Corollary 4.5 can easily be extended to nonlinear functions F if we assume that F is a P_0-function and an R_0-function. The definition of the latter class of functions as well as some of its properties are given in the recent paper [4]; see also [31].

The following counterexample shows that it is not possible to remove the boundedness assumption of the solution set $\mathcal{S}$ in Theorem 4.3 without destroying the boundedness of the path $\mathcal{P}$. This contrasts sharply with what happens in the case of monotone complementarity problems, where we always have the boundedness of the trajectory if the solution set is nonempty.

Example 4.6. *Let $F : \mathbb{R}^2 \to \mathbb{R}^2$ be defined by $F(x) := Mx + q$, where*

$$M := \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad q := \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

Obviously, F is a P_0-function. The solution set $\mathcal{S}$ is given by

$$\mathcal{S} := \{(x_1, x_2) \,|\, (x_1, 1), x_1 \geq 0\} \cup \{(x_1, x_2) \,|\, (0, x_2), x_2 \geq 1\},$$

i.e., the solution set is unbounded. It is easy to see that $x(\varepsilon) := (1/\varepsilon, 0)$ is the unique solution of the corresponding regularized problem NCP(F_ε). Obviously, $x(\varepsilon)$ is neither convergent nor bounded for $\varepsilon \to 0$. Even worse, the distance of $x(\varepsilon)$ to the solution set $\mathcal{S}$ does not go to zero since dist$(x(\varepsilon)|\mathcal{S}) = 1$ for every $\varepsilon > 0$.

5. Inexact regularization methods.

In the previous section, we illustrated several properties of the trajectory $\mathcal{P}$ which suggest that the original problem NCP(F) can be solved by calculating the exact solutions of a sequence of regularized problems NCP(F_ε) for a sequence of parameters ε converging to 0. From a practical point of view, however, it is usually not possible to solve the regularized problems NCP(F_ε) exactly for each $\varepsilon > 0$. In the following, we therefore present an algorithm which only requires inexact solutions of these subproblems and which nevertheless preserves all the convergence properties of its exact counterpart.

Algorithm 5.1 (inexact regularization method).

(S.0) *Choose $\varepsilon_0 > 0, \alpha_0 \geq 0$, and set $k := 0$.*

(S.1) *Compute an approximate solution $x^k \in \mathbb{R}^n$ of NCP(F_{ε_k}) such that*

$$\Psi_{\varepsilon_k}(x^k) \leq \alpha_k.$$

(S.2) *Terminate the iteration if a suitable stopping criterion is satisfied.*

(S.3) *Choose $\varepsilon_{k+1} > 0, \alpha_{k+1} \geq 0$, set $k \leftarrow k+1$, and go to (S.1).*

Obviously, if we take $\alpha_k = 0$ at each iteration, we have $x^k = x(\varepsilon_k)$. Note that a point x^k satisfying $\Psi_{\varepsilon_k}(x^k) \leq \alpha_k$ can easily be obtained by, e.g., applying any unconstrained minimization technique to Ψ_{ε_k}. In fact, the level sets of Ψ_{ε_k} are compact and every stationary point $\bar{x}$ of Ψ_{ε_k} is such that $\Psi_{\varepsilon_k}(\bar{x}) = 0$. Therefore, every suitable minimization algorithm will produce a minimizing sequence and the point x^k can be surely determined in a finite number of steps. This situation reflects the fact that the perturbed problems are well-posed and this, in turn, is one of the main motivations for using regularization methods.

To establish a result generalizing Theorem 4.3 to Algorithm 5.1 we need some further technical results.

LEMMA 5.2. *Let $C \subset \mathbb{R}^n$ be a compact set. Then, for every $\delta > 0$, there exists a $\bar{\varepsilon} > 0$ such that*

$$|\Psi_\varepsilon(x) - \Psi(x)| \leq \delta$$

for all $x \in C$ and all $\varepsilon \in [0, \bar{\varepsilon}]$.

Proof. The function $\Psi_\varepsilon(x)$ viewed as a function of both x and ε is continuous on the compact set $C \times [0, \bar{\varepsilon}]$. The lemma is then an immediate consequence of the fact that every continuous function on a compact set is uniformly continuous there. □

Finally, we also restate a version of the famous mountain pass theorem which is suitable for our purposes and which can easily be derived from standard statements of this theorem; see, e.g., Theorem 9.2.7 in [24].

THEOREM 5.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and coercive. Let $C \subset \mathbb{R}^n$ be a nonempty and compact set and define m to be the least value of f on the (compact) boundary of C:*

$$m := \min_{x \in \partial C} f(x).$$

Assume further that there are two points $a \in C$ and $b \notin C$ such that $f(a) < m$ and $f(b) < m$. Then there exists a point $c \in \mathbb{R}^n$ such that $\nabla f(c) = 0$ and $f(c) \geq m$.

In the convergence analysis of Algorithm 5.1, we will implicitly assume that Algorithm 5.1 generates an infinite sequence so that the termination criterion in step $(S.2)$ is never active. The following result is our main convergence theorem for Algorithm 5.1. To the authors' knowledge this convergence theorem is new even for monotone complementarity problems.

THEOREM 5.4. *Let F be a P_0-function and assume that the solution set $\mathcal{S}$ of NCP(F) is nonempty and bounded. Suppose that a sequence $\{x^k\}$ is generated according to Algorithm 5.1. If $\varepsilon_k \to 0$ and $\alpha_k \to 0$, then $\{x^k\}$ remains bounded, and every accumulation point of $\{x^k\}$ is a solution of NCP(F).*

Proof. We first note that it follows from a simple continuity argument that every accumulation point of the sequence $\{x^k\}$ is a solution of NCP(F). Hence it remains to be shown that $\{x^k\}$ is a bounded sequence. Assume that the sequence $\{x^k\}$ is not bounded. Then, subsequencing if necessary, we have $\{\|x^k\|\} \to \infty$. Hence there exists a compact set $C \subset \mathbb{R}^n$ with $\mathcal{S} \subset \text{int}C$ and

$$x^k \notin C \tag{5.1}$$

for all k sufficiently large. Let $a \in \mathcal{S}$ be an arbitrary solution of NCP(F). Then we have

$$\Psi(a) = 0.$$

Since

$$\bar{m} := \min_{x \in \partial C} \Psi(x) > 0,$$

we can apply Lemma 5.2 with $\delta := \bar{m}/4$ and conclude that

$$\Psi_{\varepsilon_k}(a) \leq \frac{1}{4}\bar{m} \tag{5.2}$$

and

$$m := \min_{x \in \partial C} \Psi_{\varepsilon_k}(x) \geq \frac{3}{4}\bar{m} \tag{5.3}$$

for all k sufficiently large. Since $\Psi_{\varepsilon_k}(x^k) \leq \alpha_k$ by step $(S.1)$ of Algorithm 5.1, we have

$$\Psi_{\varepsilon_k}(x^k) \leq \frac{1}{4}\bar{m} \tag{5.4}$$

for all k large enough since $\alpha_k \to 0$ by our assumption. Now let us fix an index k such that (5.1)–(5.4) hold. Applying the mountain pass Theorem 5.3 with $b := x^k$, we obtain the existence of a vector $c \in \mathbb{R}^n$ such that

$$\nabla \Psi_{\varepsilon_k}(c) = 0 \quad \text{and} \quad \Psi_{\varepsilon_k}(c) \geq \frac{3}{4}\bar{m} > 0.$$

In view of Proposition 3.1(c), however, the stationary point c of Ψ_{ε_k} must be a global minimizer of Ψ_{ε_k} which gives us the desired contradiction. □

Obviously, Theorem 4.3 follows from Theorem 5.4 by taking $\alpha_k = 0$ for all k and using Theorem 4.1. Also, Corollaries 4.4 and 4.5 can easily be extended to the inexact framework.

Corollary 5.5. *Assume that F is a P_0-function and suppose that a sequence $\{x^k\}$ is generated according to Algorithm* 5.1. *Suppose that $\varepsilon_k \to 0$ and $\alpha_k \to 0$. Then, if NCP(F) has a unique solution $\bar{x}$, we have*

$$\lim_{\varepsilon_k \to 0} x^k = \bar{x}.$$

Corollary 5.6. *Let $F(x) = Mx + q$ be an affine mapping with $M \in \mathbb{R}^{n \times n}$ being a P_0- and R_0-matrix. Assume that $\{x^k\}$ is any sequence generated by Algorithm* 5.1 *such that $\varepsilon_k \to 0$ and $\alpha_k \to 0$. Then the sequence $\{x^k\}$ is bounded, and every accumulation point of the sequence $\{x^k\}$ is a solution of NCP(F).*

If F is a monotone function such that NCP(F) is strictly feasible (i.e., there exists a vector $\hat{x} \in \mathbb{R}^n$ such that $\hat{x} > 0$ and $F(\hat{x}) > 0$), then it is known ([17, Theorem 3.4]) that NCP(F) has a nonempty and bounded solution set. Hence we also obtain the following corollary from our main result (Theorem 5.4) of this section.

Corollary 5.7. *Assume that F is a monotone function such that NCP(F) is strictly feasible. Suppose that $\varepsilon_k \to 0$ and $\alpha_k \to 0$. Then any sequence $\{x^k\}$ generated by Algorithm* 5.1 *remains bounded, and every accumulation point of $\{x^k\}$ is a solution of NCP(F).*

We finally stress that, as far as we know, the inexact regularization method, Algorithm 5.1, investigated in this section is the first (implementable) algorithm which guarantees that a solution of a P_0-function complementarity problem with a bounded and nonempty solution set can actually be computed.

6. Final remarks. In this paper we have shown that, under appropriate assumptions, regularization methods can be successfully applied to P_0 complementarity problems. However, some properties that hold in the monotone case are lost. In particular, when the solution set of the problem is unbounded we can no longer guarantee that the trajectory generated by the regularization method is bounded. There is an open question which we think could be interesting to investigate further. When the solution trajectory $x(\varepsilon)$ is bounded, does it converge and, if it does converge, to which element? In the monotone case, $x(\varepsilon)$ always converges to the least l_2-norm solution of NCP(F). In the P_0 case, the least l_2-norm solution can even be not unique, since the solution set is not necessarily convex.

After the completion of this paper, several contributions appeared on these issues. In particular, in [30] it has been shown that every limit point of $x(\varepsilon)$ is a weak Pareto minimal element of $\mathcal{S}$ and that $x(\varepsilon)$ actually converges if F is polynomial. We also mention that several results presented in this paper have been carefully examined and generalized to a wider class of problems in [16, 26, 27] and that the results of section 5 made it possible to develop superlinearly convergent algorithms for the solution of P_0 complementarity problems with bounded solution sets; see [26, 29].

REFERENCES

[1] S. C. Billups, *Algorithms for Complementarity Problems and Generalized Equations*, Ph.D. Thesis, Computer Sciences Department, University of Wisconsin, Madison, WI, 1995.

[2] S. C. Billups and M. C. Ferris, *QPCOMP: A quadratic programming based solver for mixed complementarity problems*, Math. Programming, 76 (1997), pp. 533–562.

[3] B. Chen and P. T. Harker, *A noninterior continuation method for quadratic and linear programming*, SIAM J. Optim., 3 (1993), pp. 503–515.

[4] B. Chen and P. T. Harker, *Smooth approximations to nonlinear complementarity problems*, SIAM J. Optim., 7 (1997), pp. 403–420.

[5] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983. (Reprinted as Classics in Applied Math. 5, SIAM, Philadelphia, PA, 1990.)

[6] R. W. Cottle, J.-S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, Boston, 1992.

[7] T. De Luca, F. Facchinei, and C. Kanzow, *A semismooth equation approach to the solution of nonlinear complementarity problems*, Math. Programming, 75 (1996), pp. 407–439.

[8] A. L. Dontchev and T. Zolezzi, *Well-Posed Optimization Problems*, Lecture Notes in Math. 1543, Springer-Verlag, New York, 1993.

[9] A. A. Ebiefung, *New perturbation results for solving the linear complementarity problem with P_0-matrices*, Appl. Math. Lett., 11 (1998), pp. 37–39.

[10] F. Facchinei, *Structural and stability properties of P_0 nonlinear complementarity problems*, Math. Oper. Res., 23 (1998), pp. 735–745.

[11] F. Facchinei and J. Soares, *A new merit function for nonlinear complementarity problems and a related algorithm*, SIAM J. Optim., 7 (1997), pp. 225–247.

[12] M. C. Ferris and C. Kanzow, *Recent Developments in the Solution of Nonlinear Complementarity Problems*, preprint, Institute of Applied Mathematics, University of Hamburg, Hamburg, Germany, 1998.

[13] A. Fischer, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.

[14] C. Geiger and C. Kanzow, *On the resolution of monotone complementarity problems*, Comput. Optim. Appl., 5 (1996), pp. 155–173.

[15] M. S. Gowda, *Private communication*, 1997.

[16] M. S. Gowda and M. A. Tawhid, *Existence and limiting behavior of trajectories associated with P_0-equations*, Comput. Optim. Appl., to appear.

[17] P. T. Harker and J.-S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications*, Math. Programming, 48 (1990), pp. 161–220.

[18] C. Kanzow, *Global convergence properties of some iterative methods for linear complementarity problems*, SIAM J. Optim., 6 (1996), pp. 326–341.

[19] J. Kyparisis, *Uniqueness and differentiability of parametric nonlinear complementarity problems*, Math. Programming, 36 (1986), pp. 105–113.

[20] N. Megiddo and M. Kojima, *On the existence and uniqueness of solutions in nonlinear complementarity theory*, Math. Programming, 12 (1977), pp. 110–130.

[21] J. J. Moré, *Coercivity conditions in nonlinear complementarity problems*, SIAM Rev., 16 (1974), pp. 1–16.

[22] J. J. Moré, *Classes of functions and feasibility conditions in nonlinear complementarity problems*, Math. Programming, 6 (1974), pp. 327–338.

[23] J. J. Moré and W. C. Rheinboldt, *On P- and S-functions and related classes of n-dimensional nonlinear mappings*, Linear Algebra Appl., 6 (1973), pp. 45–68.

[24] R. S. Palais and C.-L. Terng, *Critical Point Theory and Submanifold Geometry*, Lectures Notes in Math. 1353, Springer-Verlag, Berlin, 1988.

[25] B. T. Polyak, *Introduction to Optimization*, Optimization Software Inc., New York, NY, 1987.

[26] H. D. Qi, *A regularized smoothing Newton method for box constrained variational inequality problems with P_0 functions*, J. Optim. Theory Appl., to appear.

[27] G. Ravindran and M. S. Gowda, *Regularization of P_0-functions in Box Variational Inequality Problems*, Research Report 97-07, Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD, August 1997 (revised October 1997).

[28] P. K. Subramanian, *A note on least two norm solutions of monotone complementarity problems*, Appl. Math. Lett., 1 (1988), pp. 395–397.

[29] D. Sun, *A regularization Newton method for solving nonlinear complementarity problems*, Appl. Math. Optim., to appear.

[30] R. Sznajder and M. S. Gowda, *On the limiting behavior of the trajectory of regularized solutions of a P_0-complementarity problem*, in Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Norwell, MA, 1998, pp. 371–379.

[31] P. Tseng, *An infeasible path-following method for monotone complementarity problems*, SIAM J. Optim., 7 (1997), pp. 386–402.

[32] V. Venkateswaran, *An algorithm for the linear complementarity problem with a P_0-matrix*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 967–977.

SIAM J. Control Optim.
Vol. 37, No. 4, pp. 1162–1175

MAXIMUM PRINCIPLE FOR A HYBRID SYSTEM VIA SINGULAR PERTURBATIONS[*]

G. GRAMMEL[†]

Abstract. We prove a maximum principle of Pontryagin type for the time optimal control of a hybrid system. The system is nonlinear and consists of a controlled coupled ODE/PDE. The control input acts via the boundary and the interior of the domain. Systems of this type frequently arise in modeling population dynamics in a contaminated environment. For the regularization we use Ekeland's variational principle along with a singular perturbation technique. For this purpose we introduce a new trajectory on an additional exterior domain, whose size may be considered as a singular perturbation parameter. It turns out that the maximum principle is stable with respect to the size of the additional exterior domain. The technique allows us to obtain necessary optimality conditions without involving measure boundary data.

Key words. time optimal boundary control, Pontryagin's maximum principle, Ekeland's variational principle, Ekeland metric, singular perturbation

AMS subject classifications. 49K22, 35B25, 35K22, 93C20

PII. S0363012998332640

1. Introduction.

This paper is concerned with necessary optimality conditions for parabolic boundary control problems. Its main purpose is to provide a regularization technique via singular perturbations to obtain optimality conditions for the time optimal control problem.

In order to obtain a Pontryagin type maximum principle, one usually has to study equations with measure boundary data. The reason is that, for nonlinear or nonconvex problems with state restrictions, one cannot manage without strong variations (or spike perturbations) of the control. This necessarily leads to Dirac measures in the variational system describing the directional derivatives; see [6]. In contrast to nonlinear ODE, where the Dirac impulse just causes a jump of the trajectory, there is no similar interpretation for the boundary control of a parabolic equation. Furthermore, Dirac impulses in the boundary data cause technical difficulties (see [6]) and obviously reduce the regularity of the solutions. For *linear* systems, the explicit representation of the solutions via Green's functions (see, e.g., [7]) still allows us to overcome these difficulties; see [5], [6]. But for *nonlinear* systems, boundary element methods seem to be not as far developed. The situation becomes even more serious when pointwise state constraints are considered (see [10], [3]). Then the adjoint system necessarily has Dirac data, and duality considerations cannot be carried through when the variational system is not smooth enough. For this reason another type of variation recently has been introduced in [10], [3]. Here the variation is not concentrated locally around one point but somehow distributed over the domain, with the effect that the variational system is smooth and even looks like the variational system of a weak variation (L^∞-perturbation) of the control. This technique leads to an integrated form of the maximum principle, which, under slight continuity assumptions, can be transformed into a pointwise maximum principle.

[*] Received by the editors January 9, 1998; accepted for publication (in revised form) October 29, 1998; published electronically May 21, 1999.

http://www.siam.org/journals/sicon/37-4/33264.html

[†] Institut für Informatik und Praktische Mathematik, Christian–Albrechts–Universität zu Kiel, Olshausen Str. 40, 24098 Kiel, Germany (ggr@numerik.uni-kiel.de).

We propose a different method to derive a maximum principle without involving measure boundary data in the variational system. The basic idea is quite simple: since it doesn't cause too much trouble to deal with strong variations for ODEs we interpret the boundary term as the steady state of a fast trajectory of an additional ODE. So, instead of investigating the original system, we approximate it by a singularly perturbed augmented system, obtain necessary optimality conditions for the singularly perturbed system via strong variations, and later let the singular perturbation tend to zero.

The proposed regularization procedure is demonstrated by a hybrid model of mathematical biology, introduced in [2], [1], but it is by far not restricted to this model. Roughly speaking, it works for any boundary control problem which is well-posed with respect to L^1-perturbations (with a common L^∞-bound) of the boundary data—a property which is easily derived (also for semilinear parabolic equations) whenever the boundary problem is well-posed with respect to Ekeland- *and* L^∞-perturbations of the boundary data. In contrast, the standard approach requires well-posedness for the case where L^1-data converge in the distributional sense to a Dirac measure.

The mathematical model of the present paper consists of a coupled pair of a (linear) parabolic equation with Neumann type boundary conditions and a (nonlinear) ODE in a Banach space:

$$(1)\qquad \left\{\begin{array}{rcl} \frac{\partial}{\partial t}Y(x,t) &=& \Delta Y(x,t),\\ \frac{\partial}{\partial N(y)}Y(y,t) &=& \int_\Omega K(x,y,u(t))^T Z(x,t)dx,\\ \frac{\partial}{\partial t}Z(x,t) &=& f(Z(x,t),Y(x,t),u(t)),\\ Y(x,0) &=& Y^0(x),\\ Z(x,0) &=& Z^0(x), \end{array}\right.$$

where $x \in \Omega$, $y \in \partial\Omega$, $t \in (0,T]$ and $\frac{\partial}{\partial N(y)}$ denotes the outer normal derivative. Thus the control input $u(t) \in U$ acts via the boundary and the interior of the domain. We are interested in the following *time optimal control problem.*

Given a function $L : L^1(\Omega; R^n) \to R$, find the control function that steers $Z(\cdot,t) \in L^1(\Omega; R^n)$ in minimal time to the target set, which is given by

$$\{Z \in L^1(\Omega; R^n) : L(Z) = 0\}.$$

The singular perturbation technique is based on the introduction of a new fast boundary trajectory $X(y,t)$, which leads to the *singularly perturbed system*

$$(2)\qquad \left\{\begin{array}{rcl} \frac{\partial}{\partial t}Y(x,t) &=& \Delta Y(x,t),\\ \frac{\partial}{\partial N(y)}Y(y,t) &=& X(y,t) - Y(y,t),\\ Y(x,0) &=& Y^0(x), \end{array}\right.$$

$$(3)\qquad \left\{\begin{array}{rcl} \epsilon(y)\frac{\partial}{\partial t}X(y,t) &=& \int_\Omega K(x,y,u(t))^T Z(x,t)dx - (X(y,t) - Y(y,t)),\\ \frac{\partial}{\partial t}Z(x,t) &=& f(Z(x,t),Y(x,t),u(t)),\\ X(y,0) &=& 0,\\ Z(x,0) &=& Z^0(x). \end{array}\right.$$

Here $\epsilon(y) > 0$ is assumed to be small in order to reflect that, for a given "slow" control u, $X(y,t)$ moves fast towards its (control dependent) equilibrium, which is given by

$$X(y,t) = \int_\Omega K(x,y,u(t))^T Z(x,t)dx + Y(y,t).$$

Note that, in general, the approximation of a nonlinear control system by singularly perturbed systems causes delicate problems, since a “fast” control may force the fast trajectory to perform oscillations, and more general averaging techniques are required; see, e.g., [8], [13], [9]. Nevertheless, to prove the maximum principle via singular perturbations, we can still manage with the steady state approach à la Tychonov (see, e.g., [12]) for two reasons: first, by the linearity of the artificial fast equation for X in (3); second, by the condition of the maximum principle itself that an optimal control for the original time optimal problem *exists*.

Surely, our proposed method is not the only possibly one. First, by the linearity of the diffusion equation, strong variations together with boundary element methods (see, e.g., [6]) seem to be applicable, even if the proof that the nonsmooth variational system really describes the directional derivatives seems to be quite involved. Second, the method of “distributed variations,” as in [3], [10], could be applicable, even for semilinear diffusions, but with the obstacle of constructing a suitable location for the perturbation, which should be quite complicated considering the coupled dynamics. The advantage of our method is that the analysis towards necessary optimality conditions involves a singularly perturbed variational system without Dirac data. This approach is extendable to semilinear diffusions and easily manages complicated coupled dynamics. Furthermore, the singular perturbation parameter has a physical meaning and the regularization method hence implies a sensitivity analysis for the maximum principle against model errors.

Interpretation of the model. Models of this kind appear when modeling population dynamics in a contaminated environment. For example, $Z(x,t) \in R$ could be a population density or, more generally, if in R^n, the densities of several characteristics of a population, that cause an environmental contamination density $Y(x,t)$ by polluting the boundary area. $Y(x,t)$ enters the domain via the boundary and its dynamics within the domain are described by a diffusion. On the other hand, the contamination causes changes in the population characteristics and the circle is closed. In [1] a slightly simpler model is investigated to describe the dynamics of viral diseases. There it is assumed that the infectious agent is transported via diffusion and necessary optimality conditions for optimal control problems *without* terminal state restrictions are derived.

There is a crucial difficulty in connection with the hybrid character of the system (1). Whereas the ODE describes different characteristics of a population distribution, it clearly defines a domain $\Omega \subset R^2$ of interest. In [1] the area consists of a city at the seashore. But if, additionally, a diffusion process is considered, it is not obvious how to relate it to the domain Ω. By this reason any reasonable model should be stable with respect to parameters describing the boundary conditions.

In our singular perturbation approach, we divide the environment into two parts. First, we have the domain Ω, in which the differential equations for $Y(x,t)$, the density of the diffusing material, and for $Z(x,t)$, the characteristics of the population, hold. It is separated by the boundary $\partial\Omega$ from a new outer domain, in which $X(y,t)$, the new outer density of the diffusing material, is assumed to be “radially constant”: every boundary point $y \in \partial\Omega$ corresponds to a line of length $\epsilon(y) > 0$ in the outer domain on which the density is spatially constant. If the outer domain is thin, i.e., if $\epsilon(y) > 0$ is small, the density $X(y,t)$ changes rapidly with respect to time. When $\epsilon(y)$ tends to zero, the density $X(y,t)$ converges rapidly towards its equilibrium. The flow through the boundary $\partial\Omega$ is then determined by the difference $X(y,t) - Y(y,t)$ of outer and inner concentrations. Whereas the term $\int_\Omega K(x,y,u(t))^T Z(x,t)dx$ generates an outer concentration $X(y,t)$, which naturally changes faster, the smaller the size $\epsilon(y)$ of the outer domain. Thus the singular perturbation approach also gives an interpretation for the source term $\int_\Omega K(x,y,u(t))^T Z(x,t)dx$. It measures just the absolute quantity of contaminated material transported to the boundary area. Finally, we remark that a meaningful functional L, describing the terminal set $L(Z) = 0$, is given by $L(Z) = \int_\Omega Z^i(x)dx - k$, for $i = 1, \dots, n$, and a natural number $k \in N$. In the model above, $Z^i(x,t) \in R$ could be the density of the *affected* population, and thus the optimal control problem would consist of reducing the number of affected individuals to a given number $k \in N$ in minimal time.

Organization of the paper. Section 2 contains the setting and the maximum principles for the singularly perturbed system (2), (3) and for the original system (1). In section 3 we briefly show the unique existence of solutions for the singularly perturbed system. The variational system for the singularly perturbed system is introduced in section 4. The last two sections contain the proofs of the maximum principles.

2. Preliminaries and main results.

Throughout the paper we assume the following:

- The admissible control functions $u \in L^\infty([0,T];U)$ are measurable with values in a compact subset $U \subset R^m$ and are equipped with the Ekeland metric (λ denotes the Lebesgue measure on $[0,T]$):

$$d(u_1, u_2) = \lambda\{t \in [0,T] : u_1(t) \neq u_2(t)\}.$$

- The domain $\Omega \subset R^2$ is bounded with smooth boundary $\partial\Omega$.
- The function $f \in C^{1,1,0}(R^n \times R \times R^m; R^n)$ is continuous, bounded, and continuously differentiable with respect to the first and second arguments with bounded derivative. For all $Y \in R$ and $v \in U$ the components of $f(0,Y,v) \in R^n$ are nonnegative.
- The function $\epsilon(\cdot) \in C(\partial\Omega; R)$ is positive and continuous.
- The kernel $K \in C(\overline{\Omega} \times \partial\Omega \times R^m; R^n)$ is continuous with nonnegative components.
- The initial distribution $Y^0 \in C(\overline{\Omega}; R_0)$ is nonnegative and continuous.
- The initial distribution $Z^0 \in L^\infty(\Omega; R^n)$ is measurable, essentially bounded with essentially nonnegative components.
- The function $L : L^1(\Omega, R^n) \to R$ is continuously Frechet differentiable.

Remark 2.1. We did not make any attempts to optimize this setting. There are possibly many ways to weaken the regularity conditions above. In particular, we mention that the space dimension for the domain Ω could be any natural number; in just the special model it is 2. The control range U could be any compact metric space so that more complicated controls, also depending on the spatial variable, could be taken into consideration.

In order to formulate the Pontryagin maximum principle, we define the *adjoint system* with terminal conditions

$$(4) \quad \left\{ \begin{array}{rcl} \frac{\partial}{\partial t}P(x,t) & = & -\Delta P(x,t) - f_Y(Z_0(x,t), Y_0(x,t), u(t))^T Q(x,t), \\ \frac{\partial}{\partial N(y)}P(y,t) & = & 0, \\ \frac{\partial}{\partial t}Q(x,t) & = & -f_Z(Z_0(x,t), Y_0(x,t), u(t))^T Q(x,t) \\ & & - \int_{\partial\Omega} K(x,y,u(t))P(y,t)dy, \\ P(x,T) & = & 0, \\ Q(x,T) & = & \Lambda. \end{array} \right.$$

We denote solutions of the system (1) by Y_0, Z_0 and solutions of the adjoint system by P_0, Q_0. The maximum principle for the time optimal control problem then becomes the following theorem.

THEOREM 2.1. *If $u_0^* : [0, T_0] \to U$ is a time optimal control function for the system* (1) *with arrival time $T = T_0$, then there is a terminal condition $\Lambda \in L^\infty(\Omega; R^n)$ for the adjoint system* (4), *such that for almost all times $s \in [0, T_0]$ and all control values $v \in U$ the maximum condition*

$$\begin{aligned} 0 \leq & \int_{\partial\Omega} \int_{\Omega} (K(x, y, v) - K(x, y, u_0^*(s)))^T Z_0(x, s) P_0(y, s) dx dy \\ & + \int_{\Omega} Q_0(x, s)^T (f(Z_0(x, s), Y_0(x, s), v) - f(Z_0(x, s), Y_0(x, s), u_0^*(s)) dx \end{aligned}$$

is valid. The multiplier Λ fulfills the transversality condition

$$\Lambda = \operatorname{sgn}(L(Z^0)) \nabla L(Z_0(\cdot, T_0)(u_0^*)) \in L^\infty(\Omega; R^n).$$

We prove this Pontryagin maximum principle by regularizing the original system by singularly perturbed systems, so naturally we introduce a *singularly perturbed adjoint system* with terminal conditions:

$$(5) \qquad \left\{ \begin{array}{rcl} \frac{\partial}{\partial t} P(x, t) & = & -\Delta P(x, t) - f_Y(Z_\epsilon(x, t), Y_\epsilon(x, t), u(t))^T Q(x, t), \\ \frac{\partial}{\partial N(y)} P(y, t) & = & R(y, t) - P(y, t), \\ \epsilon(y) \frac{\partial}{\partial t} R(y, t) & = & R(y, t) - P(y, t), \\ \frac{\partial}{\partial t} Q(x, t) & = & -f_Z(Z_\epsilon(x, t), Y_\epsilon(x, t), u(t))^T Q(x, t) \\ & & - \int_{\partial\Omega} K(x, y, u(t)) R(y, t) dy, \\ P(x, T) & = & 0, \\ R(y, T) & = & 0, \\ Q(x, T) & = & \Lambda. \end{array} \right.$$

Throughout the paper, we will denote trajectories of the singularly perturbed system (2), (3) by $Y_\epsilon, X_\epsilon, Z_\epsilon$ and trajectories of the singularly perturbed adjoint system (5) by $P_\epsilon, R_\epsilon, Q_\epsilon$. Although we do not need it for the proof of Theorem 2.1, we also state the maximum principle for the singularly perturbed system because we obtain the proof of the following theorem as a byproduct.

THEOREM 2.2. *If $u^* : [0, T] \to U$ is a time optimal control function for the singularly perturbed system* (2), (3) *with arrival time $T \geq 0$, then there is a terminal condition $\Lambda \in L^\infty(\Omega; R^n)$ for the singularly perturbed adjoint system* (5), *such that for almost all times $s \in [0, T]$ and all control values $v \in U$ the maximum condition*

$$\begin{aligned} 0 \leq & \int_{\partial\Omega} \int_{\Omega} (K(x, y, v) - K(x, y, u^*(s)))^T Z_\epsilon(x, s) R_\epsilon(y, s) dx dy \\ & + \int_{\Omega} Q_\epsilon(x, s)^T (f(Z_\epsilon(x, s), Y_\epsilon(x, s), v) - f(Z_\epsilon(x, s), Y_\epsilon(x, s), u^*(s))) dx \end{aligned}$$

is valid. The multiplier Λ fulfills the transversality condition

$$\Lambda = \operatorname{sgn}(L(Z^0)) \nabla L(Z_\epsilon(\cdot, T)(u^*)) \in L^\infty(\Omega; R^n).$$

The regularization procedure via singular perturbations requires some explanation, especially since the system under consideration is *nonlinear.* For nonlinear singularly perturbed control systems, it is known that, even if the fast subsystem has an exponentially stable equilibrium for any control value $v \in U$ and any frozen slow state, the correct limit system is richer than the one obtained with the standard Tychonov approach. The explanation for this nonlinear phenomenon is that open-loop controls that oscillate between two values can force the fast motion to oscillate, if the oscillation of the open-loop control becomes faster as the perturbation becomes smaller. On the other hand, by the nonlinearity of the system, oscillations of the fast motions produce additional averaged dynamics of the slow motions, which are not captured by the steady state approach à la Tychonov. Hence, in general, the original system (1) is not rich enough to contain all limits of Y_ϵ, Z_ϵ trajectories of the singularly perturbed system (2), (3), as the perturbation tends to zero. A correct limit system for (2), (3) should take into account fast oscillations and could be considered as a generalized relaxed system.

However, in connection with the maximum principle, these difficulties do not occur, since we are interested in only one special control, namely, the optimal control of the time optimal problem. We do not need to care about the possible better trajectories of the singularly perturbed system and their limits (which would be trajectories of a generalized relaxed system) simply by the fact that the maximum principle presumes the existence of an optimal control.

3. Unique solutions of the singularly perturbed systems. In this section we sketch a proof for the unique existence of solutions of the singularly perturbed (adjoint) system and continuous dependence on the control. The corresponding statements for the original (adjoint) system are implicitly proved in Lemma 6.1 in the last section. We introduce appropriate function spaces

$$\begin{aligned} Y_\epsilon, P_\epsilon, Y_0, P_0 &\in C([0,T]; C(\overline{\Omega}; R)) \cap L^2((0,T); H^1(\Omega)),\\ X_\epsilon, R_\epsilon &\in C([0,T]; C(\partial\Omega; R)),\\ Z_\epsilon, Q_\epsilon, Z_0, Q_0 &\in C([0,T]; L^\infty(\Omega; R^n)). \end{aligned}$$

All function spaces are equipped with the standard topology.

PROPOSITION 3.1. *The singularly perturbed system* (2), (3) *and adjoint system* (5) *have unique solutions* $(Y_\epsilon, X_\epsilon, Z_\epsilon)(u)$ *for any* $T \geq 0$ *and any control function* $u \in L^\infty([0,T]; U)$. *The solutions depend continuously on the control function.*

Proof. In order to show the unique existence of solutions $(Y_\epsilon, X_\epsilon, Z_\epsilon)$ of (2), (3), we define the solution operators

$$S_Y : C([0,T]; C(\partial\Omega; R) \times L^\infty(\Omega; R^n)) \to C([0,T]; C(\overline{\Omega}; R)) \cap L^2((0,T); H^1(\Omega)),$$

which map trajectories (X, Z) to solutions Y of the PDE (2). Conversely, we define

$$S^u_{(X,Z)} : C([0,T]; C(\overline{\Omega}; R)) \cap L^2((0,T); H^1(\Omega)) \to C([0,T]; C(\partial\Omega; R) \times L^\infty(\Omega; R^n)),$$

which map trajectories Y (for a given control function u) to solutions (X, Z) of the ODE (3). Note that the operator $S^u_{(X,Z)}$ also depends on the size $\epsilon(y) > 0$ of the additional boundary region. For $T \geq 0$ small enough, the composition

$$S^u_{(X,Z)} \circ S_Y : C([0,T]; C(\partial\Omega; R) \times L^\infty(\Omega; R^n)) \to C([0,T]; C(\partial\Omega; R) \times L^\infty(\Omega; R^n))$$

is a (continuously differentiable) contraction for any control $u \in L^\infty([0,T];U)$: the operator S_Y is affine linear and continuous (see [11] or [3]), thus continuously differentiable and Lipschitz. A standard application of the Gronwall lemma also shows that the nonlinear operator $S^u_{(X,Z)}$ is continuously differentiable and Lipschitz since f is. The Lipschitz constant of the operator $S^u_{(X,Z)}$ defined by the integral equation

$$\begin{cases} X(y,t) &= \int_0^t (\int_\Omega K(x,y,u(\tau))^T Z(x,\tau)dx - (X(y,\tau)+Y(y,\tau)))d\tau, \\ Z(x,t) &= Z^0(x) + \int_0^t f(Z(x,\tau),Y(x,\tau),u(\tau))d\tau \end{cases}$$

becomes arbitrary small, as the time horizon $[0,T]$ becomes smaller. We conclude that, for $T \geq 0$ small enough, the singularly perturbed system (2), (3) has for any control function $u \in L^\infty([0,T];U)$ a unique solution $(X_\epsilon, Z_\epsilon)(u)$, which is just the fixed point of the contraction $S^u_{(X,Z)} \circ S_Y$. Similarly, $Y_\epsilon(u)$ is the fixed point of the contraction $S_Y \circ S^u_{(X,Z)}$. A standard application of the Gronwall lemma shows that the nonlinear operator $S^u_{(X,Z)}$ depends continuously on $u \in L^\infty([0,T];U)$ (in the sense of pointwise convergence). Thus, also, the fixed point of the contraction $S^u_{(X,Z)} \circ S_Y$ depends continuously on $u \in L^\infty([0,T];U)$. For arbitrary $T \geq 0$ we can subdivide the interval $[0,T]$ into smaller subintervals and construct appropriate contractions on each subinterval to show successively the existence of a unique solution on the whole time interval. The corresponding statement for the adjoint system can be proved in essentially the same way and the proof is finished. □

4. The singularly perturbed variational system. The main advantage of our singular perturbation approach is that the variational system, which describes the variations or directional derivatives of the trajectories corresponding to strong variations of the control, contains Dirac impulses only in the ODE part. An investigation of the more complicated variational system for the original system (1) is no longer necessary. In the following we introduce the variational system for the singularly perturbed system and manifest its connections with the singularly perturbed adjoint system.

We use the strong variation for control functions. For a control value $v \in U$, a time $s \in (0,T]$ and $h \in (0,s]$, the strong variation u^h of the control function u is defined by

$$u^h(t) := \begin{cases} v & \text{for almost all } t \in [s-h,s], \\ u(t) & \text{for almost all } t \in [0,s-h] \cup [s,T]. \end{cases} \tag{6}$$

Then the *singularly perturbed variational system* for the limits

$$(V_\epsilon, U_\epsilon, W_\epsilon) = \lim_{h \to 0+} \frac{1}{h}((Y_\epsilon, X_\epsilon, Z_\epsilon)(u^h) - (Y_\epsilon, X_\epsilon, Z_\epsilon)(u)) \tag{7}$$

becomes for $t \in [s,T]$

$$\begin{cases} \frac{\partial}{\partial t}V(x,t) &= \Delta V(x,t), \\ \frac{\partial}{\partial N(y)}V(y,t) &= U(y,t) - V(y,t), \\ \epsilon(y)\frac{\partial}{\partial t}U(y,t) &= \int_\Omega K(x,y,u(t))^T W(x,t)dx - (U(y,t)-V(y,t)), \\ \frac{\partial}{\partial t}W(x,t) &= f_Z(Z_\epsilon(x,t),Y_\epsilon(x,t),u(t))W(x,t), \\ & \quad +f_Y(Z_\epsilon(x,t),Y_\epsilon(x,t),u(t))V(x,t) \\ V(x,s) &= 0, \\ U(y,s) &= \frac{1}{\epsilon(y)}\int_\Omega (K(x,y,v)-K(x,y,u(s)))^T Z(x,s)dx, \\ W(x,s) &= f(Z_\epsilon(x,s),Y_\epsilon(x,s),v) - f(Z_\epsilon(x,s),Y_\epsilon(x,s),u(s)). \end{cases} \tag{8}$$

For $t \in [0, s)$ the variation $(V_\epsilon, U_\epsilon, W_\epsilon)$ is obviously equal to zero; thus, we have the usual jump at time $t = s$, which is caused by a Dirac impulse. We show that the solution of the singularly perturbed variational system (8) in fact describes the variation of the trajectories with respect to the strong variation of the control. The function spaces for the trajectories of the variational system are

$$\begin{aligned} V_\epsilon &\in C([s,T]; C(\overline{\Omega}; R)) \cap L^2((0,T); H^1(\Omega)), \\ U_\epsilon &\in C([s,T]; C(\partial\Omega; R)), \\ W_\epsilon &\in C([s,T]; L^\infty(\Omega; R^n)). \end{aligned}$$

All function spaces are equipped with the standard topology. We omit showing the unique existence of solutions to (8), since the proof is completely analogous to the proof of Proposition 3.1.

LEMMA 4.1. *Let $u^h \in L^\infty([0,T]; U)$ be the strong variation of the control $u \in L^\infty([0,T]; U)$. Then the limit* (7) *exists and is a solution of the singularly perturbed variational system* (8).

Proof. Since the operator $S^u_{(X,Z)} \circ S_Y$ is continuously differentiable, we can write (again for $T > 0$ small enough)

$$\begin{aligned} &\frac{1}{h}((X_\epsilon, Z_\epsilon)(u^h) - (X_\epsilon, Z_\epsilon)(u)) \\ &\quad = \frac{1}{h}\left(S^{u^h}_{(X,Z)} \circ S_Y((X_\epsilon, Z_\epsilon)(u^h)) - S^u_{(X,Z)} \circ S_Y((X_\epsilon, Z_\epsilon)(u))\right) \\ &\quad = \frac{1}{h}\left(S^{u^h}_{(X,Z)} \circ S_Y((X_\epsilon, Z_\epsilon)(u^h)) - S^{u^h}_{(X,Z)} \circ S_Y((X_\epsilon, Z_\epsilon)(u))\right) \\ &\qquad + \frac{1}{h}\left(S^{u^h}_{(X,Z)} \circ S_Y((X_\epsilon, Z_\epsilon)(u)) - S^u_{(X,Z)} \circ S_Y((X_\epsilon, Z_\epsilon)(u))\right) \\ &\quad = D_{(X,Z)}\left(S^{u^h}_{(X,Z)} \circ S_Y((X_\epsilon, Z_\epsilon)(u))\right)\left[\frac{1}{h}((X_\epsilon, Z_\epsilon)(u^h) - (X_\epsilon, Z_\epsilon)(u))\right] + \frac{o(h)}{h} \\ &\qquad + \frac{1}{h}\left(S^{u^h}_{(X,Z)} \circ S_Y - S^u_{(X,Z)} \circ S_Y\right)((X_\epsilon, Z_\epsilon)(u)). \end{aligned}$$

A standard application of the Gronwall lemma shows the convergence

$$\lim_{h \to 0+} \frac{1}{h}\left(S^{u^h}_{(X,Z)} \circ S_Y - S^u_{(X,Z)} \circ S_Y\right)((X_\epsilon, Z_\epsilon)(u)) = (U_2, W_2)$$

in $C([s,T]; C(\partial\Omega; R) \times C(\overline{\Omega}; R^n))$, where (U_2, W_2) fulfill the differential equation

$$\left\{ \begin{array}{rcl} \epsilon(y)\frac{\partial}{\partial t}U_2(y,t) &=& \int_\Omega K(x,y,u(t))^T W_2(x,t)dx - U_2(y,t), \\ \frac{\partial}{\partial t}W_2(x,t) &=& f_Z(Z_\epsilon(x,t), Y_\epsilon(x,t), u(t))W_2(x,t), \\ U_2(y,s) &=& \frac{1}{\epsilon(y)}\int_\Omega (K(x,y,v) - K(x,y,u(s)))^T Z_\epsilon(x,s)dx, \\ W_2(x,s) &=& f(Z_\epsilon(x,s), Y_\epsilon(x,s), v) - f(Z_\epsilon(x,s), Y_\epsilon(x,s), u(s)). \end{array} \right.$$

At the same time we have the convergence of the bounded linear operator (in the sense of the operator norm)

$$\lim_{h \to 0+} D_{(X,Z)}\left(S^{u^h}_{(X,Z)} \circ S_Y((X_\epsilon, Z_\epsilon)(u))\right) = D_{(X,Z)}\left(S^u_{(X,Z)} \circ S_Y((X_\epsilon, Z_\epsilon)(u))\right),$$

where the limit operator maps $(U,W)\in C([s,T];C(\partial\Omega;R)\times L^\infty(\Omega;R^n))$ to solutions (U_1,W_1) of the system

$$\left\{\begin{array}{rcl} \frac{\partial}{\partial t}V(x,t) & = & \Delta V(x,t),\\ \frac{\partial}{\partial N}V(y,t) & = & (U(y,t)-V(y,t)),\\ \epsilon(y)\frac{\partial}{\partial t}U_1(y,t) & = & \int_\Omega K(x,y,u(t))^T W_1(x,t)dx-(U_1(y,t)-V(y,t)),\\ \frac{\partial}{\partial t}W_1(x,t) & = & f_Z(Z_\epsilon(x,t),Y_\epsilon(x,t),u(t))W_1(x,t)\\ & & +f_Y(Z_\epsilon(x,t),Y_\epsilon(x,t),u(t))V(x,t),\\ V(x,s) & = & 0,\\ U_1(y,s) & = & 0,\\ W_1(x,s) & = & 0. \end{array}\right.$$

Thus $\frac{1}{h}((X_\epsilon,Z_\epsilon)(u^h)-(X_\epsilon,Z_\epsilon)(u))$ is a fixed point of a contraction that depends continuously on h. Note that the contraction constant does not depend on h! We conclude that the limit $(U_\epsilon,W_\epsilon)=\lim_{h\to 0+}\frac{1}{h}((X_\epsilon,Z_\epsilon)(u^h)-(X_\epsilon,Z_\epsilon)(u))$ exists and fulfills the equation

$$(U_\epsilon,W_\epsilon)=D_{(X,Z)}\left(S^u_{(X,Z)}\circ S_Y((X_\epsilon,Z_\epsilon)(u))\right)[(U,W)]+(U_2,W_2).$$

It follows that $(U_\epsilon,W_\epsilon)=(U_1,W_1)+(U_2,W_2)$ is a solution of the variational system and the proof is finished. □

We close this section with a lemma describing the relationship between the singularly perturbed variational system (8) and the singularly perturbed adjoint system (5).

Lemma 4.2. *For all times $s\in(0,T]$, all times $t\in[s,T]$, and all control values $v\in U$ the trajectories of the singularly perturbed variational system* (8) *and of the singularly perturbed adjoint system* (5) *fulfill*

$$\begin{aligned} &\int_\Omega (P_\epsilon(x,s)V_\epsilon(x,s)+Q_\epsilon(x,s)^T W_\epsilon(x,s))dx+\int_{\partial\Omega}\epsilon(y)R_\epsilon(y,s)U_\epsilon(y,s)dy\\ &\qquad=\int_\Omega (P_\epsilon(x,t)V_\epsilon(x,t)+Q_\epsilon(x,t)^T W_\epsilon(x,t))dx+\int_{\partial\Omega}\epsilon(y)R_\epsilon(y,t)U_\epsilon(y,t)dy\\ &\qquad=\int_\Omega Q_\epsilon(x,T)^T W_\epsilon(x,T)dx. \end{aligned}\tag{9}$$

Proof. We can formally calculate

$$\begin{aligned} &\frac{d}{dt}\int_\Omega P_\epsilon(x,t)V_\epsilon(x,t)dx\\ &\qquad=\int_\Omega\left(\frac{\partial}{\partial t}P_\epsilon(x,t)V_\epsilon(x,t)+P_\epsilon(x,t)\frac{\partial}{\partial t}V_\epsilon(x,t)\right)dx\\ &\qquad=\int_\Omega(-\Delta P_\epsilon(x,t)V_\epsilon(x,t)+P_\epsilon(x,t)\Delta V_\epsilon(x,t))dx\\ &\qquad\quad-\int_\Omega f_Y(Z_\epsilon(x,t),Y_\epsilon(x,t),u(t))^T Q_\epsilon(x,t)V_\epsilon(x,t)dx\\ &\qquad=\int_{\partial\Omega}\left(\frac{\partial}{\partial N(y)}V_\epsilon(y,t)P_\epsilon(y,t)-V_\epsilon(y,t)\frac{\partial}{\partial N(y)}P_\epsilon(y,t)\right)dy\\ &\qquad\quad-\int_\Omega f_Y(Z_\epsilon(x,t),Y_\epsilon(x,t),u(t))^T Q_\epsilon(x,t)V_\epsilon(x,t)dx \end{aligned}$$

$$
\begin{aligned}
&= \int_\Omega (U_\epsilon(y,t)P_\epsilon(y,t) - V_\epsilon(y,t)R_\epsilon(y,t))dy \\
&\quad - \int_\Omega f_Y(Z_\epsilon(x,t), Y_\epsilon(x,t), u(t))^T Q_\epsilon(x,t)V_\epsilon(x,t)dx
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{d}{dt}\int_{\partial\Omega} \epsilon(y)R_\epsilon(y,t)U_\epsilon(y,t)dy \\
&\quad = \int_{\partial\Omega} \epsilon(y)\left(\frac{\partial}{\partial t}R_\epsilon(y,t)U_\epsilon(y,t) + R_\epsilon(y,t)\frac{\partial}{\partial t}U_\epsilon(y,t)\right)dy \\
&\quad = \int_{\partial\Omega} (R_\epsilon(y,t)V_\epsilon(y,t) - P_\epsilon(y,t)U_\epsilon(y,t))dy \\
&\qquad + \int_{\partial\Omega} R_\epsilon(y,t)\int_\Omega K(x,y,u(t))^T W_\epsilon(x,t)dxdy
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{d}{dt}\int_\Omega Q_\epsilon(x,t)^T W_\epsilon(x,t)dx \\
&\quad = \int_\Omega \left(\frac{\partial}{\partial t}Q_\epsilon(x,t)^T W_\epsilon(x,t) + Q_\epsilon(x,t)^T\frac{\partial}{\partial t}W_\epsilon(x,t)\right)dx \\
&\quad = -\int_\Omega\int_{\partial\Omega} K(x,y,u(t))^T R_\epsilon(y,t)dyW_\epsilon(x,t)dx \\
&\qquad + \int_\Omega Q_\epsilon(x,t)^T f_Y(Z_\epsilon(x,t), Y_\epsilon(x,t), u(t))V_\epsilon(x,t)dx.
\end{aligned}
$$

For the present this calculation is justified for smooth control functions and data (cf. [7]). We conclude that for times $t \in (s,T]$

$$
\frac{d}{dt}\left(\int_\Omega (P_\epsilon(x,t)V_\epsilon(x,t) + Q_\epsilon(x,t)^T W_\epsilon(x,t))dx + \int_{\partial\Omega} \epsilon(y)R_\epsilon(y,t)U_\epsilon(y,t)dy\right) = 0
$$

holds, from which it follows that

$$
\begin{aligned}
&\int_\Omega (P_\epsilon(x,t)V_\epsilon(x,t) + Q_\epsilon(x,t)^T W_\epsilon(x,t))dx + \int_{\partial\Omega} \epsilon(y)R_\epsilon(y,t)U_\epsilon(y,t)dy \\
&\quad = \int_\Omega (P_\epsilon(x,s)V_\epsilon(x,s) + Q_\epsilon(x,s)^T W_\epsilon(x,s))dx + \int_{\partial\Omega} \epsilon(y)R_\epsilon(y,s)U_\epsilon(y,s)dy
\end{aligned}
$$

is constant for all $t \in [s,T]$. Since we can approximate any admissible control function with respect to the Ekeland metric by continuous control functions, and the terminal condition Λ and initial condition Z^0 by smooth functions, we conclude that the expression above is constant and the proof of the lemma is finished. □

5. Proof of Theorem 2.2.

The proof of the Pontryagin maximum principle for the singularly perturbed system is standard and mainly follows the unifying concept of [5], with the only differences being that by introducing the adjoint variables our maximum principle has the familiar form and that the (smooth) target set gives more information on the terminal adjoint state.

Let u^* be the optimal control for the singularly perturbed system (2), (3) and $T>0$ the arrival time. We take a sequence of times $t_n \in [0,T)$ with $t_n \to T$ and define on $L^\infty((0,T);U)$ the (positive) functional

$$F_n(u) := |L(Z_\epsilon(\cdot,t_n)(u))|.$$

We define $\delta_n := F_n(u^*) > 0$. By Ekeland's variational principle (cf. [4]), there is a control u_n with

$$d(u_n,u^*) \le \sqrt{\delta_n}, \qquad F_n(u) \ge F_n(u_n) - \sqrt{\delta_n} d(u,u_n).$$

The second inequality holds especially for the strong variation u_n^h of the control function u_n; see (6). Then the inequality can be written as

$$\frac{|L(Z_\epsilon(\cdot,t_n)(u_n^h))| - |L(Z_\epsilon(\cdot,t_n)(u_n))|}{h} \ge -\sqrt{\delta_n}\frac{d(u_n^h,u_n)}{h}.$$

Taking the limit $h\to 0+$ we obtain

$$\frac{L(Z_\epsilon(\cdot,t_n)(u_n))}{|L(Z_\epsilon(\cdot,t_n)(u_n))|}\langle \nabla L(Z_\epsilon(\cdot,t_n)(u_n), W_\epsilon(\cdot,t_n)(u_n)\rangle \ge -\sqrt{\delta_n}, \tag{10}$$

where $W_\epsilon(u_n)$ is the strong variation of the trajectory $Z_\epsilon(u_n)$; see (6). The inequality above holds for all $v\in U$ and all Lebesgue points $s\in[0,T]$ of u_n, thus for all $s\in A(n)\subset[0,T]$, where $A(n)$ is a set of full Lebesgue measure in $[0,T]$. Taking a subsequence, we can assume that $d(u_n,u^*)\le 2^{-n}$. Then we can take the limit $n\to\infty$ in (10) and get

$$\operatorname{sgn}(L(Z^0))\langle \nabla L(Z_\epsilon(\cdot,T)(u^*)), W_\epsilon(\cdot,T)(u^*)\rangle \ge 0, \tag{11}$$

where $W_\epsilon(u^*)$ is the strong variation of the trajectory $Z_\epsilon(u^*)$. Now we define the terminal condition of the adjoint system as

$$\Lambda = \operatorname{sgn}(L(Z^0))\nabla L(Z_\epsilon(\cdot,T)(u^*)).$$

Then the maximum condition (11) can be written as

$$\operatorname{sgn}(L(Z^0))\int_\Omega Q_\epsilon(x,T)(u^*)^T W_\epsilon(x,T)(u^*)dx \ge 0,$$

and using Lemma 4.2 and the initial conditions at time $t=s$ of the variational system (8) we finish the proof. □

6. Proof of Theorem 2.1. The proof of Theorem 2.1 is based on an approximation of the system (1) by the singularly perturbed system (2), (3). For this reason we need a stability statement for small perturbations $\epsilon(y)=\epsilon_n>0$ *and* controls that are close to a given nominal control.

The following lemma is a key tool. It states that, if the perturbation of the control is adapted to the singular perturbation of the system, the trajectories of the singularly perturbed system converge to trajectories of the (unperturbed) system.

LEMMA 6.1. *For any sequence $\epsilon_n\to 0$ and $u_n\in B_{\epsilon_n^2}(u)$ we have the convergence*

$$\begin{cases} Z_{\epsilon_n}(u_n)\to Z_0(u) & \text{in} \quad C([0,T_0];L^\infty(\Omega;R^n)),\\ Y_{\epsilon_n}(u_n)\to Y_0(u) & \text{in} \quad C([0,T_0];C(\overline{\Omega};R))\cap L^2((0,T);H^1(\Omega)),\end{cases} \tag{12}$$

$$(13) \qquad \begin{cases} Q_{\epsilon_n}(u_n) \to Q_0(u) & \text{in} \quad C([0,T_0];L^\infty(\Omega;R^n)), \\ R_{\epsilon_n}(u_n) \to P_0(u) & \text{in} \quad C([0,T_0];C(\partial\Omega;R)). \end{cases}$$

Proof. We will show (12). We consider the (ϵ_n-dependent) solution operator $S^u_{(X,Z)}$, which maps trajectories Y to solutions (X,Z) of (3). For the fixed control function u we can write

$$\begin{cases} Z(x,t) &= Z^0(x) + \int_0^t f(Z(x,r),Y(x,r),u(r))dr, \\ X(y,t) &= \int_0^t \frac{e^{-(t-r)/\epsilon_n}}{\epsilon_n}(\int_\Omega K(x,y,u(r))Z(x,r)dx + Y(y,r))dr. \end{cases}$$

We note that $r \mapsto \frac{e^{-(t-r)/\epsilon_n}}{\epsilon_n}$ serves as mollifier, since, as $\epsilon_n \to 0$, we have

$$\int_0^t \frac{e^{-(t-r)/\epsilon_n}}{\epsilon_n}dr = 1 - e^{-t/\epsilon_n} \to 1, \qquad \int_{t-\sqrt{\epsilon_n}}^t \frac{e^{-(t-r)/\epsilon_n}}{\epsilon_n}dr = 1 - e^{-1/\sqrt{\epsilon_n}} \to 1.$$

Thus, as $\epsilon_n \to 0$, we have the convergence in $L^1((0,T_0);C(\partial\Omega;R))$

$$\left(t \mapsto \int_0^t \frac{e^{-(t-r)/\epsilon_n}}{\epsilon_n}\left(\int_\Omega K(x,y,u(r))Z(x,r)dx + Y(y,r)\right)dr\right)$$
$$\to \left(t \mapsto \int_\Omega K(x,y,u_n(t))Z(x,t)dx + Y(y,t)\right).$$

We now take into account variations of the control function. First, we note that, as $\epsilon_n \to 0$,

$$\int_{t-{\epsilon_n}^2}^t \frac{e^{-(t-r)/\epsilon_n}}{\epsilon_n}dr = 1 - e^{-\sqrt{\epsilon_n}} \to 0.$$

Thus, if we take $u_n \in B_{\epsilon_n^2}(u)$, we still have

$$\left|\int_0^t \frac{e^{-(t-r)/\epsilon_n}}{\epsilon_n}\int_\Omega (K(x,y,u(r))) - K(x,y,u(r))Z(x,r)dxdr\right| = O(\sqrt{\epsilon_n}).$$

Hence, as $\epsilon_n \to 0$, we get the convergence in $L^1((0,T_0);C(\partial\Omega;R))$

$$\left(t \mapsto \int_0^t \frac{e^{-(t-r)/\epsilon_n}}{\epsilon_n}\left(\int_\Omega K(x,y,u_n(r))Z(x,r)dx + Y(y,r)\right)dr\right)$$
$$\to \left(t \mapsto \int_\Omega K(x,y,u(t))Z(x,t)dx + Y(y,t)\right).$$

Note that both the converging sequence and the limit are (uniformly bounded) in $L^\infty((0,T_0);C(\partial\Omega;R))$. With respect to the natural topology of $L^\infty((0,T_0);C(\partial\Omega;R))$ for X the solution operator S_Y, which maps trajectories (X,Z) to solutions Y of (2), is continuous. Also with respect to the Ekeland topology of $L^\infty((0,T_0);C(\partial\Omega;R))$ for X we get continuity of S_Y (see Theorem 5.1 in [3]). Both continuity properties imply continuity of S_Y with respect to the $L^1((0,T_0);C(\partial\Omega;R))$ topology for X! Just use the fact that a sequence X_n in $L^\infty((0,T_0);C(\partial\Omega;R))$, which converges in $L^1((0,T_0);C(\partial\Omega;R))$, necessarily uniformly approximates the limit on a sequence of sets $A_n \subset (0,T_0)$ with increasing Lebesgue measure $\lambda(A_n) \to T_0$. All in all, we conclude that the fixed point of the contraction $S_Y \circ S^{u_n}_{(X,Z)}$ tends in $L^1([0,T];C(\partial\Omega)) \times$

$C([0,T]; L^\infty(\Omega; R^n))$ to $t \mapsto (\int_\Omega K(x,y,u)Z_0(x,t)(u)dx + Y_0(y,t)(u), Z_0(x,t)(u))$, and similarly the fixed point of the contraction $S^{u_n}_{(X,Z)} \circ S_Y$ tends to $Y_0(u)$. The proof of (13) is completely analogous. We just have to take into account that the equation for R in the singularly perturbed adjoint system is stable in backwards time and the same singular perturbation technique works again. The proof of the lemma is finished. □

Let u_0^* be the optimal control for the system (1) and $T_0 > 0$ the arrival time. We take sequences of times $t_n \in [0, T_0)$ and of perturbation parameters $\epsilon_n > 0$, with $t_n \to T_0$ and $\epsilon_n \to 0$, as $n \to \infty$. We define the functional

$$F_n(u) := |L(Z_{\epsilon_n}(\cdot, t_n)(u))|$$

obtained with the trajectory $Z_{\epsilon_n}(u)$ of the singularly perturbed system (2), (3) with $\epsilon(y) := \epsilon_n > 0$ for all $y \in \partial\Omega$.

Lemma 6.2. *For every sequence $t_n \to T_0$, $t_n \in [0, T_0)$, there is a sequence $\epsilon_n \to 0$, $\epsilon_n > 0$, and a sequence of radii $\mu_n \to 0$, $\mu_n > 0$, such that, for $n \in N$ large enough, the functional F_n is positive on the ball $B_{\mu_n}(u_0^*)$. $\delta_n := F_n(u_0^*)$ especially is positive, for $n \in N$ large enough. Furthermore $\delta_n := F_n(u_0^*) \to 0$, as $n \to \infty$.*

Proof. According to Lemma 6.1, applied to $u = u_0^*$, we have

$$Z_{\epsilon_n}(u_n) \to Z_0(u_0^*)$$

if $u_n \in B_{\epsilon_n^2}(u_0^*)$. It follows that

$$L(Z_{\epsilon_n}(\cdot, t_n)(u_n)) \to L(Z_0(\cdot, t_n)(u_0^*))$$

uniformly in $t_n \in [0, T_0]$, if $u_n \in B_{\epsilon_n^2}(u_0^*)$. By optimality of u_0^* we have

$$|L(Z_0(\cdot, t_n)(u_0^*))| > 0$$

for every $t_n \in [0, T_0)$. Thus, if we choose $0 < \mu_n \le \epsilon_n^2$ small enough, the functional F_n is positive on $B_{\mu_n}(u_0^*)$. It also follows that $\delta_n \to 0$ and the proof of the lemma is finished. □

Since $\delta_n = F_n(u_0^*)$ is positive, for $n \in N$ large enough, we get, as in the proof of Theorem 2.2 for a $u_n \in B_{\mu_n}(u_0^*) \cap B_{\sqrt{\delta_n}}(u_0^*)$,

$$\frac{L(Z_{\epsilon_n}(\cdot, t_n)(u_n))}{|L(Z_{\epsilon_n}(\cdot, t_n)(u_n))|} \langle \nabla L(Z_{\epsilon_n}(\cdot, t_n)(u_n)), W_{\epsilon_n}(\cdot, t_n)(u_n) \rangle \ge -\sqrt{\delta_n},$$

where $W_{\epsilon_n}(u_n)$ is the strong variation of the trajectory $Z_{\epsilon_n}(u_n)$; see (7). We conclude by Lemma 4.2 and by the initial conditions of (8) that

$$\begin{aligned} -\sqrt{\delta_n} \le {} & \int_\Omega Q_{\epsilon_n}(x,s)\,(f(Z_{\epsilon_n}(x,s), Y_{\epsilon_n}(x,s), v) - f(Z_{\epsilon_n}(x,s), Y_{\epsilon_n}(x,s), u_n(s)))\,dx \\ & + \int_{\partial\Omega} R_{\epsilon_n}(y,s) \int_\Omega (K(x,y,v) - K(x,y,u_n(s)))Z_{\epsilon_n}(x,s)dy, \end{aligned} \tag{14}$$

where $Q_{\epsilon_n} = Q_{\epsilon_n}(u_n)$ and $R_{\epsilon_n} = R_{\epsilon_n}(u_n)$ are solutions of the adjoint system

$$(15) \quad \left\{ \begin{array}{rcl} \frac{\partial}{\partial t}P(x,t) & = & -\Delta P(x,t) - f_Y(Z_{\epsilon_n}(x,t), Y_{\epsilon_n}(x,t), u_n(t))^T Q(x,t), \\ \frac{\partial}{\partial N(y)}P(y,t) & = & R(y,t) - P(y,t), \\ \epsilon(y)\frac{\partial}{\partial t}R(y,t) & = & R(y,t) - P(y,t), \\ \frac{\partial}{\partial t}Q(x,t) & = & -f_Z(Z_{\epsilon_n}(x,t), Y_{\epsilon_n}(x,t), u(t))^T Q(x,t) \\ & & - \int_{\partial\Omega} K(x,y,u_n(t))R(y,t)dy, \\ P(x,t_n) & = & 0, \\ R(y,t_n) & = & 0, \\ Q(x,t_n) & = & \mathrm{sgn}(L(Z^0))\nabla L(Z_{\epsilon_n}(\cdot,t_n)(u_n)) \end{array} \right.$$

obtained with terminal conditions at time $t = t_n$, with the control u_n and with $\epsilon(y) = \epsilon_n$. Since, as a slight adaptation of the proof of Lemma 6.1 shows, we have the convergence of the trajectories of (15) to the trajectories of (4), obtained with terminal conditions at time $t = T_0$ and with the control u_0^*, thus the maximum principle follows from (14) and the proof is finished. □

REFERENCES

[1] V. Arnautu, V. Barbu, and V. Capasso, *Controlling the spread of a class of epidemics*, Appl. Math. Optim., 20 (1989), pp. 297–317.

[2] V. Capasso and K. Kunisch, *A reaction–diffusion system arising in modelling man–environment diseases*, Quart. Appl. Math., 46 (1988), pp. 431–450.

[3] E. Casas, *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations*, SIAM J. Control Optim., 35 (1997), pp. 1297–1327.

[4] I. Ekeland, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.

[5] H. O. Fattorini, *A unified theory of necessary conditions for nonlinear nonconvex control systems*, Appl. Math. Optim., 15 (1987), pp. 141–185.

[6] H. O. Fattorini and T. Murphy, *Optimal problems for nonlinear parabolic boundary control systems*, SIAM J. Control Optim., 32 (1994), pp. 1577–1596.

[7] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Inc., Englewood Cliffs, NJ, 1964.

[8] V. G. Gaitsgory, *Suboptimization of singularly perturbed control systems*, SIAM J. Control Optim., 30 (1992), pp. 1228–1249.

[9] G. Grammel, *Averaging of singularly perturbed systems*, Nonlinear Anal., 28 (1997), pp. 1851–1865.

[10] B. Hu and J. Yong, *Pontryagin maximum principle for semilinear and quasilinear parabolic equations with pointwise state constraints*, SIAM J. Control Optim., 33 (1995), pp. 1857–1880.

[11] O. A. Ladyzhenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, AMS, Providence, RI, 1968.

[12] A. N. Tychonov, *Systems of differential equations containing small parameters in the derivatives*, Mat. Sb., 31 (1952), pp. 575–586.

[13] A. Vigodner, *Limits of singularly perturbed control problems with statistical dynamics of fast motions*, SIAM J. Control Optim., 35 (1997), pp. 1–28.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 4, pp. 1176–1194

PRIMAL-DUAL STRATEGY FOR CONSTRAINED OPTIMAL CONTROL PROBLEMS*

MAÏTINE BERGOUNIOUX[†], KAZUFUMI ITO[‡], AND KARL KUNISCH[§]

Abstract. An algorithm for efficient solution of control constrained optimal control problems is proposed and analyzed. It is based on an active set strategy involving primal as well as dual variables. For discretized problems sufficient conditions for convergence in finitely many iterations are given. Numerical examples are given and the role of the strict complementarity condition is discussed.

Key words. active set, augmented Lagrangian, primal-dual method, optimal control

AMS subject classifications. 49J20, 49M29

PII. S0363012997328609

1. Introduction and formulation of the problem.

In the recent past significant advances have been made in efficiently solving nonlinear optimal control problems. Most of the proposed methods are based on variations of the sequential quadratic programming (SQP) technique; see for instance [HT, KeS, KuS, K, T] and the references therein. The SQP-algorithm is sequential and each of its iterations requires the solution of a quadratic minimization problem subject to linearized constraints. If these auxiliary problems contain inequality constraints with infinite dimensional image space, then their solution is still a significant challenge.

In this paper we propose an algorithm for the solution of infinite dimensional quadratic problems with linear equality constraints and pointwise affine inequality constraints. It is based on an active set strategy involving primal and dual variables. It thus differs significantly from conventional active set strategies that involve primal variables only; see [Sch] for example. In practice the proposed algorithm behaves like an infeasible one. The iterates of the algorithm violate the constraints up to the next-to-last iterate. The algorithm stops at a feasible and optimal solution.

Within this paper we do not aim for generality but rather we treat as a model problem a unilateral control constraint optimal control problem related to elliptic partial differential equations. The distributed nature of this problem, which is reflected in the fact that it behaves like an obstacle problem for the biharmonic equation, makes it difficult to analyze.

Let us briefly outline the contents of the paper. The algorithm will be presented in section 2. We prove that if the algorithm produces the same active set in two consecutive iterates, then the optimal solution has been obtained. In section 3 we shall give sufficient conditions which guarantee that an augmented Lagrangian functional behaves as a decreasing merit function for the algorithm. In practice this implies finite

*Received by the editors October 13, 1997; accepted for publication (in revised form) September 29, 1998; published electronically May 21, 1999.
http://www.siam.org/journals/sicon/37-4/32860.html

[†]UMR-CNRS 6628, Université d'Orleans, U.F.R. Sciences, B.P. 6759, F-45067 Orléans Cedex 2, France (maitine@labomath.univ-orleans.fr). The research of this author was supported in part by EEC, HCM contract CHRX-CT94-0471.

[‡]Department of Mathematics, North Carolina State University, Raleigh, NC 27695 (Kito@eos.ncsu.edu).

[§]Institut für Mathematik, Universität Graz, A-8010 Graz, Austria (Kunisch@kfunigraz.ac.at). The research of this author was supported in part by EEC, HCM contract CHRX-CT94-0471 and Fonds zur Förderung der wissenschaftlichen Forschung, UF8, Optimization and Control.

step convergence of the discretized problem. Section 4 is devoted to showing that for a minor modification of the algorithm the cost functional is increasing until the feasible optimal solution is reached. In section 5 several numerical examples are given. For most examples the algorithm behaves extremely efficiently and typically converges in less than five iterations. Thus, to present interesting cases the majority of the test examples is in some sense extreme: either the strict complementarity condition is violated or the cost of the control is nearly zero.

To describe the problem, let Ω be an open, bounded subset of $\mathbb{R}^N$, $N \leq 3$, with smooth boundary Γ and consider the following distributed optimal control problem:

$$(\mathcal{P}) \qquad \min \quad J(y,u) = \frac{1}{2}\int_\Omega (y - z_d)^2\, dx + \frac{\alpha}{2}\int_\Omega (u - u_d)^2\, dx,$$

$$(1.1) \qquad -\Delta y = u \text{ in } \Omega, \quad y = 0 \quad \text{on } \Gamma,$$

$$(1.2) \qquad u \in U_{ad} \subset L^2(\Omega),$$

where $z_d,\ u_d \in L^2(\Omega),\ \alpha > 0$, and $U_{ad} = \{\ u \in L^2(\Omega) \mid u(x) \leq b(x) \text{ a.e. in } \Omega\},\ b \in L^\infty(\Omega)$.

It is well known that for every $u \in L^2(\Omega)$ system (1.1) has a unique solution $y = \mathcal{T}(u)$ in $H^2(\Omega) \cap H^1_o(\Omega)$.

REMARK 1.1. *To emphasis the basic ideas of the proposed approach we treated the rather simple problem* $(\mathcal{P})$. *Many generalizations are possible. In particular, the analysis of this paper can be extended to the case where* $-\Delta$ *in* (1.1) *is replaced by any strictly elliptic second order differential operator. The algorithm itself can be easily adapted to other optimal control problems involving, for example, ordinary differential equations. Its numerical efficiency as well as the convergence analysis require some additional research.*

It is standard that problem $(\mathcal{P})$ has a unique solution (y^*, u^*) characterized by the following optimality system:

$$\begin{cases} -\Delta y^* = u^* \text{ in } \Omega, \quad y^* \in H^1_o(\Omega), \\ -\Delta p^* = z_d - y^* \text{ in } \Omega, \quad p^* \in H^1_o(\Omega), \\ (\alpha(u^* - u_d) - p^*, u - u^*) \geq 0 \qquad \text{for all } u \in U_{ad}, \end{cases}$$

where $(\cdot,\cdot)$ denotes the $L^2(\Omega)$-inner product.

Let us give an equivalent formulation for this optimality system, which is essential to motivate the forthcoming algorithm.

THEOREM 1.1. *The unique solution* (y^*, u^*) *to problem* $(\mathcal{P})$ *is characterized by*

$$(\mathcal{S}) \qquad \begin{cases} -\Delta y^* = u^* \ \text{in } \Omega, \quad y^* \in H^1_o(\Omega), \\ -\Delta p^* = z_d - y^* \ \text{in } \Omega, \quad p^* \in H^1_o(\Omega), \\ u^* = u_d + \dfrac{p^* - \lambda^*}{\alpha}, \\ \lambda^* = c\left[u^* + \dfrac{\lambda^*}{c} - \Pi\left(u^* + \dfrac{\lambda^*}{c}\right)\right] = c \max\left(0, u^* + \dfrac{\lambda^*}{c} - b\right) \end{cases}$$

for every $c > 0$. *Here* Π *denotes the projection of* $L^2(\Omega)$ *onto* U_{ad}.

Proof. We refer to [IK]. □

We point out that the last equation in $(\mathcal{S})$,

$$\lambda^* = c\left[u^* + \frac{\lambda^*}{c} - \Pi\left(u^* + \frac{\lambda^*}{c}\right)\right], \tag{1.3}$$

is equivalent to

$$\lambda^* \in \partial I_{U_{ad}}(u^*), \tag{1.4}$$

where ∂I_C denotes the subdifferential of the indicator function I_C of a convex set C. This follows from general properties of convex functions (see [IK] for example) and can also easily be verified directly for the convex function $I_{U_{ad}}$. The replacement of the well-known differential inclusion (1.4) [B] in the optimality system for $(\mathcal{P})$ by (1.3) is an essential ingredient of the algorithm that we shall propose.

Here and below, order relations like "max" and "$\leq$" between elements of $L^2(\Omega)$ are understood in the pointwise almost everywhere (a.e.) sense.

Let us interpret the optimality system $(\mathcal{S})$. From $-\Delta y^* = u_d + \frac{p^*-\lambda^*}{\alpha}$ it follows that $p^* = \alpha[-\Delta y^* - u_d] + \lambda^*$ and hence

$$-\alpha\Delta y^* - \Delta^{-1}y^* + \lambda^* = \alpha\ u_d - \Delta^{-1}z_d.$$

It follows that

$$\begin{aligned} \alpha u^* + \Delta^{-2}u^* + \lambda^* &= \alpha u_d - \Delta^{-1}z_d\ , \\ \lambda^* &= c\max\left(0, u^* + \frac{\lambda^*}{c} - b\right) \qquad \text{for all } c > 0, \end{aligned}$$

which implies the highly distributed nature of the optimal control. Setting $\mathcal{H} = \alpha I + \Delta^{-2}$ and $f = \alpha u_d - \Delta^{-1}z_d$, system $(\mathcal{S})$ can be expressed as

$$(\mathcal{S}_1) \quad \left\{ \begin{aligned} \mathcal{H}u^* + \lambda^* &= f, \\ \lambda^* &= c\max\left(0, u^* + \frac{\lambda^*}{c} - b\right) \qquad \text{for all } c > 0. \end{aligned} \right.$$

We observe that by setting $u = -\Delta y$, system $(\mathcal{S})$ constitutes an optimality system for the variational inequality,

$$\left\{ \begin{array}{l} \min \dfrac{\alpha}{2}\displaystyle\int_\Omega |\Delta y|^2 dx + \dfrac{1}{2}\displaystyle\int_\Omega |y - (z_d - \alpha\ \Delta u_d)|^2 dx, \\ y \in H_o^1(\Omega) \cap H^2(\Omega), \\ -\Delta y \leq b, \end{array} \right.$$

the regularity of which was studied in [BS].

2. Presentation of the algorithm. In this section we present the primal-dual active set algorithm and discuss some of its basic properties. Let us introduce the active and inactive sets for the solution to $(\mathcal{P})$ and define

$$\mathcal{A}^* = \{\ x \mid u^*(x) = b \ \text{ a.e. }\} \text{ and } \mathcal{I}^* = \{\ x \mid u^*(x) < b \ \text{ a.e. }\}.$$

The proposed strategy is based on (1.3). Given (u_{n-1}, λ_{n-1}) the active set for the current iterate is chosen as

$$\mathcal{A}_n = \left\{\ x \mid u_{n-1}(x) + \frac{\lambda_{n-1}(x)}{c} > b \ \text{ a.e. } \right\}.$$

We recall that $\lambda^* \geq 0$ and in the case of strict complementarity $\lambda^* > 0$ on $\mathcal{A}^*$. The complete algorithm is specified next.

Algorithm.

1. Initialization: choose $y_o,\ u_o$, and λ_o and set $n = 1$.
2. Determine the following subsets of Ω:

$$\mathcal{A}_n = \left\{\ x \mid u_{n-1}(x) + \frac{\lambda_{n-1}(x)}{c} > b\ \right\},\ \mathcal{I}_n = \left\{\ x \mid u_{n-1}(x) + \frac{\lambda_{n-1}(x)}{c} \le b\ \right\}.$$

3. If $n \ge 2$ and $\mathcal{A}_n = \mathcal{A}_{n-1}$, then STOP.
4. Else, find $(y_n, p_n) \in H^1_o(\Omega) \times H^1_o(\Omega)$ such that

$$-\Delta y_n = \begin{cases} b & \text{in } \mathcal{A}_n, \\ u_d + \dfrac{p_n}{\alpha} & \text{in } \mathcal{I}_n, \end{cases}$$
$$-\Delta p_n = z_d - y_n \quad \text{in } \Omega$$

and set

$$u_n = \begin{cases} b & \text{in } \mathcal{A}_n, \\ u_d + \dfrac{p_n}{\alpha} & \text{in } \mathcal{I}_n. \end{cases}$$

5. Set $\lambda_n = p_n - \alpha(u_n - u_d)$, update $n = n + 1$, and goto 2.

The existence of the triple (y_n, u_n, p_n) satisfying the system of step 4 of the algorithm follows from the fact that it constitutes the optimality system for the auxiliary problem

$$(\mathcal{P}_{aux}) \qquad \min\ \{\ J(y,u) \mid y \in H^1_o(\Omega),\ -\Delta y = u \text{ in } \Omega,\ u = b \text{ on } \mathcal{A}_n\ \},$$

which has (y_n, u_n) as unique solution.

We may use different initialization schemes. The one that was used most frequently is the following:

$$\text{(2.1)} \qquad \begin{cases} u_o = b, \\ -\Delta y_o = u_o,\ y_o \in H^1_o(\Omega), \\ -\Delta p_o = z_d - y_o,\ p_o \in H^1_o(\Omega), \\ \lambda_o = \max(0, \alpha(u_d - b) + p_o). \end{cases}$$

This choice of initialization has the property of feasibility. Alternatively, we tested the algorithm with initialization as the solution of the unconstrained problem, i.e.,

$$\text{(2.2)} \qquad \begin{cases} \lambda_o = 0, \\ -\Delta y_o = u_d + \dfrac{p_o}{\alpha}, & y_o \in H^1_o(\Omega), \\ -\Delta p_o = z_d - y_o, & p_o \in H^1_o(\Omega), \\ u_o = u_d + \dfrac{p_o}{\alpha}. \end{cases}$$

For all examples the first initialization behaved better than or equal to the second.

The initialization process (2.1) has the property that the first set $\mathcal{A}_1$ is always included in the active set $\mathcal{A}^*$ of problem $(\mathcal{P})$. More precisely we have the following lemma.

Lemma 2.1. *If (u_o, y_o, λ_o) are given by (2.1) with $u_o \ge u^*$, then $\lambda_o \le \lambda^*$. In addition, if $u_o = b$, then $\mathcal{A}_1 \subset \mathcal{A}^*$.*

Proof. The proof is the construction

$$\lambda_o = \max(0, \alpha(u_d - u_o) + p_o) = \max(0, \alpha(u_d - u_o) + \Delta^{-1}(y_o - z_d))$$

and as a consequence of $(\mathcal{S})$,

$$\lambda^* = \alpha(u_d - u^*) + p^* = \alpha(u_d - u^*) + \Delta^{-1}(y^* - z_d) = \alpha(u_d - u^*) - \Delta^{-2}u^* - \Delta^{-1}z_d \ \geq 0.$$

It follows that $\lambda^* - \lambda_o = \lambda^* \geq 0$ if $\alpha(u_d - u_o) + \Delta^{-1}(y_o - z_d) \leq 0$, and $\lambda^* - \lambda_o = \alpha(u_o - u^*) + \Delta^{-2}(u_o - u^*) + \alpha(u_d - u_o) + \Delta^{-1}(y_o - z_d)$ else.

If $u_o \geq u^*$, the maximum principle yields $\Delta^{-2}(u_o - u^*) \geq 0$ and

$$\lambda^* - \lambda_o \begin{cases} = \ \lambda^* \geq 0 & \text{if } \alpha(u_d - u_o) + \Delta^{-1}(y_o - z_d) \leq 0, \\ \geq \alpha(u_d - u_o) + \Delta^{-1}(y_o - z_d) \geq 0 & \text{else.} \end{cases}$$

Therefore $\lambda_o \leq \lambda^*$.

In addition, if $u_o = b$, then $u_o + \frac{\lambda_o}{c} = b + \frac{\lambda_o}{c} > b$ on $\mathcal{A}_1$. Consequently $\lambda_o > 0$ on $\mathcal{A}_1$ and $\lambda^* > 0$. It follows that $\mathcal{A}_1 \subset \mathcal{A}^*$ and the proof is complete. □

A first convergence result which also justifies the stopping criterion in step 3 is given in the following theorem.

Theorem 2.1. *If there exists $n \in \mathbb{N} - \{0\}$ such that $\mathcal{A}_n = \mathcal{A}_{n+1}$, then the algorithm stops and the last iterate satisfies*

$$(\mathcal{S}_n) \qquad \begin{cases} -\Delta y_n = u_n = \begin{cases} b & \textit{in } \mathcal{A}_n, \\ u_d + \dfrac{p_n}{\alpha} & \textit{in } \Omega - \mathcal{A}_n, \end{cases} \\ -\Delta p_n = z_d - y_n \quad \textit{in } \Omega, \\ \lambda_n = p_n - \alpha(u_n - u_d), \ u_n \in U_{ad}, \end{cases}$$

with

$$\lambda_n = 0 \ \textit{on } \mathcal{I}_n \qquad \textit{and} \qquad \lambda_n > 0 \ \textit{on } \mathcal{A}_n. \tag{2.3}$$

Therefore, the last iterate is the solution of the original optimality system $(\mathcal{S})$.

Proof. If there exists $n \in \mathbb{N} - \{0\}$ such that $\mathcal{A}_n = \mathcal{A}_{n+1}$, then it is clear that the algorithm stops and the last iterate satisfies $(\mathcal{S}_n)$ by construction, except possibly for $u_n \in U_{ad}$. Thus we have to prove $u_n \in U_{ad}$ and (2.3).

- On $\mathcal{I}_n$ we have $\lambda_n = 0$ by step 5 of the algorithm. Moreover, $u_n + \frac{\lambda_n}{c} = u_n \leq b$, since $\mathcal{I}_n = \mathcal{I}_{n+1}$.
- On $\mathcal{A}_n$ we get $u_n = b$ and $u_n + \frac{\lambda_n}{c} > b$ since $\mathcal{A}_n = \mathcal{A}_{n+1}$. Therefore $\lambda_n > 0$ on $\mathcal{A}_n$ and $u_n \in U_{ad}$.

To prove that the last iterate is a solution of the original optimality system $(\mathcal{S})$, it remains to show that

$$\lambda_n = c\left[u_n + \frac{\lambda_n}{c} - \Pi\left(u_n + \frac{\lambda_n}{c}\right)\right].$$

- On $\mathcal{I}_n$ we have $\lambda_n = 0$ and $u_n + \frac{\lambda_n}{c} = u_n \leq b$. It follows that

$$u_n + \frac{\lambda_n}{c} - \Pi\left(u_n + \frac{\lambda_n}{c}\right) = u_n - \Pi(u_n) = 0 = \lambda_n.$$

- On $\mathcal{A}_n$ we get $u_n = b$, $\lambda_n > 0$ and therefore

$$c\left[u_n + \frac{\lambda_n}{c} - \Pi\left(u_n + \frac{\lambda_n}{c}\right)\right] = c\left[b + \frac{\lambda_n}{c} - b\right] = \lambda_n. \qquad \square$$

Now we give a structural property of the algorithm.

LEMMA 2.2. *If u_n is feasible for some $n \in \mathbb{N} - \{0\}$ (i.e., $u_n \le b$), then $\mathcal{A}_{n+1} \subset \mathcal{A}_n$.*

Proof. On $\mathcal{I}_n$ we get $\lambda_n = 0$ by construction so that $u_n + \frac{\lambda_n}{c} = u_n \le b$ (because of feasibility). This implies $\mathcal{I}_n \subset \mathcal{I}_{n+1}$ and consequently $\mathcal{A}_{n+1} \subset \mathcal{A}_n$. □

Note that Theorem 2.1 and in particular (2.3) do not utilize or imply strict complementarity. In fact, if (2.3) holds, then the set of x for which $u_n(x) = b$ and $\lambda_n(x) = 0$ is contained in $\mathcal{I}_n$.

We end this section with simple cases, where we may conclude easily that the algorithm is convergent.

THEOREM 2.2. *For initialization* (2.1), *the algorithm converges in one iteration in the following cases:*

1. $z_d \le 0,\ u_d = 0,\ b \ge 0$ *and the solution to* $-\alpha\Delta u - \Delta^{-1}u = z_d$ *is nonpositive.*

2. $z_d \ge 0,\ b \le 0,\ u_d > b$ *or* $z_d \ge 0,\ b \le 0,\ u_d \ge b$ *and* $z_d + \Delta^{-1}b$ *is not zero as element in* $L^2(\Omega)$.

Proof. Let us first examine case 1. The maximum principle implies that $-\Delta^{-1}u_o \ge 0$. Consequently $z_d + \Delta^{-1}u_o \le 0$ and by a second application of the maximum principle

$$-\Delta^{-1}(z_d + \Delta^{-1}u_o) \le 0.$$

Together with the fact that $u_d - b = -b \le 0$, this implies

$$\lambda_o = \max(0, \alpha(u_d - b) - \Delta^{-1}(z_d + \Delta^{-1}u_o)) = 0.$$

Therefore $\mathcal{A}_1 = \emptyset$ and $\mathcal{I}_1 = \Omega$.

Using the first iteration we obtain $u_1 = \frac{p_1}{\alpha}$ in Ω. Moreover, $-\Delta y_1 = u_1$ and $-\Delta p_1 = z_d - y_1$ imply that

$$-\alpha\Delta u_1 - \Delta^{-1}u_1 = z_d.$$

By assumption u_1 is feasible. Therefore $\mathcal{A}_2 = \mathcal{A}_1 = \emptyset$ and by Theorem 2.1 the algorithm stops at the solution to $(\mathcal{P})$.

Now we consider case 2. By assumption and due to (2.1) we have $z_d \ge 0,\ b \le 0,\ \lambda_o \ge 0$, and $\mathcal{A}_1 = \{\ \lambda_o > 0\ \}$. Due to the maximum principle $-\Delta^{-1}u_o \le 0$ and

$$p_o = -\Delta^{-1}(z_d - y_o) = -\Delta^{-1}[z_d - (-\Delta^{-1}u_o)] \ge 0.$$

Moreover, if $z_d + \Delta^{-1}b$ is not the zero element in $L^2(\Omega)$, then $p_o > 0$ in Ω and $\alpha(u_d - b) + p_o > \alpha(u_d - b)$.

If $u_d > b$ or ($u_d = b$ and $z_d + \Delta^{-1}b \ne 0$), then $\lambda_o = \max(0, \alpha(u_d - b) + p_o) > 0$ in Ω (and $\lambda_o = 0$ on $\partial\Omega$). Consequently $\mathcal{A}_1 = \Omega$ and $u_1 = b$, $\lambda_1 = -\Delta^{-1}(z_d + \Delta^{-1}b) + \alpha(u_d - b) > 0$. This yields $\mathcal{A}_2 = \mathcal{A}_1 = \Omega$ and the algorithm stops. □

3. Convergence analysis.

3.1. The continuous case. The convergence analysis of the algorithm is based on the decrease of appropriately chosen merit functions. For that purpose we define the following augmented Lagrangian functions:

$$L_c(y, u, \lambda) = J(y, u) + (\lambda, \hat{g}_c(u, \lambda)) + \frac{c}{2}\|\hat{g}_c(u, \lambda)\|^2 \quad \text{and} \quad \hat{L}_c(y, u, \lambda) = L_c(y, u, \lambda^+),$$

where $(\cdot,\cdot)$ is the $L^2(\Omega)$-inner product, $\|\cdot\|$ is the $L^2(\Omega)$-norm, $\lambda^+ = \max(\lambda, 0)$, and $\hat{g}_c(u, \lambda) = \max(g(u), -\frac{\lambda}{c})$ with $g(u) = u - b$. Further $(\cdot,\cdot)_{|S}$ and $\|\cdot\|_{|S}$ denote the L^2-inner product and norm on a measurable subset $S \subset \Omega$. Note that the mapping

$$u \mapsto (\lambda, \hat{g}_c(u, \lambda)) + \frac{c}{2}\|\hat{g}_c(u, \lambda)\|^2$$

is $\mathcal{C}^1$, which is not the case for the function given by

$$u \mapsto (\lambda, g(u)) + \frac{c}{2}\|\max(g(u),0)\|^2.$$

The following relationship between primal and dual variables will be essential.

LEMMA 3.1. *For all $n \in \mathbb{N} - \{0\}$ and $(y,u) \in H^1_o(\Omega) \times L^2(\Omega)$ satisfying $-\Delta y = u$ we have*

$$J(y_n,u_n) - J(y,u) = -\frac{1}{2}\|y-y_n\|^2 - \frac{\alpha}{2}\|u-u_n\|^2 + (\lambda_n, u-u_n)_{|\mathcal{A}_n}\,. \tag{3.1}$$

Proof. Using $\|a\|^2 - \|b\|^2 = -\|a-b\|^2 + 2\,(a-b,a)$ and steps 4 and 5 of the algorithm, we find that

$$\begin{aligned}J(y_n,u_n)-J(y,u) &= -\frac{1}{2}\|y-y_n\|^2-\frac{\alpha}{2}\|u-u_n\|^2+(y_n-y,y_n-z_d)+\alpha\,(u_n-u,u_n-u_d)\\ &= -\frac{1}{2}\|y-y_n\|^2-\frac{\alpha}{2}\|u-u_n\|^2+(\Delta(y_n-y),p_n)+\alpha\,(u_n-u,u_n-u_d)\\ &= -\frac{1}{2}\|y-y_n\|^2-\frac{\alpha}{2}\|u-u_n\|^2+(u_n-u,-p_n+\alpha(u_n-u_d))\\ &= -\frac{1}{2}\|y-y_n\|^2-\frac{\alpha}{2}\|u-u_n\|^2+(u-u_n,\lambda_n)\,.\end{aligned}$$

Because $\lambda_n = 0$ on $\mathcal{I}_n$ the result follows. □

Let us define

$$\mathcal{S}_{n-1} = \{\ x \in \mathcal{A}_{n-1} \mid \lambda_{n-1}(x) \le 0\ \} \ \text{ and } \mathcal{T}_{n-1} = \{\ x \in \mathcal{I}_{n-1} \mid u_{n-1}(x) > b(x)\ \}.$$

These two sets can be paraphrased by calling $\mathcal{S}_{n-1}$ the set of elements that the active set strategy predicts to be active at level $n-1$ but the Lagrange multiplier indicates should be inactive and by calling $\mathcal{T}_{n-1}$ the set of elements that was predicted to be inactive but the $(n-1)$st iteration level corrects it to be active. We note that

$$\Omega = (\mathcal{I}_{n-1}\backslash\mathcal{T}_{n-1}) \cup \mathcal{T}_{n-1} \cup \mathcal{S}_{n-1} \cup (\mathcal{A}_{n-1}\backslash\mathcal{S}_{n-1}) \tag{3.2}$$

defines a decomposition of Ω in mutually disjoint sets. Moreover, we have the following relation between these sets at each level n:

$$\mathcal{I}_n = (\mathcal{I}_{n-1}\backslash\mathcal{T}_{n-1}) \cup \mathcal{S}_{n-1}, \qquad \mathcal{A}_n = (\mathcal{A}_{n-1}\backslash\mathcal{S}_{n-1}) \cup \mathcal{T}_{n-1}. \tag{3.3}$$

In fact, as $\Omega = \mathcal{I}_n \cup \mathcal{A}_n$ it is sufficient to prove that

$$(\mathcal{I}_{n-1}\backslash\mathcal{T}_{n-1}) \cup \mathcal{S}_{n-1} \subset \mathcal{I}_n \qquad \text{and} \qquad (\mathcal{A}_{n-1}\backslash\mathcal{S}_{n-1}) \cup \mathcal{T}_{n-1} \subset\ \mathcal{A}_n,$$

that is

$$\mathcal{S}_{n-1} \subset \mathcal{I}_n \quad \text{and} \quad \mathcal{T}_{n-1} \subset\ \mathcal{A}_n.$$

Since $\mathcal{S}_{n-1} \subset \mathcal{A}_{n-1}$ we find $u_{n-1} = b$ on $\mathcal{S}_{n-1}$. From the definition of $\mathcal{S}_{n-1}$ we conclude that $\lambda_{n-1} \le 0$ so that $u_{n-1} + \frac{\lambda_{n-1}}{c} \le b$. This implies $\mathcal{S}_{n-1} \subset \mathcal{I}_n$. The verification of $\mathcal{T}_{n-1} \subset\ \mathcal{A}_n$ is quite similar.

For the convenience of the reader we present these sets in Figure 3.1.

In Figure 3.1 the shaded region depicts $\mathcal{I}_n$ and the white region is $\mathcal{A}_n$. Table 3.1 depicts the signs of primal and dual variables for two consecutive iteration levels.

Below, $\|\Delta^{-1}\|$ will denote the operator norm of Δ^{-1} in $\mathcal{L}(L^2(\Omega))$.

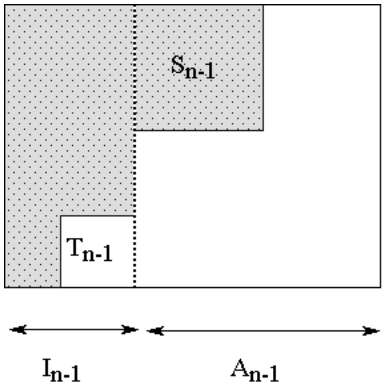


FIG. 3.1. *Decomposition of Ω at levels $n-1$ and n.*

TABLE 3.1

	λ_{n-1}	λ_n	u_{n-1}	u_n
$\mathcal{T}_{n-1} = \mathcal{I}_{n-1} \cap \mathcal{A}_n$	0		$> b$	$= b$
$\mathcal{S}_{n-1} = \mathcal{A}_{n-1} \cap \mathcal{I}_n$	≤ 0	0	$= b$	
$\mathcal{I}_{n-1} \backslash \mathcal{T}_{n-1}$ $(\subset \mathcal{I}_n)$	0	0	$\leq b$	
$\mathcal{A}_{n-1} \backslash \mathcal{S}_{n-1}$ $(\subset \mathcal{A}_n)$	> 0		$= b$	$= b$

THEOREM 3.1. *If $\mathcal{A}_n \neq \mathcal{A}_{n-1}$ and*

$$\alpha + \gamma \leq c \leq \alpha - \frac{\alpha^2}{\gamma} + \frac{\alpha^2}{\|\Delta^{-1}\|^2} \tag{3.4}$$

for some $\gamma > 0$, then $\hat{L}_c(y_n, u_n, \lambda_n) \leq \hat{L}_c(y_{n-1}, u_{n-1}, \lambda_{n-1})$. In addition, if the second inequality of (3.4) *is strict, then either $\hat{L}_c(y_n, u_n, \lambda_n) < \hat{L}_c(y_{n-1}, u_{n-1}, \lambda_{n-1})$ or the algorithm stops at the solution to* $(\mathcal{S})$*.*

Proof. A short computation gives

$$(\lambda, \hat{g}_c(u, \lambda)) + \frac{c}{2}\|\hat{g}_c(u, \lambda)\|^2$$

$$= \left(\frac{1}{\sqrt{c}}\lambda, \sqrt{c}\ \hat{g}_c(u, \lambda)\right) + \frac{1}{2}\left(\sqrt{c}\ \hat{g}_c(u, \lambda), \sqrt{c}\ \hat{g}_c(u, \lambda)\right)$$

$$= \frac{1}{2}\ \left\|\sqrt{c}\max\left(g(u), -\frac{\lambda}{c}\right) + \frac{1}{\sqrt{c}}\lambda\right\|^2 - \frac{1}{2c}\|\lambda\|^2$$

$$= \frac{1}{2}\left\|\max\left(\sqrt{c}\ g(u), -\frac{\lambda}{\sqrt{c}}\right) + \frac{1}{\sqrt{c}}\ \lambda\right\|^2 - \frac{1}{2c}\|\lambda\|^2$$

$$= \frac{1}{2c}\|\max(c\ g(u) + \lambda, 0)\|^2 - \frac{1}{2c}\|\lambda\|^2.$$

Moreover, for all (y, u, λ) we find

$$L_c(y, u, \lambda) = J(y, u) + \frac{1}{2c}\|\max(c\,g(u) + \lambda, 0)\|^2 - \frac{1}{2c}\|\lambda\|^2. \tag{3.5}$$

By assumption $\mathcal{A}_n \neq \mathcal{A}_{n-1}$ and hence $\mathcal{S}_{n-1} \cup \mathcal{T}_{n-1} \neq \emptyset$. Using (3.5) we get

$$\begin{aligned}
&\hat{L}_c(y_n, u_n, \lambda_n) - \hat{L}_c(y_{n-1}, u_{n-1}, \lambda_{n-1})\\
&= J(y_n, u_n) - J(y_{n-1}, u_{n-1})\\
&+\frac{1}{2c}\left[\|\max(c\;g(u_n) + \lambda_n^+, 0)\|^2 - \|\lambda_n^+\|^2 - \|\max(c\;g(u_{n-1}) + \lambda_{n-1}^+, 0)\|^2 + \|\lambda_{n-1}^+\|^2\right]
\end{aligned}$$

and by (3.1)

$$\begin{aligned}
&\hat{L}_c(y_n, u_n, \lambda_n) - \hat{L}_c(y_{n-1}, u_{n-1}, \lambda_{n-1})\\
&= -\frac{1}{2}\|y_{n-1} - y_n\|^2 - \frac{\alpha}{2}\|u_{n-1} - u_n\|^2 + (u_{n-1} - u_n, \lambda_n)_{\mathcal{T}_{n-1}}\\
&+\frac{1}{2c}\left[\|\max(c\;g(u_n) + \lambda_n^+, 0)\|^2 - \|\lambda_n^+\|^2 - \|\max(c\;g(u_{n-1}) + \lambda_{n-1}^+, 0)\|^2 + \|\lambda_{n-1}^+\|^2\right].
\end{aligned} \tag{3.6}$$

It will be convenient to introduce

$$\begin{aligned}
d(x) = |\max(c\,g(u_n(x)) + \lambda_n^+(x), 0)|^2 - |\lambda_n^+(x)|^2 - |\max(c\,g(u_{n-1}(x))\\
+ \lambda_{n-1}^+(x), 0)|^2 + |\lambda_{n-1}^+(x)|^2.
\end{aligned}$$

Let us estimate d on the four distinct subsets of Ω according to (3.2).

1. On $\mathcal{I}_{n-1}\backslash\mathcal{T}_{n-1}$ we have $\lambda_n(x) = \lambda_{n-1}(x) = 0,\ u_{n-1}(x) \leq b(x)\ (g(u_{n-1}(x)) \leq 0)$, and

$$d(x) = |\max(c\,g(u_n(x)), 0)|^2 - |\max(c\,g(u_{n-1}(x)), 0)|^2 \leq c^2|u_n(x) - u_{n-1}(x)|^2.$$

Moreover, as $\lambda_n = p_n - \alpha(u_n - u_d) = 0$ and $\lambda_{n-1} = p_{n-1} - \alpha(u_{n-1} - u_d) = 0$ we have $u_n(x) - u_{n-1}(x) = \frac{p_n(x) - p_{n-1}(x)}{\alpha}$ so that

$$|u_n(x) - u_{n-1}(x)| \leq \frac{1}{\alpha}|p_n(x) - p_{n-1}(x)| \quad \text{on } \mathcal{I}_{n-1}\backslash\mathcal{T}_{n-1}.$$

2. On $\mathcal{S}_{n-1}$, $\lambda_n(x) = 0$, $\lambda_{n-1}(x) \leq 0$, $g(u_{n-1}(x)) = 0$ so that $d(x) = |\max(c\,g(u_n(x)), 0)|^2$. Here we used the positivity of λ^+ to get $\lambda_{n-1}^+(x) = 0$. To estimate $d(x)$ in detail we consider first the case where $u_n(x) \geq b(x)$. Since $x \in \mathcal{S}_{n-1} \subset \mathcal{I}_n$ we obtain $\lambda_n(x) = p_n(x) - \alpha[u_n(x) - u_d(x)] = 0$ and hence $u_n(x) = \frac{p_n(x)}{\alpha} + u_d(x)$. Moreover, $\lambda_{n-1}(x) = p_{n-1}(x) - \alpha[u_{n-1}(x) - u_d(x)] \leq 0$ so that $u_d(x) - b(x) \leq -\frac{p_{n-1}(x)}{\alpha}$, where we used $u_{n-1}(x) = b(x)$. Since by assumption $u_n(x) \geq b$ these estimates imply

$$\begin{aligned}
|u_n(x) - u_{n-1}(x)| &= u_n(x) - b(x)\\
&= \frac{p_n(x)}{\alpha} + u_d(x) - b(x) \leq \frac{p_n(x)}{\alpha} - \frac{p_{n-1}(x)}{\alpha} = \frac{1}{\alpha}\,|p_n(x) - p_{n-1}(x)|.
\end{aligned}$$

In addition it is clear that on the set $\mathcal{I}_n$

$$d(x) = |\max(c\,g(u_n(x)), 0)|^2 \leq c^2|u_n(x) - u_{n-1}(x)|^2.$$

In the second case, $u_n(x) < b(x)$ so that $\max(c\,g(u_n(x)), 0) = 0$ and $d(x) = 0$. Finally, we have a precise estimate on the whole set $\mathcal{I}_n$. Let us denote

$$\mathcal{I}_n^* = \mathcal{I}_{n-1}\backslash\mathcal{T}_{n-1} \cup \{x \in \mathcal{S}_{n-1} \mid u_n(x) \geq b(x)\} \;= \mathcal{I}_n\backslash\{x \in \mathcal{S}_{n-1} \mid u_n(x) < b(x)\};$$

then

$$\int_{\mathcal{I}_n} d(x)\;dx = \int_{\mathcal{I}_n^*} d(x)\;dx = c^2\|\max(g(u_n), 0)\|^2_{\mathcal{I}_n^*} \leq c^2\;\|u_n - u_{n-1}\|^2_{\mathcal{I}_n^*}. \tag{3.7}$$

We note that we have proved in addition that

$$\|u_n - u_{n-1}\|_{\mathcal{I}_n^*} \le \frac{\|\Delta^{-1}\|}{\alpha}\|y_n - y_{n-1}\|. \tag{3.8}$$

3. On $\mathcal{A}_{n-1}\backslash\mathcal{S}_{n-1}$ we have $g(u_{n-1}(x)) = g(u_n(x)) = 0,\ \lambda_{n-1}(x) > 0$ and hence

$$d(x) = |\max(\lambda_n^+(x), 0)|^2 - |\lambda_n^+(x)|^2 \le 0. \tag{3.9}$$

4. On $\mathcal{T}_{n-1}$ we have $\lambda_{n-1}(x) = 0,\ g(u_n(x)) = 0,\ g(u_{n-1}(x)) > 0$ and thus

$$d(x) = -c^2|g(u_{n-1}(x))|^2 = -c^2|u_n(x) - u_{n-1}(x)|^2. \tag{3.10}$$

Next we estimate the term $(\lambda_n, u_{n-1} - u_n)_{\mathcal{T}_{n-1}}$ in (3.6):

$$\begin{aligned}(\lambda_n, u_{n-1} - u_n)_{\mathcal{T}_{n-1}} &= (\lambda_n - \lambda_{n-1}, u_{n-1} - u_n)_{\mathcal{T}_{n-1}}\\ &= (p_n - p_{n-1}, u_{n-1} - u_n)_{\mathcal{T}_{n-1}} + \alpha\|u_n - u_{n-1}\|^2_{\mathcal{T}_{n-1}}\end{aligned}$$

and therefore

$$\begin{aligned}&(\lambda_n, u_{n-1} - u_n)_{\mathcal{T}_{n-1}}\\ &\qquad \le \|\Delta^{-1}\|\ \|y_n - y_{n-1}\|_\Omega\|u_n - u_{n-1}\|_{\mathcal{T}_{n-1}} + \alpha\|u_n - u_{n-1}\|^2_{\mathcal{T}_{n-1}}.\end{aligned} \tag{3.11}$$

Inserting (3.7)–(3.11) into (3.6) we find

$$\begin{aligned}&\hat{L}_c(y_n, u_n, \lambda_n) - \hat{L}_c(y_{n-1}, u_{n-1}, \lambda_{n-1})\\ &\le -\frac{1}{2}\|y_{n-1} - y_n\|^2 - \frac{\alpha}{2}\|u_{n-1} - u_n\|^2_{\mathcal{I}_n^*} - \frac{\alpha}{2}\|u_{n-1} - u_n\|^2_{\mathcal{I}_n\backslash\mathcal{I}_n^*} - \frac{\alpha}{2}\|u_{n-1} - u_n\|^2_{\mathcal{T}_{n-1}}\\ &+\|\Delta^{-1}\|\ \|y_n - y_{n-1}\|_\Omega\|u_n - u_{n-1}\|_{\mathcal{T}_{n-1}} + \alpha\|u_n - u_{n-1}\|^2_{\mathcal{T}_{n-1}}\\ &+\frac{c}{2}\|u_{n-1} - u_n\|^2_{\mathcal{I}_n^*} - \frac{c}{2}\|u_{n-1} - u_n\|^2_{\mathcal{T}_{n-1}}.\end{aligned} \tag{3.12}$$

Using $ab \le \frac{1}{2}(\frac{a^2}{\rho} + \rho b^2)$ for every $\rho > 0$ and relation (3.8), we get for $c \ge \alpha$

$$\begin{aligned}&\hat{L}_c(y_n, u_n, \lambda_n) - \hat{L}_c(y_{n-1}, u_{n-1}, \lambda_{n-1})\\ &\le -\frac{1}{2}\|y_{n-1} - y_n\|^2 + \frac{(c-\alpha)}{2}\|u_{n-1} - u_n\|^2_{\mathcal{I}_n^*} + \frac{(\alpha - c)}{2}\|u_{n-1} - u_n\|^2_{\mathcal{T}_{n-1}}\\ &+\frac{\|\Delta^{-1}\|}{2\rho}\|y_{n-1} - y_n\|^2 + \frac{\rho\|\Delta^{-1}\|}{2}\|u_{n-1} - u_n\|^2_{\mathcal{T}_{n-1}}\\ &\le -\frac{1}{2}\|y_{n-1} - y_n\|^2 + \frac{(c-\alpha)\|\Delta^{-1}\|^2}{2\alpha^2}\|y_{n-1} - y_n\|^2\\ &+\frac{\alpha - c + \rho\|\Delta^{-1}\|}{2}\|u_{n-1} - u_n\|^2_{\mathcal{T}_{n-1}} + \frac{\|\Delta^{-1}\|}{2\rho}\|y_{n-1} - y_n\|^2\\ &= \frac{1}{2}\left[(c-\alpha)\frac{\|\Delta^{-1}\|^2}{\alpha^2} + \frac{\|\Delta^{-1}\|}{\rho} - 1\right]\|y_{n-1} - y_n\|^2\\ &+\frac{1}{2}(\alpha + \rho\|\Delta^{-1}\| - c)\|u_{n-1} - u_n\|^2_{\mathcal{T}_{n-1}}.\end{aligned}$$

Setting $\gamma = \rho\|\Delta^{-1}\|$, then $\hat{L}_c(y_n, u_n, \lambda_n) \le \hat{L}_c(y_{n-1}, u_{n-1}, \lambda_{n-1})$ provided that

$$\left[\left[\frac{(c-\alpha)}{\alpha^2} + \frac{1}{\gamma}\right]\|\Delta^{-1}\|^2 - 1\right] \le 0 \quad \text{and} \quad \alpha + \gamma - c \le 0.$$

The latter condition is equivalent to

$$\alpha + \gamma \leq c \leq \alpha - \frac{\alpha^2}{\gamma} + \frac{\alpha^2}{\|\Delta^{-1}\|^2}. \tag{3.4}$$

If the second inequality is strict, then $\hat{L}_c(y_n, u_n, \lambda_n) < \hat{L}_c(y_{n-1}, u_{n-1}, \lambda_{n-1})$ except if $y_n = y_{n-1}$. In this latter case $u_n = u_{n-1}$ and the algorithm stops at the solution to $(\mathcal{S})$. □

REMARK 3.1. *Note that for the choice $\gamma = \alpha$ condition* (3.4) *is equivalent to*

$$2\,\alpha \leq c \leq \frac{\alpha^2}{\|\Delta^{-1}\|^2}. \tag{3.13}$$

REMARK 3.2. *If there exists γ such that* (3.4) *holds, then necessarily*

$$c > \alpha \geq 2\|\Delta^{-1}\|^2$$

holds. Indeed, assume that $\alpha < 2\|\Delta^{-1}\|^2$. Then

$$\alpha + \gamma < \alpha - \frac{\alpha^2}{\gamma} + 2\alpha,$$

that is

$$\gamma^2 - 2\alpha\gamma + \alpha^2 = (\gamma - \alpha)^2 < 0,$$

which is a contradiction.

3.2. The discrete case. So far we have given a sufficient condition for $\hat{L}_c$ to act as a merit function for which the algorithm has a strict descent property. In particular this eliminates the possibility of chattering of the algorithm: it will not return to the same active set a second time. If the control and state spaces are discretized, then the descent property can be used to argue convergence in a finite number of steps. More precisely, assume that a finite difference or finite element–based approximation to $(\mathcal{P})$ results in

$$(\mathcal{P}^{N,M}) \qquad \begin{array}{l} \min \quad J^{N,M}(Y,U) = \frac{1}{2}\|M_1^{\frac{1}{2}}(Y - Z_d)\|_{\mathbb{R}^N}^2 + \frac{\alpha}{2}\|M_2^{\frac{1}{2}}(U - U_d)\|_{\mathbb{R}^M}^2, \\ S\,Y = M_3\,U, \\ U \leq B. \end{array}$$

Here Y and Z_d denote vectors in $\mathbb{R}^N$ corresponding to the discretization of y and z_d, and U, U_d, and B denote vectors in $\mathbb{R}^M$, corresponding to the discretizations of u, u_d, and b. Furthermore M_1, S, and M_2 are, respectively, $N \times N$, $N \times N$, and $M \times M$ positive definite matrices while M_3 is an $N \times M$ matrix. The norms in $(\mathcal{P}^{N,M})$ denote Euclidean norms and the inequality is understood coordinatewise. Finally, it is assumed that M_2 is a diagonal matrix. It is simple to argue the existence of a solution (Y^*, U^*) to $(\mathcal{P}^{N,M})$. A first order optimality system is given by

$$\left\{ \begin{array}{rcl} S\,Y^* &=& M_3\,U^*, \\ S\,P^* &=& -M_1(Y^* - Z_d), \\ U^* &=& U_d + \frac{1}{\alpha}M_2^{-1}(M_3^\top P^* - \Lambda^*), \\ \Lambda^* &=& c\max\left(0, U^* + \frac{1}{c}\Lambda^* - B\right), \end{array} \right. \tag{3.14}$$

with $(P^*, \Lambda^*) \in \mathbb{R}^N \times \mathbb{R}^M$ for every $c > 0$. Here max is understood coordinatewise. The algorithm for the discretized problem is given next.

Discretized algorithm.

1. Initialization: choose Y^o, U^o, and Λ^o, and set $n = 1$.
2. Determine the following subsets of $\{1, \dots, M\}$:

$$A_n = \left\{ i \mid U_i^{n-1} + \frac{1}{c}\Lambda_i^{n-1} > B_i \right\}, \quad I_n = \{1, \dots, M\} \backslash A_n.$$

3. If $n \geq 2$ and $A_n = A_{n-1}$, then STOP.
4. Else, find $(Y^n, P^n) \in \mathbb{R}^N \times \mathbb{R}^N$ such that

$$\begin{aligned} S\,Y^n &= M_3 \begin{cases} B & \text{in } A_n, \\ U_d + \dfrac{1}{\alpha} M_2^{-1} M_3^\top \; P^n & \text{in } I_n, \end{cases} \\ S\,P^n &= -M_1(Y^n - Z_d) \end{aligned}$$

and set

$$U^n = \begin{cases} B & \text{in } A_n, \\ U_d + \dfrac{1}{\alpha} M_2^{-1} M_3^\top \; P^n & \text{in } I_n. \end{cases}$$

5. Set $\Lambda^n = M_3^\top P^n - \alpha M_2 (U^n - U_d)$, update $n = n + 1$, and goto 2.

The following corollary describing properties of the discretized algorithm can be obtained with techniques analogous to those utilized above for analyzing the continuous algorithm. We shall denote

$$\underline{m_2} = \min_i (M_2)_{i,i}, \qquad \overline{m_2} = \max_i (M_2)_{i,i}, \qquad \text{and } K = \|M_2^{-1} M_3^\top\| \; \|S^{-1} M_1\|.$$

Corollary 3.1. *If*

$$\overline{m_2}\,(\alpha + \gamma) \leq c < \alpha\,\underline{m_2} - \frac{\alpha^2}{\gamma} + \frac{\alpha^2 \|M_1\|}{K} \tag{3.15}$$

holds for some $\gamma > 0$, then the discretized algorithm converges in finitely many steps to the solution of $(\mathcal{P}^N)$.

Proof. First we observe that if the discretized algorithm stops in step 3, then the current iterate gives the unique solution. Then we show with an argument analogous to that of the proof of Theorem 3.1 that with (3.15) holding, we have $L_c^{N,M}(Y_n, U_n, \Lambda_n) < L_c^{N,M}(Y_{n-1}, U_{n-1}, \Lambda_{n-1})$ or $(Y_n, U_n) = (Y_{n-1}, U_{n-1})$, where the discretized merit function is given by

$$\begin{aligned} L_c^{N,M}(Y, U, \Lambda) = {} & \frac{1}{2} \|M_1^{\frac{1}{2}} \; (Y - Z_d)\|^2_{\mathbb{R}^N} \\ & + \frac{\alpha}{2} \|M_2^{\frac{1}{2}} \; (U - U_d)\|^2_{\mathbb{R}^M} \\ & + (\Lambda, \hat{g}_c(U, \Lambda))_{\mathbb{R}^M} + \frac{c}{2} \|\hat{g}_c(U, \Lambda)\|^2_{\mathbb{R}^M} \end{aligned}$$

with $\hat{g}_c(U, \Lambda) = (\max(U_1 - B_1, -\frac{\Lambda_1}{c}), \dots, \max(U_M - B_M, -\frac{\Lambda_M}{c}))^\top$. If $(Y_n, U_n) = (Y_{n-1}, U_{n-1})$, then $A_{n+1} = A_n$ and the discretized algorithm stops at the solution. The case $L_c^{N,M}(Y_n, U_n, \Lambda_n) < L_c^{N,M}(Y_{n-1}, U_{n-1}, \Lambda_{n-1})$ cannot occur for infinitely many n since there are only finitely many different combinations of active index sets. In fact, assume that there exists $p < n$ such that $A_n = A_p$ and $I_n = I_p$. Since (Y_n, U_n)

is a solution of the optimality system of step 4 if and only if (Y_n, U_n) is the unique solution of

$$\min\{\ J^{N,M}(y,u) \mid S\ Y = M_3\ U,\ U = B \ \text{ in } A_n\ \},$$

it follows that $Y_n = Y_p$, $U_n = U_p$, and $\Lambda_n = \Lambda_p$. This contradicts $L_c^{N,M}(Y_n, U_n, \Lambda_n) < L_c^{N,M}(Y_p, U_p, \Lambda_p)$ and ends the proof. □

REMARK 3.3. *Note that for* $\gamma = \frac{\alpha}{\underline{m_2}}$ *condition* (3.15) *is satisfied if*

$$\overline{m_2}\,\alpha\left(1 + \frac{1}{\underline{m_2}}\right) \le c < \frac{\alpha^2 \|M_1\|}{K}.$$

Therefore, one can choose $c = \overline{m_2}\,\alpha\,(1 + \frac{1}{\underline{m_2}})$ *for any*

$$\alpha > \frac{\overline{m_2} K}{\|M_1\|}\left(1 + \frac{1}{\underline{m_2}}\right).$$

4. Ascent properties of the algorithm. In the previous section sufficient conditions for convergence of the algorithm in terms of α, c, and $\|\Delta^{-1}\|$ were given. Numerical experiments showed that the algorithm converges also for values of α, c, and $\|\Delta^{-1}\|$ which do not satisfy the conditions of Theorem 3.1. In fact the only possibility of constructing an example for which the algorithm has some difficulties (which will be made precise in the following section) is based on violating the strict complementarity condition.

Thus one is challenged to further justify theoretically the efficient behavior of the algorithm. In the tests that were performed it was observed that the cost functional was always increasing so that in practice the algorithm behaves like an infeasible algorithm. To parallel theoretically this behavior of the algorithm as far as possible, we slightly modify the algorithm. For the modified algorithm an ascent property of the cost J will be shown.

MODIFIED ALGORITHM.

1. Initialization: choose u_o, y_o, and λ_o; set $n = 1$.
2. (a) Determine the following subsets of Ω:

$$\mathcal{A}_n = \left\{\ x \mid u_{n-1}(x) + \frac{\lambda_{n-1}(x)}{c} > b \right\},\quad \mathcal{I}_n = \left\{\ x \mid u_{n-1}(x) + \frac{\lambda_{n-1}(x)}{c} \le b \right\},$$

(b) and find $(\tilde{y}, \tilde{p}) \in H_o^1(\Omega) \times H_o^1(\Omega)$ such that

$$\begin{aligned} -\Delta \tilde{y} &= \begin{cases} b & \text{in } \mathcal{A}_n, \\ u_d + \dfrac{\tilde{p}}{\alpha} & \text{in } \mathcal{I}_n, \end{cases} \\ -\Delta \tilde{p} &= z_d - \tilde{y} \ \text{ in } \Omega, \end{aligned}$$

and set

$$\tilde{u} = \begin{cases} b & \text{in } \mathcal{A}_n, \\ u_d + \dfrac{\tilde{p}}{\alpha} & \text{in } \mathcal{I}_n. \end{cases}$$

3. $\tilde{\lambda} = \tilde{p} - \alpha(\tilde{u} - u_d)$.

4. Set

$$\widetilde{\mathcal{A}} = \left\{ x \mid \tilde{u}(x) + \frac{\tilde{\lambda}(x)}{c} > b \right\}.$$

If $\widetilde{\mathcal{A}} = \mathcal{A}_n$, then STOP, else goto 5.

5. Check for $J(\tilde{y}, \tilde{u}) > J(y_{n-1}, u_{n-1})$.
   (a) If $J(\tilde{y}, \tilde{u}) > J(y_{n-1}, u_{n-1})$, then

   $n = n+1,\ y_n = \tilde{y},\ u_n = \tilde{u},\ \lambda_n = \tilde{\lambda}$, and goto 2(a).

   (b) Otherwise, determine

$$\mathcal{T}_{n-1} = \{\ x \in \mathcal{I}_{n-1} \mid u_{n-1}(x) > b\ \}.$$

   - If measure of $\mathcal{T}_{n-1}$ is null, then STOP;
   - else set

$$\mathcal{A}_n = \mathcal{A}_{n-1} \cup \mathcal{T}_{n-1}, \quad \mathcal{I}_n = \mathcal{I}_{n-1} \backslash \mathcal{T}_{n-1},$$

   then goto 2(b).

THEOREM 4.1. *If the modified algorithm stops in step* 4, *then* $(\tilde{u}, \tilde{y}, \tilde{\lambda})$ *is the solution to* $(\mathcal{S})$. *If it never stops in step* 5(b), *then the sequence* $J(y_n, u_n)$ *(*$n \geq 2$*) is strictly increasing and converges to some* J^*.

Proof. Let us first assume that the algorithm stops in step 4. In case $\mathcal{A}_n$ is calculated from 2(a), then $(\tilde{u}, \tilde{y}, \tilde{\lambda})$ is the solution to $(\mathcal{S})$ by Theorem 2.1. If $\mathcal{A}_n$ is determined from 5(b), then an argument analogous to that used in the proof of Theorem 2.1 allows us to argue that again $(\tilde{u}, \tilde{y}, \tilde{\lambda})$ is the solution to $(\mathcal{S})$.

Next we assume that the algorithm never stops in step 4. Let us consider an iteration level, where the check for ascent in step 5(a) is not passed. Consequently $\mathcal{A}_n$ and $\mathcal{I}_n$ are redefined according to step 5(b) and $(\tilde{y}, \tilde{u})$ are recalculated from 2(b). We have already noticed that $(\tilde{y}, \tilde{u})$ is a solution of the optimality system of step 2(b) if and only if $(\tilde{y}, \tilde{u})$ is the unique solution of

$$(\mathcal{P}_{aux}) \qquad \min\{\ J(y,u) \mid -\Delta y = u \text{ in } \Omega, y \in H^1_o(\Omega),\ u = b \text{ in } \mathcal{A}_n\ \}.$$

Since $\mathcal{A}_n = \mathcal{A}_{n-1} \cup \mathcal{T}_{n-1}$ strictly contains $\mathcal{A}_{n-1}$ it necessarily follows that

$$(4.1) \qquad J(y_{n-1}, u_{n-1}) \leq J(\tilde{y}, \tilde{u}).$$

It will be shown next that equality in (4.1) is impossible. In fact if $J(\tilde{y}, \tilde{u}) = J(y_{n-1}, u_{n-1})$, then due to uniqueness of the solution to $(\mathcal{P}_{aux})$ it follows that $(\tilde{y}, \tilde{u}) = (y_{n-1}, u_{n-1})$ and consequently $\tilde{\lambda} = \lambda_{n-1}$. On $\mathcal{A}_n = \mathcal{A}_{n-1} \cup \mathcal{T}_{n-1}$, we get $\tilde{u} = b = u_{n-1}$. This implies that $u_{n-1} = b$ on $\mathcal{T}_{n-1}$ and gives a contradiction to the assumption that the measure of $\mathcal{T}_{n-1}$ is non-null. Hence $J(y_{n-1}, u_{n-1}) = J(\tilde{y}, \tilde{u})$ is impossible. Together with (4.1) it follows that $J(y_{n-1}, u_{n-1}) < J(\tilde{y}, \tilde{u})$ and thus the sequence $\{J(y_n, u_n)\}$ generated by the modified algorithm is strictly increasing. The pair (y_b, b) with $-\Delta y_b = b$ in Ω is feasible for all $(\mathcal{P}_{aux})$ so that $J(y_n, u_n) \leq J(y_b, b)$. It follows that $J(y_n, u_n)$ is convergent to some J^*.

We note in addition that $\tilde{u}$ is feasible since $\tilde{u} = u_{n-1} = u_{n-1} + \frac{\lambda_{n-1}}{c} \leq b$ on $\mathcal{I}_n$ $(\lambda_{n-1} = \tilde{\lambda} = 0$ on $\mathcal{I}_n)$. □

The previous result can be strengthened in the case where $(\mathcal{P})$ is discretized as in subsection 3.1.

Table 5.1
Example 1(a): $u_d \equiv 0$, $\alpha = 10^{-2}$, $c = 10^{-1}$.

Iteration	$\max(u_n - b)$	Size of $\mathcal{A}_n$	$J(y_n, u_n)$	$L_c(y_n, u_n, \lambda_n)$	$\hat{L}_c(y_n, u_n, \lambda_n)$
1	4.8708e-02	1250	4.190703e-02	4.190785e-02	4.190785e-02
2	5.8230e-05	1331	4.190712e-02	4.190712e-02	4.190712e-02
3	0.0000e+00	1332	4.190712e-02	4.190712e-02	4.190712e-02
4	0.0000e+00	1332	4.190712e-02	4.190712e-02	4.190712e-02

Corollary 4.1. *If the modified algorithm is discretized as described in the previous section and if it never stops in step* 5(b), *then the (discretized) solution is obtained in finitely many steps.*

Proof. Unless the algorithm stops in step 4, the values of $J^N(Y_n, U_n)$ $(n \geq 2)$ are strictly increasing. As argued in the proof of Corollary 3.1, at each level of the iteration the minimization is carried out over an active set different from all those that have been computed before. As there are only finitely many different possibilities for active sets, the modified algorithm terminates in step 4 at the unique solution of $(\mathcal{S})$. □

We have not found a numerical example in which the modified algorithm terminates in step 5(b).

5. Numerical experiments. In this section we report on numerical tests with the proposed algorithm. For these tests we chose $\Omega =]0,1[\times]0,1[$ and the five-point finite difference approximation of the Laplacian. Unless otherwise specified the discretization was carried out on a uniform mesh with grid size 1/50.

For the chosen dimension $\|\Delta^{-1}\| = \frac{1}{2\pi^2}$ so that $\frac{1}{\|\Delta^{-1}\|^2} = 4\pi^4 \simeq 390$. Relation (3.13), which is required for the applicability of Theorem 3.1, is satisfied if $\alpha \geq 5.\ 10^{-3}$ to get the convergence via Theorem 3.1. Nevertheless we have also tested the method for smaller values of α.

The tests were performed on a Hewlett-Packard workstation using the MATLAB package.

5.1. Example 1. We set

$$z_d(x_1, x_2) = \sin\ (2\pi x_1)\ \sin\ (2\pi x_2)\ \exp(2x_1)/6, \quad b \equiv 0.$$

Several tests for different values for α, c, and u_d were performed. We present two of them. In the first one (3.13) is satisfied with strict inequalities; see Table 5.1.

Plots of the optimal state and optimal control are shown in Figure 5.1.

We present in Table 5.2 a second example where (3.13) is not fulfilled because α is too small; in addition u_d has been chosen infeasible.

Although the size of the set $\mathcal{A}_n$, in the sense of number of grid points, in $\mathcal{A}_n$ is increasing, the sequence $\mathcal{A}_n$ does not increase monotonically. More precisely, points in $\mathcal{A}_n$ at iteration n may not belong to $\mathcal{A}_{n+1}$ at iteration $n+1$.

We observe numerically that the algorithm stops as soon as an iterate is feasible. Thus the sequence of iterates is not feasible until it reaches the solution. We could say that we have an "outer" method. We must also underline that unlike with classical primal active set methods, the primal-dual method that we propose can move a lot of points from one iteration to the next.

We compared the new algorithm to an Uzawa method for the augmented Lagrangian with Gauss–Seidel splitting. For convenience we recall that algorithm.

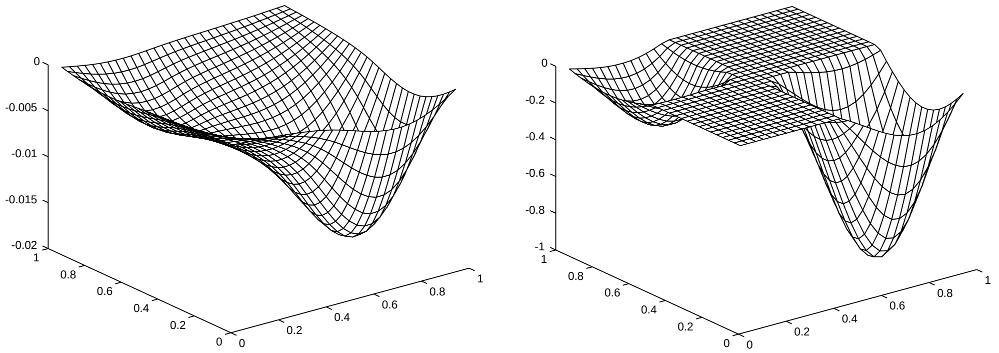


FIG. 5.1. *Optimal state (left), optimal control (right).*

TABLE 5.2
Example 1(b): $u_d \equiv 1,\ \alpha = 10^{-6},\ c = 10^{-2}$.

Iteration	$\max(u_n - b)$	Size of $\mathcal{A}_n$	$J(y_n, u_n)$	$L_c(y_n, u_n, \lambda_n)$	$\hat{L}_c(y_n, u_n, \lambda_n)$
1	5.0986e+02	1250	1.734351e-02	9.858325e+00	9.858325e+00
2	4.4728e+02	1487	2.089663e-02	7.688683e+00	7.688683e+00
3	3.6796e+02	1677	2.375001e-02	5.612075e+00	5.612075e+00
4	5.8313e+02	1831	2.603213e-02	4.526200e+00	4.526200e+00
5	6.7329e+02	1944	2.782111e-02	3.657995e+00	3.657995e+00
6	5.3724e+02	2039	2.911665e-02	2.402021e+00	2.402021e+00
7	3.6175e+02	2098	2.981378e-02	1.191161e+00	1.191161e+00
8	1.5071e+02	2146	3.011540e-02	3.678089e-01	3.678089e-01
9	6.5928e+01	2178	3.018832e-02	7.796022e-02	7.796022e-02
10	2.3420e+01	2196	3.019715e-02	3.344241e-02	3.344241e-02
11	3.4889e+00	2208	3.019762e-02	3.022994e-02	3.022994e-02
12	0.0000e+00	2210	3.019762e-02	3.019762e-02	3.019762e-02
13	0.0000e+00	2210	3.019762e-02	3.019762e-02	3.019762e-02

ALGORITHM UGS.

- Step 1. Initialization: Set $n = 1$ and choose $\gamma > 0$.
  Choose $q_o \in L^2(\Omega)$ and $u_{-1} \in L^2(\Omega)$.
- Step 2. Choose $k_n \in \mathbb{N}$, set $u_n^{-1} = u_{n-1}$, and for $j = 0, \ldots, k_n$

$$\begin{aligned} y_n^j &= \text{ Arg min } \{\ L_\gamma(y, u_n^{j-1}, q_n) \mid y \in H^2(\Omega) \cap H_o^1(\Omega)\ \}, \\ u_n^j &= \text{ Arg min } \{\ L_\gamma(y_n^j, u, q_n) \mid u \in U_{ad}\ \}. \end{aligned}$$

  End of the inner loop: $y_n = y_n^{k_n}, u_n = u_n^{k_n}$.
- Step 3.

$$q_{n+1} = q_n + \frac{\rho}{k_n + 1} \sum_{j=0}^{k_n} (Ay_n^j - u_n^j), \qquad \text{where } \rho \in (0, 2\gamma],$$

where

$$L_\gamma(y, u, q) = J(y, u) + (q, Ay - u)_{L^2(\Omega)} + \frac{\gamma}{2} \|Ay - u\|_{L^2(\Omega)}^2.$$

For this algorithm a detailed convergence analysis was given in [BK]. Due to the splitting technique the second constrained minimization in Step 2 can be carried out by a simple algebraic manipulation. Algorithm UGS is an iterative algorithm that approximates the solution (y^*, u^*), whereas the new algorithm obtains the exact (discretized) solution. For Example 1(a) (Table 5.1) the computing time was 61

Table 5.3
$u_o \equiv 0\ (\equiv b)$.

Iteration	$\max(u_n - b)$	size of $\mathcal{A}_n$	$J(y_n, u_n)$	$L_c(y_n, u_n, \lambda_n)$	$\hat{L}_c(y_n, u_n, \lambda_n)$
1	4.4409e-15	1385	4.296739e-02	4.296739e-02	4.296739e-02
2	1.2546e-14	160	4.296739e-02	4.296739e-02	4.296739e-02
3	3.2752e-15	2078	4.296739e-02	4.296739e-02	4.296739e-02
4	4.5519e-15	2308	4.296739e-02	4.296739e-02	4.296739e-02
5	4.5242e-15	1613	4.296739e-02	4.296739e-02	4.296739e-02
6	4.3299e-15	1787	4.296739e-02	4.296739e-02	4.296739e-02

seconds whereas Algorithm UGS with accuracy set at 10^{-3} was stopped after 105 minutes. At that moment the difference between the algorithm and Algorithm UGS was

$$|J_{ugs} - J(y^*, u^*)| \approx 4.10^{-8},\ \|y_{ugs} - y^*\|_{L^\infty} \approx 8.10^{-7},\ \text{and}\ \|u_{ugs} - u^*\|_{L^\infty} \approx 4.10^{-6},$$

where the index "ugs" refers to the result from Algorithm UGS. For Example 1(b) (Table 5.2) the algorithm took 191 seconds whereas Algorithm UGS was stopped after 120 minutes.

5.2. Example 2. The desired state z_d, b is set as in the previous example and $\alpha = 10^{-2}$, $c = 10^{-1}$. See Table 5.3. This example has been constructed such that there is no strict complementarity at the solution. More precisely we have set $u_d = b - \frac{1}{\alpha}[-\Delta^{-1} z_d + \Delta^{-2} b]$ so that the exact solution of problem $(\mathcal{P})$ is $u^* = b = 0$ and $\lambda^* = 0$ and hence λ^* is not positive where the constraint is active. This example was considered by means of the optimality system $(\mathcal{S})$ of Theorem 1.1.

In this example the canonical initial guess u_o coincides with the solution u^*. From Table 5.3 we observe that u_n, $J(y_n, u_n)$, $L_c(y_n, u_n)$, and $\hat{L}_c(y_n, u_n)$ remain constant while the active sets $\mathcal{A}_n$ chatter. For different initial guesses for u_o the same type of behavior is observed, the algorithm always reaches the optimal value for u and J in one iteration, and if the stopping criterion of the algorithm was based on the coincidence of two consecutive values of J, it would stop after one iteration. The chattering of active sets is due to lack of strict complementarity and machine precision. Let us briefly consider this phenomenon and note at first that the signs in the algorithm are set such that at the limit we should have $\Omega = \mathcal{I}^*$ (all inactive with $\lambda^* = u^* = 0$). If $x \in \mathcal{A}_{n-1}$, then $u_{n-1}(x) = 0$ by step 4 and $\lambda_{n-1}(x) = \pm\varepsilon$, with ε equal to the computer epsilon, will decide whether $x \in \mathcal{A}_n$ or $\mathcal{I}_n$, although for numerical purposes the exact pair for (u, λ) is already obtained. If $x \in \mathcal{I}_{n-1}$, then $\lambda_{n-1} = 0$ and $u_{n-1}(x) = \pm\varepsilon$ will decide whether $x \in \mathcal{A}_n$ or $\mathcal{I}_n$, while the influence of this choice on J or L_c is of the order of ε^2, i.e., it is numerically zero. Therefore we decided to replace "$> b$" in the definition of $\mathcal{A}_n$ by "$> b - \varepsilon$ " (and $\mathcal{I}_n = \Omega \backslash \mathcal{A}_n$): the algorithm now behaves as expected and stops after two iterations.

5.3. Example 3. We have seen with Example 1 (Tables 5.1 and 5.2) that the augmented Lagrangian function decreases during iterations. We show with this example that the augmented Lagrangian function may not decrease although the method is convergent and provides the exact solution. Let us precise the data:

$$z_d = \begin{cases} 200\ x_1 x_2\ (x_1 - \frac{1}{2})^2\ (1 - x_2) & \text{if } 0 < x_1 \le 1/2, \\ 200\ x_2\ (x_1 - 1)(x_1 - \frac{1}{2})^2\ (1 - x_2) & \text{if } 1/2 < x_1 \le 1, \end{cases}$$

$$u_d \equiv 0,\ b \equiv 1,\ c = 10^{-2}.$$

Table 5.4
Example 3(a): $\alpha = 10^{-6}$, $u_o \equiv 1$ ($\equiv b$).

Iteration	$\max(u_n - b)$	Size of $\mathcal{A}_n$	$J(y_n, u_n)$	$L_c(y_n, u_n, \lambda_n)$	$\hat{L}_c(y_n, u_n, \lambda_n)$
1	4.1995e+02	1100	3.314755e-02	9.645226e+00	9.645226e+00
2	3.8057e+02	1370	3.672870e-02	7.943326e+00	7.943326e+00
3	3.6453e+02	1300	3.963515e-02	7.393744e+00	7.393744e+00
4	3.7512e+02	1400	4.249987e-02	7.809205e+00	7.809205e+00
5	3.8952e+02	1500	4.555558e-02	8.300084e+00	8.300084e+00
6	3.9452e+02	1600	4.880515e-02	8.320358e+00	8.320358e+00
7	3.8004e+02	1700	5.203947e-02	7.485445e+00	7.485445e+00
8	3.3858e+02	1800	5.490267e-02	5.699382e+00	5.699382e+00
9	2.6458e+02	1898	5.701220e-02	3.286759e+00	3.286759e+00
10	1.5311e+02	1986	5.811845e-02	1.093548e+00	1.093548e+00
11	8.3048e+01	2040	5.834162e-02	3.099587e-01	3.099587e-01
12	1.5809e+01	2086	5.839423e-02	5.959874e-02	5.959874e-02
13	0.0000e+00	2098	5.839438e-02	5.839438e-02	5.839438e-02
14	0.0000e+00	2098	5.839438e-02	5.839438e-02	5.839438e-02

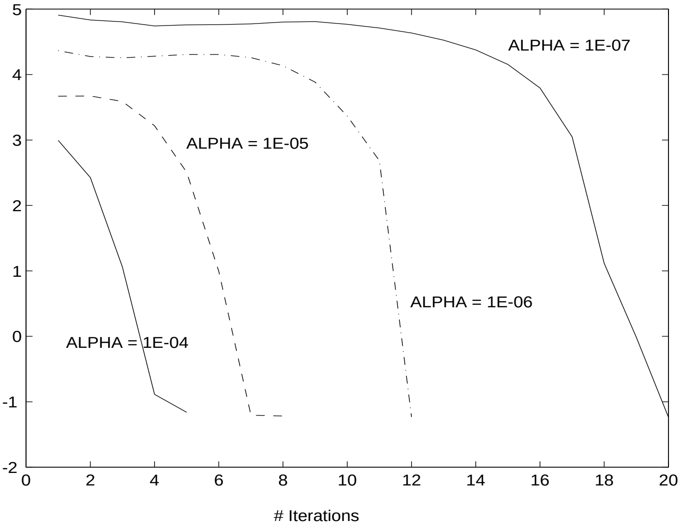


Fig. 5.2. *Influence of α on the behavior of L_c (logarithmic scale).*

See Table 5.4. The solution was obtained in 210 seconds.

The plot in Figure 5.2 shows the influence of α on the behavior of the Lagrangian function L_c.

We see that during the first iterations the augmented Lagrangian function does not decrease if α is too small.

However, if the initialization point is close enough to the solution, then this function decreases. We have tested initialization points different from b which were closer to the solution and obtained decrease of L_c. As an example we give in Table 5.5 the results for $\alpha = 10^{-10}$ with an initialization according to (2.1) but with u_o the solution for $\alpha = 10^{-5}$.

Note that the total number of iterations including the initialization with $\alpha = 10^{-5}$ to obtain the solution corresponding to $\alpha = 10^{-10}$ is equal to 18. If one computes the solution with initialization $u_o = b$, the number of iterations is 27 and L_c decreases

TABLE 5.5
Example 3(b): $\alpha = 10^{-10}$, u_o *given by the solution to* $(\mathcal{P})$ *for* $\alpha = 10^{-5}$.

Iteration	$\max(u_n - b)$	Size of $\mathcal{A}_n$	$J(y_n, u_n)$	$L_c(y_n, u_n, \lambda_n)$
1	1.6605e+03	1986	5.696032e-02	4.889158e+01
2	1.4741e+03	2034	5.750110e-02	2.948470e+01
3	1.1542e+03	2082	5.781067e-02	1.299992e+01
4	6.8931e+02	2130	5.793424e-02	2.631407e+00
5	1.6713e+02	2168	5.795024e-02	2.198494e-01
6	1.1931e+02	2172	5.795048e-02	1.276798e-01
7	7.0091e+01	2176	5.795058e-02	7.857522e-02
8	2.0618e+01	2180	5.795061e-02	5.958497e-02
9	0.0000e+00	2182	5.795061e-02	5.795061e-02
10	0.0000e+00	2182	5.795061e-02	5.795061e-02

after iteration 12. Thus a good initial guess can decrease the number of iterations to obtain the solution. This process was repeated successfully for smaller values of α up to $\alpha = 10^{-15}$ as well.

REFERENCES

[B] V. Barbu, *Analysis and Control of Non Linear Infinite Dimensional Systems*, Math. Sci. Engrg. 190, Academic Press, New York, 1993.

[BK] M. Bergounioux and K. Kunisch, *Augmented Lagrangian techniques for elliptic state constrained optimal control problems*, SIAM J. Control Optim. 35, (1997), pp. 1524–1543.

[BS] H. Brézis and G. Stampacchia, *Remarks on some fourth order variational inequalities*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 4 (1977), pp. 363–371.

[HT] M. Heinkenschloss and F. Tröltzsch, *Analysis of the Lagrange-SQP-Newton Method for the Control of a Phase Field Equation*, preprint, Virginia Polytechnic Institute, Blacksburg, VA, 1997.

[IK] K. Ito and K. Kunisch, *Augmented Lagrangian Formulation of Nonsmooth Convex Optimization in Hilbert Spaces*, in Control of Partial Differential Equations and Applications, E. Casas, ed., Lecture Notes in Pure and Appl. Math. 174, Marcel Dekker, New York, 1995, pp. 107–117.

[KeS] C.T. Kelley and E. Sachs, *Approximate quasi-Newton methods*, Math. Programming, 48 (1990), pp. 41–70.

[KuS] K. Kunisch and E. W. Sachs, *Reduced SQP methods for parameter identification problems*, SIAM J. Numer. Anal., 29 (1992), pp. 1793–1820.

[K] F.S. Kupfer, *An infinite-dimensional convergence theory for reduced SQP-methods in Hilbert spaces*, SIAM J. Optim., 6 (1996), pp. 126–163.

[Sch] K. Schittkowski, *On the convergence of a sequential quadratic programming method with an augmented Lagrangian line search function*, Math. Operationsforsch. Statist. Ser. Optim., 14 (1983), pp. 197–216.

[T] F. Tröltzsch, *An SQP-method for optimal control of a nonlinear heat equation*, Control Cybernet., 23 (1994), pp. 268–288.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 4, pp. 1195–1221

REGULARITY OF THE MINIMUM TIME FUNCTION AND MINIMUM ENERGY PROBLEMS: THE LINEAR CASE[*]

FAUSTO GOZZI[†] AND PAOLA LORETI[‡]

Abstract. We prove new results on the continuity properties and on reachable sets of the minimum time function associated with a linear dynamical system in a separable Hilbert space H. Part of the results are new also in the finite dimensional case. The regularity results are stated thanks to the result we obtain on the connection between the minimum time function and the minimum energy.

Key words. null controllability, minimum time function, minimum energy problems

AMS subject classifications. 49R20, 93B05, 93B06, 93C20

PII. S0363012996312763

1. Introduction.

This paper is devoted to the study of continuity properties of the minimum time function associated with a linear dynamical system in finite and infinite dimensions. We consider a separable Hilbert space H (the state space) with norm $|\cdot|$ and scalar product $\langle\cdot,\cdot\rangle$ and we consider a controlled dynamical system in H whose behavior is described by the following linear Cauchy problem in H:

$$\begin{cases} y'(t) = Ay(t) + Bu(t); \quad t > 0, \\ y(0) = x \in H, \end{cases} \tag{1}$$

where $A : D(A) \subset H \to H$ is a closed linear operator that generates a strongly continuous semigroup e^{tA} on H. Moreover, $u : [0, +\infty[\to U$ is a measurable function with values in another Hilbert space U (the space of control parameters) and $B : U \to H$ is a continuous linear operator. Given $u \in L^1_{loc}(0, +\infty; U)$, there exists a unique mild solution to (1) in $[0, +\infty)$ given by (see, e.g., Pazy [28])

$$y(t; x, u) = e^{tA}x + \int_0^t e^{(t-s)A}Bu(s)ds. \tag{2}$$

Now we assume that the control strategies $u(\cdot)$ are taken in a given set of admissibility $\mathcal{U}_{ad}$ and consider the following minimum time problem:

Given $x \in H$, find the minimal time needed to bring x to 0 *by applying an admissible control strategy.*

The function $T(x)$ giving the minimal time to bring x to 0 is set equal to $+\infty$ when it is not possible to steer x to 0 in finite time by applying admissible control strategies, and it is called the Bellman function of the system (1). Obviously, the function T will depend on the set $\mathcal{U}_{ad}$ of admissible control strategies. We will consider the case where $\mathcal{U}_{ad}$ is a ball centered at $x = 0$ of the space $L^p(0, +\infty; U)$, $p \in (1, +\infty]$ and

[*] Received by the editors December 2, 1996; accepted for publication (in revised form) June 5, 1998; published electronically May 21, 1999.
http://www.siam.org/journals/sicon/37-4/31276.html

[†] Dipartimento di Matematica, Università di Pisa, Via F. Buonarroti n.2, 56127 Pisa, Italy (gozzi@dm.unipi.it). Current address: Dipartimento di Matematica per le Decisioni, Universitá di Roma "La Sapienza," via del Castro Laurenziano 9, Roma, Italy.

[‡] Istituto per le Applicazioni del Calcolo "Mauro Picone," Consiglio Nazionale delle Ricerche, Viale del Policlinico 137, 00161 Roma, Italy. Current address: Dipartimento MeMoMat, Università degli Studi di Roma "La Sapienza," Via A. Scarpa, 16, 00161 Roma, Italy (loreti@dmmm.uniroma1.it).

call the associated Bellman function $T_p(x)$. The regularity properties of the Bellman function give important information on the system (1) and have been studied in many papers when the dimension of H is finite and $p = +\infty$.

Petrov [29] studied more general classes of admissible controls and stated that a necessary and sufficient condition for the continuity of the minimal time function is the local null controllability. In our context this is equivalent to the so-called Kalman condition, i.e., that the matrix $[B, AB, ..., A^{n-1}B]$ is full rank.

Petrov also describes necessary and sufficient conditions under which the Bellman function is Lipschitz continuous in a neighborhood of the origin. Other papers related to Lipschitz continuity of the Bellman function $T_\infty(x)$ are, e.g., Cannarsa and Sinestrari [8, 9], which study various properties of the minimum time function in the case when the state equation is nonlinear, and Hajek [20], which studies the case of linear systems.

On the other hand, in many cases the minimal time function is neither differentiable nor Lipschitz continuous. The Hölder continuity was investigated by Liverovskii [23, 24], Petrenko [30], Ranguin [32, 33], and by Gyurkovics [19] who showed that, when the system (1) satisfies the Kalman condition, the Hölder exponent of T_∞ is related to the minimum number $k \leq n-1$ such that the matrix $[B, AB, ..., A^k B]$ has full rank. We refer moreover to Bacciotti [1, sections 25, 33] and to Bianchini and Stefani [6] and the references therein for a discussion of the relationship between the Hölder continuity of the minimum time function and geometrical properties of the system also in the nonlinear case.

Other papers related to regularity properties of $T_\infty(x)$ in finite dimension are Bardi and Soravia [5] concerning Hölder continuity in the nonlinear case; Rampazzo and Sartori [31] concerning the case of L^1 controls; Bardi [4] concerning the dynamic programming equation for T_∞; and Conti [10]. See also Hermes and Lasalle; Lee and Markus; and Li and Yong [21, 22, 25] for a wider exposition of the argument.

The minimum time problem has been studied also in the case when the dimension of H is infinite and $p = +\infty$. Carja [12] proved the continuity of the Bellman function T_∞ and various properties of the reachable set when the system is linear and null controllable, while in Carja [14] the linear case with $B = I$ is considered. In Carja [13, 15] Hölder continuity results of T_∞ in the linear case are proved. In these last two papers Hölder continuity results for T_p, $p < +\infty$ are also proved; however, part of the method used to prove these results seems to be incorrect (see our section 3.2). We also recall Tauraso [38] which contains continuity results for the minimum time function of linear systems such as (1) in the case of $p = 2$ in connection with controllability properties of the system.

Other papers on the minimum time problem in infinite dimension are, e.g., Fattorini [17] and Krabs [26, 27], which study existence and properties of optimal controls that we will use in this paper. We recall also the papers of Barbu [2, 3] in which the dynamic programming equation and the maximum principle in some infinite dimensional case are studied.

The aim of this paper is to study the continuity properties of the Bellman function T_p, $p \in (1, +\infty]$, in the case when the state space H is a separable Hilbert space by generalizing the results of the papers mentioned above. We point out that in finite dimension the natural approach (used in the papers mentioned above) is to connect the continuity properties of T_p in a neighborhood of $x = 0$ with the minimum number $k \leq n-1$ such that the matrix $[B, AB, ..., A^k B]$ has full rank. Since in infinite dimension this is not possible, we use a different approach: we prove (see Theorem

4.1) that the function T_p is implicitly defined by the minimum energy function (that we will introduce later; see also section 2); we prove new regularity results for the minimum energy function and we use them to prove various continuity properties of T_p.

The connection between minimum time problems and minimum energy problems has been observed in [38], where the case $p = 2$ is studied and a lower semicontinuity result for T_2 is proved, and in [15, Remark 3.1], where it is used to prove Hölder continuity of T_∞ in some particular cases. However, in both cases, a precise formulation of this connection and of its consequences is not given. Here the mentioned relationship is established in a precise way for $p \in (1, +\infty]$ (see Theorem 4.1). To well exploit this connection we study the regularity properties of the minimum energy. Some results of this kind are available in the literature (see section 3.1) but we were able to prove more. Our new results on the minimum energy are interesting also in themselves and we gather them in section 3.2, Theorem 3.7. Then, using these new results we get an estimate of the growth rate of $T_p(x)$ in a neighborhood of $x = 0$ (see Theorem 4.6) which can be seen as an estimate of the modulus of continuity (and in some cases of the Hölder exponent) near this point. This estimate is then exploited in a different way when $p \in (1, +\infty)$ and when $p = +\infty$.

If $p = +\infty$ via the dynamic programming principle, the estimate of Theorem 4.6 can be transferred (locally) to the whole reachable sets obtaining local uniform continuity, an estimate of the local modulus and, in somes case, of the local Hölder exponent. The fact that T_∞ is locally uniformly continuous is already shown in [15] where some estimates of the modulus of continuity are given, too. Here the estimate proved in Theorem 4.6 gives a way to estimate the modulus of continuity of T_∞ in more general cases by using estimates for the minimum energy as $t \searrow 0$. In general it is not easy to find such estimates, but in some cases it can be done, as shown in section 5.

If $1 \leq p < +\infty$, the dynamic programming principle does not hold, in general, as shown in Example 5.14. If $p \in (1, +\infty)$, by using directly Theorem 4.1, we prove that the function T_p is continuous (Theorem 4.8) and that the reachable sets enjoy useful properties (Theorem 4.10). Moreover, by adding a new variable to the problem (i.e., the radius of the ball of admissible strategies) we can also prove a version of the dynamic programming principle which allows us again to transfer (with a different method) the estimate of Theorem 4.6 to the whole reachable sets obtaining local uniform continuity, an estimate of the local modulus and, in some cases, of the local Hölder exponent. In particular, in the finite dimensional case, using results of Seidman and Yong [35, 37] (see Theorem 3.6) that give sharp estimates for the minimum energy as $t \searrow 0$, we obtain precise estimates of the local Hölder exponent of T_p that seem to be completely new.

We remark that the case $p = 1$ is the most difficult since the space L^1 is not reflexive nor a dual and so we cannot use the characterization of controllability given in section 2. This case is not treated in this paper (see [31] for results in this direction).

Finally we recall that the behavior of the minimum energy function, in the case $p = 2$, is related to the regularity properties of the stochastic diffusion process that solves (1), where the control u is substituted by a white noise $\dot{W}$. In particular, the Kolmogorov equation associated with the process y enjoys good regularity properties depending on the behavior of the singularity of the minimum energy as $t \searrow 0$ (see [16, Chapter 9] for an extensive treatment of the subject). So, in this paper, by proving Theorem 4.1 we have indirectly shown a connection between the continuity properties of $T_2(x)$ in a neighborhood of $x = 0$ and the regularity properties of the solutions of suitable second order partial differential equations (for results in this direction in the finite dimensional case see, e.g., [18]).

The plan of the paper is the following. First we give some preliminaries and the statement of the problem in section 2. Then we recall known results about null controllability and minimum energy problems in section 3.1, while in section 3.2 we give new results about continuity and monotonicity of the minimum energy that are used in the rest of the paper. Section 4 is devoted to a statement and proof of the main results, while in section 5 we discuss the uniform continuity (and in some cases the Hölder continuity) of T_p and give some examples.

2. Preliminaries.

This section is devoted to introducing the basic concepts that will be used in the paper, i.e., the null controllability, the minimum time problem, and the minimum energy problem. Consider the control system (1), where we assume the following.

HYPOTHESIS 2.1.

1. *H and U are real separable Hilbert spaces.*
2. *$A : D(A) \subset H \to H$ is a closed linear operator that generates a C_0 semigroup e^{tA} on H satisfying $\|e^{tA}\| \leq Me^{\omega t}$ for given $M \geq 1$ and $\omega \in \mathbb{R}$.*
3. *$u : [0, +\infty) \to U$ is a locally integrable function.*
4. *$B : U \to H$ is a continuous linear operator.*

The unique mild solution (2) is continuous and can be written as

$$y(t; x, u) = e^{tA}x + \mathcal{L}_t u, \tag{3}$$

where the operator $\mathcal{L}_t$ is defined as

$$\mathcal{L}_t : L^p(0, +\infty; U) \to H, \qquad \mathcal{L}_t u = \int_0^t e^{(t-s)A}Bu(s)ds. \tag{4}$$

We will assume that the control strategies belong to a given set of admissibility $\mathcal{U}_{ad} \subset L^1_{loc}(0, +\infty; U)$ which will vary depending on the context. We begin by taking, for $p \in [1, +\infty]$, $\mathcal{U}_{ad} = \mathcal{U}_p \overset{def}{=} L^p(0, +\infty; U)$ and by considering the following null controllability problem:

Given $T > 0$ and $x \in H$, find an admissible control strategy u such that $y(T; x, u) = 0$.

In the literature we refer to [39] for the various concepts of controllability, in particular the null controllability. We now give the definition of p-null controllability.

DEFINITION 2.2. *Let $T > 0$, $x \in H$, and $p \in [1, +\infty]$. If there exists a strategy $u \in \mathcal{U}_p$ such that $y(T; x, u) = 0$, then the point x is said to be p-null controllable in time T. If this happens for every $x \in H$, then the system* (1) *is said to be p-null controllable in time T. If this happens for every $T > 0$ and $x \in H$, then the control system* (1) *is said to be simply p-null controllable.*

DEFINITION 2.3. *Given $T > 0$, $x \in H$, and $p \in [1, +\infty]$ we will denote by $\mathcal{M}(p, T, x)$ the (possibly empty) set of control strategies $u \in L^p(0, +\infty; U)$ such that $y(T; x, u) = 0$.*

Remark 2.4. The set $\mathcal{M}(p, T, x)$ is convex and closed and is nonempty if and only if x is p-null controllable in time T. Moreover, by (3) and (4) it follows that x is p-null controllable in time T if and only if $e^{TA}x \in \mathcal{L}_T(\mathcal{U}_p)$. Since $\mathcal{L}_T(\mathcal{U}_q) \subset \mathcal{L}_T(\mathcal{U}_p)$ for $q > p$, then, if a point x is q-null controllable in time T, it is also p-null controllable in time T for every $p < q$.

Finally, the set $\mathcal{L}_t(\mathcal{U}_p)$ is increasing in the variable t. This remain true if we consider, for any $\rho > 0$, the set $\mathcal{L}_t\left(\{u \in \mathcal{U}_p, \|u\|_p \leq \rho\}\right)$.

We now take by $\mathcal{U}_{ad}$ the following subset of controls:

$$\mathcal{U}_{ad} = \mathcal{U}_{p,\rho} = \{u \in \mathcal{U}_p : |u|_p \leq \rho\},$$

where ρ is a positive number. We denote by $\mathcal{R}_{p,\rho}(t)$ the set of states controllable to the rest within time t by some controls in $\mathcal{U}_{p,\rho}$. More precisely,

$$\mathcal{R}_{p,\rho}(t) = \{x \in H : \exists u \in \mathcal{U}_{p,\rho}\, s.t.\, y(t,x,u) = 0\}.$$

We observe that since H is a reflexive space, then $\mathcal{R}_{p,\rho}(t)$ is a closed set. We set

$$\mathcal{R}_{p,\rho} = \bigcup_{t \geq 0} \mathcal{R}_{p,\rho}(t)$$

and we define the minimum time function as

$$T_{p,\rho}(x) = \inf\{t \geq 0 : x \in \mathcal{R}_{p,\rho}(t)\}, \text{if } x \in \mathcal{R}_{p,\rho} \tag{5}$$

and $T_{p,\rho}(x) = +\infty$ otherwise. In some cases we will take for simplicity $\rho = 1$ and we will write T_p instead of $T_{p,1}$ and $\mathcal{R}_p$ instead of $\mathcal{R}_{p,1}$.

We recall that the continuity of the minimal time function in the origin is equivalent to the p-null controllability of system (1) as the following result, proved in [15], shows.

THEOREM 2.5. *Assume that Hypothesis* 2.1 *holds and let* $p \in (1, +\infty]$, $\rho > 0$. *Then the system* (1) *is p-null controllable if and only if the corresponding minimal time function* $T_{p,\rho}$ *is continuous at the origin.*

This result drives us to study only the case when the system (1) is p-null controllable. We end the section with the definition of the minimum energy problem.

DEFINITION 2.6. *Given* $T > 0$, $x \in H$, *and* $p \in [1, +\infty]$, *define the p-energy of a control strategy* $u \in \mathcal{M}(p, T, x)$ *as the quantity*

$$\mathcal{E}_p(u) = \left(\int_0^{+\infty} |u(s)|_U^p ds\right)^{\frac{1}{p}}, \qquad p \in [1, +\infty),$$

$$\mathcal{E}_\infty(u) = \sup_{t \geq 0} |u(t)|_U,$$

and the minimum p-energy to bring x *to* 0 *as*

$$\mathcal{E}_p^*(T, x) \overset{def}{=} \inf_{u \in \mathcal{M}(p,T,x)} \mathcal{E}_p(u).$$

The minimum p-energy problem is

find the minimum p-energy needed to bring x *to* 0 *in time* T.

We observe that the second extremum of the integral can be taken as T, since the control $u(s) = 0$ for $s > T$ is admissible and it does not modify the value of the minimum.

3. Null controllability and minimum energy.

3.1. Known results. In this section we recall some known results about null controllability and minimum energy that we will use in the rest of the paper. We start with the following if and only if conditions for the 2-null controllability of the system (1) (see, e.g., [39, Part IV, Chapter 1]).

THEOREM 3.1. *Assume Hypothesis* 2.1.

(i) *A point* $x \in H$ *is* 2*-null controllable in time* $T > 0$ *if and only if* $e^{TA}x \in Q_T^{\frac{1}{2}}(H)$, *where* $Q_T = \int_0^T e^{sA}BB^*e^{sA^*}ds$.

(ii) *The system* (1) *is* 2*-null controllable if and only if* $e^{TA}(H) \subset Q_T^{\frac{1}{2}}(H)$ $\forall T > 0$.

(iii) *The condition* (ii) *above is equivalent to*

$$\text{For every } T > 0 \text{ there exists a constant } C(T) > 0$$
$$\text{such that } |e^{TA^*}x| \leq C(T)|Q_T^{1/2}x| \ \forall x \in H.$$

In this case the operator $\Gamma(T) = Q_T^{-1/2}e^{TA}$ *is well defined and bounded on* H *and* $\|\Gamma(T)\|_{\mathcal{L}(H)}$ *is the infimum of the constants* $C(T)$ *such that condition* (iii) *holds.*

(iv) *If* $H = \mathbb{R}^n$, *then the system* (1) *is* 2*-null controllable if and only if*

$$\text{rank}\,[B, AB, A^2B, ..., A^{n-1}B] = n. \tag{6}$$

Remark 3.2. In the finite dimensional case all concepts of p-null controllability coincide with (6) which is the well-known Kalman condition.

If $p \in [1, +\infty]$, then (i), (ii) of the above theorem still hold true with the operator $\mathcal{L}_T : \mathcal{U}_p \to H$ in place of $Q_T^{\frac{1}{2}}$. Moreover, if $p \in (1, +\infty)$, then condition (iii) is true in the following form (see, e.g., [11, Theorem 2.2]):

The system (1) *is* p*-null controllable if and only if for every* $T > 0$ *there exists a constant* $C(T) > 0$ *such that*

$$|e^{TA^*}x| \leq C(T)|\mathcal{L}_T^*x| = C(T)\left(\int_0^T |B^*e^{(T-s)A^*}x|^q ds\right)^{\frac{1}{q}} \qquad \forall x \in H, \tag{7}$$

where $\frac{1}{q} + \frac{1}{p} = 1$. As in the case $p = 2$ the smallest $C(T)$ is decreasing with respect to T and blows up as $T \searrow 0$. If $p = 1, +\infty$ the equivalence above is still true when A generates a group but in general is false. In this case we have only that p-null controllability implies (7) (see, e.g., [11]).

We now give a result about existence of optimal controls that, in the case $p = 2$, gives an explicit formula for the optimal strategy. The proof can be found, e.g., in [39, Part IV, Chapter 1] for the case $p = 2$. The case $p \neq 2$ can be obtained exactly by repeating the same argument and we omit it. The case $p = +\infty$ can be proved using the same argument of [17].

THEOREM 3.3. *Let* $p \in (1, +\infty)$. *Assume Hypothesis* 2.1 *and let the system* (1) *be* p*-null controllable. Then the problem*

$$\min\left\{\int_0^T |u(s)|^p ds, u \in L^p(0, +\infty; U) : \ y(T; x, u) = 0\right\}$$

admits only one optimal solution which, in the case $p = 2$, *is given by*

$$u_{T,x}^*(s) = -\overline{B^*e^{(T-s)A^*}Q_T^{-\frac{1}{2}}}\Gamma(T)x$$

and in this case the value of the minimum is $|\Gamma(T)x|^2$*;* $\overline{B^* e^{(T-s)A^*} Q_T^{-\frac{1}{2}}}$ *denotes the closure of the operator* $B^* e^{(T-s)A^*} Q_T^{-\frac{1}{2}}$.

Let $p = +\infty$*. Assume Hypothesis* 2.1 *and let the system* (1) *be* ∞*-null controllable. Then the problem*

$$\min\{||u||_\infty, u \in L^\infty(0,+\infty;U) : \; y(T;x,u) = 0\}$$

admits at least one optimal solution.

Remark 3.4. The operator $B^* e^{(T-s)A^*} Q_T^{-\frac{1}{2}}$ is not a priori well defined on all H. However, its closure is well defined on H due to the 2-null controllability assumption (see [39, Part IV, Chapter 1]).

Moreover, we recall that if $p = 1$, in general an optimal control strategy for the minimum energy problem does not exists. For example, one can take the one dimensional case when $A = a \neq 0$ and $B = 1$.

The behavior of $\mathcal{E}_p^*(t,x)$ when $t \to 0^+$ plays an important role in the rest of the paper. We recall the following result, which is immediately provable taking the control $u(s) = -(1/T)B^{-1}e^{sA}x$. (See [16, Appendix B] for the case $p = 2$.)

PROPOSITION 3.5. *Assume that Hypothesis* 2.1 *holds and that B is onto. Then, for every* $p \in [1,+\infty]$ *the system* (1) *is p-null controllable and for every fixed* T_0 *there exists a suitable constant* $C(T_0) > 0$ *such that*

$$\mathcal{E}_p^*(T,x) \leq C(T_0)|x|T^{-1+1/p}, \qquad T \in (0,T_0].$$

In finite dimension we have the following sharp result given by Seidman [35] and by Seidman and Yong [37].

THEOREM 3.6. *If* $H = \mathbb{R}^n$*,* $p \in [1,+\infty]$*, then, for almost every* $x \in H$ *(setting* $1/p = 0$ *when* $p = +\infty$*)*

$$\mathcal{E}_p^*(t,x) \sim t^{-k-1+1/p}|x| \qquad \textit{as } t \to 0^+,$$

where k is the first integer such that rank $[B, AB, \dots, A^k B] = n$.

This theorem shows that the singularity of $\mathcal{E}_p^*(t,x)$ as $t \to 0^+$ can have only a finite range of behaviors. When the dimension of H is infinite this is not true as can be deduced by the example given in [16, pp. 275–277] which shows that, in the case $p = 2$ for every $\beta \geq 1$, we can find diagonal operators A, B such that $\|\Gamma(t)\| = o(t^{\frac{\beta}{2}})$ as $t \to 0$.

3.2. New results. The following result about the regularity of the minimal energy is interesting in itself and will be a key tool for proving properties of the minimum time function and of reachable sets. For simplicity we will write L^p for $L^p(0,+\infty;U)$.

THEOREM 3.7. *Let* $p \in (1,+\infty]$*. Assume that Hypothesis* 2.1 *is satisfied and that the system is p-null controllable. Then*

(i) *the function* $\mathcal{E}_p^*(t,\cdot)$ *is homogeneus of degree* 1 *for every* $t > 0$*;*

(ii) *the function* $\mathcal{E}_p^*(\cdot,x)$ *is strictly decreasing for every* $x \in H$*;*

(iii) *if* $p \in (1,+\infty)$*, then the function* $\mathcal{E}_p^*$ *is continuous. Moreover, for every* $\varepsilon > 0$ *there exists* C_ε *such that*

$$|\mathcal{E}_p^*(t,x_1) - \mathcal{E}_p^*(t,x_2)| \leq C_\varepsilon |x_1 - x_2| \tag{8}$$

for every $t \geq \varepsilon$, $x_1, x_2 \in H$, *and*

$$|\mathcal{E}_p^*(t_1,x) - \mathcal{E}_p^*(t_2,x)| \leq C_\varepsilon \left[|x|(t_2-t_1) + \left| x - e^{(t_2-t_1)A}x \right| \right] \tag{9}$$

for every $x \in H$, $t_2 > t_1 > \varepsilon$.

To prove the above theorem, we need first a useful representation lemma that uses a minimax theorem to give a different representation of the minimum energy function.

LEMMA 3.8. *Given* $p \in (1, +\infty)$ *consider the function*

$$L_p : [0,+\infty) \times H \times L^p \times H \mapsto \mathbb{R}, \qquad L_p(t,x;u,y) = \|u\|_p^p + \langle y, \mathcal{L}_t u + e^{tA}x \rangle.$$

Then, for every $(t,x) \in [0,+\infty) \times H$ *we have*

$$\mathcal{E}_p^*(t,x)^p = \inf_{u \in L^p} \sup_{y \in H} L_p(t,x;u,y) = \sup_{y \in H} \inf_{u \in L^p} L_p(t,x;u,y)$$

and

$$\mathcal{E}_p^*(t,x)^p = \sup_{y \in H} \left\{ \langle y, e^{tA}x \rangle - c_p \|\mathcal{L}_t^* x\|_{L^q}^q \right\}, \tag{10}$$

where $1/p + 1/q = 1$ *and* $c_p = (1/q)[1/p]^{q/p}$.

Proof of Lemma 3.8. First we observe that

$$\sup_{y \in H} L_p(t,x;u,y) = \begin{cases} \|u\|_p^p & \text{if } e^{tA}x + \mathcal{L}_t u = 0, \\ 0 & \text{otherwise} \end{cases}$$

so that

$$\inf_{u \in L^p} \sup_{y \in H} L_p(t,x;u,y) = \inf_{u \in \mathcal{M}(p,t,x)} \|u\|_p^p = \mathcal{E}_p^*(t,x)^p.$$

The proof of the other equality

$$\mathcal{E}_p^*(t,x)^p = \sup_{y \in H} \inf_{u \in L^p} L_p(t,x;u,y)$$

is much more difficult. Let us call, for brevity,

$$G_1(t,x) = \inf_{u \in L^p} \sup_{y \in H} L_p(t,x;u,y) = \mathcal{E}_p^*(t,x)^p$$

and

$$G_2(t,x) = \sup_{y \in H} \inf_{u \in L^p} L_p(t,x;u,y).$$

We want to prove that $G_1 = G_2$. We begin by observing that, by easy verification,

$$\inf_{u \in L^p} \sup_{y \in H} L_p(t,x;u,y) \geq \sup_{y \in H} \inf_{u \in L^p} L_p(t,x;u,y),$$

so $G_1 \geq G_2$ and we have only to prove the reverse inequality. Since the function L_p does not satisfy usual assumptions of minimax theorems, we consider for $\varepsilon > 0$ the approximating function $L_{p,\varepsilon}(t,x;u,y) = L_p(t,x;u,y) - \varepsilon |x|^2$. It can be easily verified that the function $L_{p,\varepsilon}$ satisfies assumptions of [7, Proposition 1] (see also [38, Chapter IV]) and so there exists a point $(u_\varepsilon, y_\varepsilon)$ (a so-called "saddle point") such that

$$\max_{y\in H}\inf_{u\in L^p(0,+\infty;U)} L_{p,\varepsilon}(t,x;u,y) = L_{p,\varepsilon}(t,x;u_\varepsilon,y_\varepsilon) = \min_{u\in L^p}\sup_{y\in H} L_{p,\varepsilon}(t,x;u,y)$$

which gives, for every $u \in L^p$ and $y \in H$,

$$L_p(t,x;u_\varepsilon,y) - \varepsilon|y|^2 \le L_p(t,x;u_\varepsilon,y_\varepsilon) - \varepsilon|y_\varepsilon|^2 \le L_p(t,x;u,y_\varepsilon) - \varepsilon|y_\varepsilon|^2$$

so that

$$\begin{aligned} L_p(t,x;u_\varepsilon,y_\varepsilon) &\le \inf_{u\in L^p} L_p(t,x;u,y_\varepsilon) \le \sup_{y\in H}\inf_{u\in L^p} L_p(t,x;u,y_\varepsilon) \\ &= G_2(t,x) \le \mathcal{E}_p^*(t,x)^p < +\infty. \end{aligned}$$

The above two inequalities give

$$0 \le \|u_\varepsilon\|_p^p = L_p(t,x;u_\varepsilon,0) \le L_p(t,x;u_\varepsilon,y_\varepsilon) - \varepsilon|y_\varepsilon|^2 \le G_2(t,x)$$

so that the set $\{u_\varepsilon;\ \varepsilon > 0\}$ is bounded and so there exists $u_0 \in L^p$ and a sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$ converging to 0 as $n \to +\infty$ such that

$$u_{\varepsilon_n} \longrightarrow u_0 \quad \text{weakly in } L^p.$$

Passing to the limit on this sequence we then obtain

$$L_p(t,x;u_0,y) \le \liminf_{n\to+\infty} L_p(t,x;u_{\varepsilon_n},y) \le \liminf_{n\to+\infty} [L_p(t,x;u_{\varepsilon_n},y) - \varepsilon_n|y|^2]$$

$$\le \liminf_{n\to+\infty} [L_p(t,x;u_{\varepsilon_n},y_{\varepsilon_n}) - \varepsilon_n|y_{\varepsilon_n}|^2] \le \liminf_{n\to+\infty} L_p(t,x;u_{\varepsilon_n},y_{\varepsilon_n}) \le G_2(t,x)$$

so that

$$G_1(t,x) \le \inf_{u\in L^p}\sup_{y\in H} L_p(t,x;u,y) \le \sup_{y\in H} L_p(t,x;u_0,y) \le G_2(t,x)$$

which gives the claim $G_1 = G_2$.

Finally, to prove the last equality of the claim it is enough to observe that, by standard calculations, we get

$$\sup_{y\in H}\inf_{u\in L^p} L_p(t,x;u,y)$$

$$= \sup_{y\in H}\left\{\langle y, e^{tA}x\rangle + \inf_{u\in L^p}\left[\|u\|_p^p + \langle y, \mathcal{L}_t u\rangle\right]\right\}.$$

Now, since $\mathcal{L}_t^* : H \mapsto L^q$, $[\mathcal{L}_t^* y](s) = B^* e^{(t-s)A^*} y$, we have

$$\|u\|_p^p + \langle y, \mathcal{L}_t u\rangle = \|u\|_p^p + \langle \mathcal{L}_t^* y, u\rangle_{\langle L^q, L^p\rangle}$$

and

$$\inf_{u\in L^p}\left[\|u\|_p^p + \langle y, \mathcal{L}_t u\rangle\right] = -\sup_{u\in L^p}\left[-\|u\|_p^p - \langle \mathcal{L}_t^* y, u\rangle\right]$$

which is the convex conjugate function (see, e.g., [34, p. 104]) of $\|u\|_p^p$ calculated at $-\mathcal{L}_t^* y$ which, by standard calculations, is equal to $c_p \|\mathcal{L}_t^* y\|_q^q$, where $c_p = (1/q)[1/p]^{q/p}$. It follows that

$$\inf_{u \in L^p} \left[\|u\|_p^p + \langle y, \mathcal{L}_t u \rangle \right] = -c_p \|\mathcal{L}_t^* y\|_q^q$$

which gives the claim. □

Remark 3.9.

(i) We observe that, in general, we cannot say that the map

$$H \mapsto \mathbb{R}, \qquad y \mapsto \langle y, e^{tA} x \rangle - c_p \|\mathcal{L}_t^* x\|_{L^q}^q$$

has a maximum point. However, if the system (1) is exactly controllable, then there exists a unique maximum point (see, e.g., [38, Chapter IV]).

(ii) From the proof of Theorem 3.7 and from Remark 3.2 we can easily deduce that a point $x \in H$ is p-null controllable in time $t > 0$ if and only if

$$G_2(t, x) = \sup_{y \in H} \left\{ \langle y, e^{tA} x \rangle - c_p \|\mathcal{L}_t^* x\|_{L^q}^q \right\} < +\infty.$$

We are now ready to give the proof of Theorem 3.7.

Proof of Theorem 3.7. (i) It is a straightforward consequence of the definition of minimum energy and of the observation that

$$\mathcal{M}(p, t, \lambda x) = \lambda \mathcal{M}(p, t, x).$$

(ii) Given $p \in (1, +\infty]$, $t_1 \geq 0$, and $x \in H$ it is clear that, if u_0 is the minimum p-energy control at (t_1, x), then for every $t_2 > t_1$ the control $\overline{u}$ defined as $\overline{u}(r) = u_0(r)$ for $r \in [0, t_1]$ and $\overline{u}(r) = 0$ for $r \in (t_1, t_2]$ belongs to $\mathcal{M}(p, t_2, x)$. This implies, in particular, that $\mathcal{E}_p^*(t_1, x) \geq \mathcal{E}_p^*(t_2, x)$ for $t_2 > t_1$.

We now want to prove the strict inequality starting from the case $p \in (1, +\infty)$. In this case we in fact have existence and uniqueness of the minimum p-energy control, so to prove the strict inequality it is enough to prove that the above control $\overline{u}$ is not optimal for $t_2 > t_1$. To this end we observe that, for $p \in (1, +\infty)$, $t \geq 0$, $x \in H$, and $u \in \mathcal{M}(p, t, x)$, we can write

$$\mathcal{M}(p, t, x) = u + \text{Ker } \mathcal{L}_t.$$

In particular, the control of minimum p-energy u^* is such that, for any given $w \in \text{Ker } \mathcal{L}_t$,

$$\|u^*\|_{L^p} = \min_{\lambda \in \mathbb{R}} \|u^* + \lambda w\|_{L^p}$$

so that, by differentiating, we get

$$\langle |u^*|^{p-2} u^*, w \rangle_{\langle L^q(0,t;U), L^p(0,t;U) \rangle} = 0.$$

Suppose now by contradiction that the control $\overline{u}$ defined above is optimal at time $t_2 > t_1$. Then, for every $w \in \text{Ker } \mathcal{L}_{t_2}$, we have

$$\langle |\overline{u}|^{p-2} \overline{u}, w \rangle_{\langle L^q(0,t_2;U), L^p(0,t_2;U) \rangle} = 0.$$

Choosing $u_1 \in L^p(0,t_2-t_1;U)$ such that

$$\mathcal{L}_{t_2-t_1}u_1 = e^{(t_2-t_1)A}\mathcal{L}_{t_1}u_0$$

(this is always possible thanks to the p-null controllability of the system (1)) the control w_1 defined as $w_1(s)=u_0(s)$ for $s\in[0,t_1]$ and $w_1(s)=u_1(s-t_1)$ for $s\in(t_1,t_2]$ belongs to Ker $\mathcal{L}_{t_2}$ and is such that

$$\langle|\overline{u}|^{p-2}\overline{u},w\rangle_{\langle L^q(0,t_2;U),L^p(0,t_2;U)\rangle} = \|u_0\|_{L^p(0,t_1;U)} > 0,$$

which is a contradiction.

Consider now the case $p=+\infty$. In this case we do not have uniqueness of the optimal control, in general, so the above argument is not a straightforward application. As before, let $0<t_1<t_2$, $x\in H$, and let u_0 be the minimum ∞-energy control at (t_1,x). Consider the control $u_\lambda\in L^\infty(0,t_1;U)$ defined as $u_\lambda(r)=(1-\lambda)u_0(r)$ for $r\in[0,t_1]$. Then $y(t_1;xu_\lambda)=\lambda e^{t_1A}x$. Now, from the ∞-null controllability assumption we know that $e^{(t_2-t_1)A}(H)\subset\mathcal{L}_{t_2-t_1}(\mathcal{U}_\infty)$ so that (see [16, Appendix B]) for every $\varepsilon>0$ there exists $\delta>0$ such that

$$e^{(t_2-t_1)A}(\{|x|<\delta\})\subset\mathcal{L}_{t_2-t_1}(\{\|u\|_\infty<\varepsilon\}).$$

Now we choose $0<\varepsilon<\|u_0\|_\infty$, $\lambda>0$ such that

$$\lambda|e^{t_1A}x|\le\delta$$

and $\hat{u}$ such that $\|\hat{u}\|_\infty<\varepsilon$, $\mathcal{L}_{t_2-t_1}\hat{u}=e^{(t_2-t_1)A}\left[\lambda e^{t_1A}x\right]$. Then the control $u_1\in L^p(0,t_2;U)$ defined as $u_1(s)=\lambda u_0(s)$ for $s\in[0,t_1]$ and $u_1(s)=\hat{u}(s-t_1)$ for $s\in(t_1,t_2]$ belong to $\mathcal{M}(\infty,t_2,x)$ and $\|u_1\|_\infty<\|u_0\|_\infty$, the claim follows.

(iii) We use the representation given in (10)

$$\mathcal{E}_p^*(t,x)^p=\sup_{y\in H}\left\{\langle y,e^{tA}x\rangle-c_p\|\mathcal{L}_t^*y\|_{L^q}^q\right\}.$$

Using this, we have for brevity $G(t,x)=\mathcal{E}_p^*(t,x)^p$, and taking $x_1\neq x_2$, with $G(t,x_1)>G(t,x_2)$,

$$\begin{aligned}G(t,x_1)-G(t,x_2)=&\sup\nolimits_{y\in H}\left\{\langle y,e^{tA}x_1\rangle-c_p\|\mathcal{L}_t^*y\|_{L^q}^q\right\}\\&-\sup\nolimits_{y\in H}\left\{\langle y,e^{tA}x_2\rangle-c_p\|\mathcal{L}_t^*y\|_{L^q}^q\right\}\end{aligned}$$

which, taking y_ε such that

$$G(t,x_1)\le\varepsilon+\langle y_\varepsilon,e^{tA}x_1\rangle-c_p\|\mathcal{L}_t^*y_\varepsilon\|_{L^q}^q,$$

implies

$$G(t,x_1)-G(t,x_2)\le\varepsilon+\langle e^{tA^*}y_\varepsilon,x_1-x_2\rangle\le\varepsilon+|e^{tA^*}y_\varepsilon||x_1-x_2|. \tag{11}$$

By the p-null controllability assumption and by Remark 3.2 we have

$$|e^{tA^*}y_\varepsilon|\le C(t)|\mathcal{L}_t^*y_\varepsilon|_{L^q} \tag{12}$$

so that, taking $\varepsilon\le G(t,x_1)$ (if $x_1=0$ we take the reverse inequality) we have

$$0\le\langle e^{tA^*}y_\varepsilon,x_1\rangle-c_p\|\mathcal{L}_t^*y_\varepsilon\|_{L^q}^q\le|e^{tA^*}y_\varepsilon||x_1|-c_p\|\mathcal{L}_t^*y_\varepsilon\|_{L^q}^q$$

$$\leq C(t)|\mathcal{L}_t^* y_\varepsilon|_{L^q}|x_1| - c_p\|\mathcal{L}_t^* y_\varepsilon\|_{L^q}^q$$

and, by Young's inequality, for any $\sigma > 0$ there exists $C_\sigma > 0$ such that

$$0 \leq \sigma|\mathcal{L}_t^* y_\varepsilon|_{L^q}^q + C_\sigma[C(t)|x_1|]^p - c_p|\mathcal{L}_t^* y_\varepsilon|_{L^q}^q$$

which finally gives, taking $\sigma = c_p/2$,

$$c_p/2|\mathcal{L}_t^* y_\varepsilon|_{L^q}^q \leq C_{c_p/2}C(t)^p|x_1|^p \tag{13}$$

and then, by (12),

$$|e^{tA^*}y_\varepsilon| \leq C(t)\left[\frac{2}{c_p}C_{c_p/2}C(t)^p|x_1|^p\right]^{1/q} = \left[\frac{2}{c_p}\right]^{1/q} C_{c_p/2}^{1/q}C(t)^p|x_1|^{p-1}.$$

By substituting the last inequality into (11) we get, taking also ε sufficiently small,

$$G(t,x_1)-G(t,x_2) \leq \varepsilon+\left[\frac{2}{c_p}\right]^{1/q} C_{c_p/2}^{1/q}C(t)^p|x_1|^{p-1}|x_1-x_2| \leq C_1(p)C(t)^p|x_1|^{p-1}|x_1-x_2|, \tag{14}$$

where $C_1(p)$ is a suitable constant depending only on p. Now we observe that, given $a > b \geq 0$ and $p \in (1,+\infty)$, we have

$$a - b \leq a^{1-p}\left[a^p - b^p\right]$$

so that, in our case,

$$\begin{aligned} \mathcal{E}_p^*(t,x_1) - \mathcal{E}_p^*(t,x_2) &\leq \mathcal{E}_p^*(t,x_1)^{1-p}\left[G(t,x_1) - G(t,x_2)\right] \\ &= G(t,x_1)^{-1/q}\left[G(t,x_1) - G(t,x_2)\right]. \end{aligned} \tag{15}$$

By (10) and (12) we get

$$G(t,x_1) \leq \sup_{y\in H}\{C(t)\|\mathcal{L}_t^* y\|_{L^q}|x_1| - c_p\|\mathcal{L}_t^* y\|_{L^q}^q\} \tag{16}$$

which gives, by standard calculations,

$$G(t,x_1) \leq c_p^{1-p}c_qC(t)^p|x_1|^p$$

so that putting the last equation and (14) into (15), we obtain

$$\mathcal{E}_p^*(t,x_1) - \mathcal{E}_p^*(t,x_2)$$

$$\leq \left[c_p^{1-p}c_qC(t)^p|x_1|^p\right]^{-1/q}C_1(p)C(t)^p|x_1|^{p-1}|x_1-x_2| = C_2(p)C(t)|x_1-x_2|,$$

where $C_2(p) = C_1(p)[c_p^{1-p}c_q]^{-1/q}$. Then (8) follows by repeating the same argument, exchanging the roles of x_1 and x_2.

To prove (9) we use a similar argument. First we fix $t_1 < t_2$ and we consider y_ε such that

$$G(t_1,x) \leq \varepsilon + \langle y_\varepsilon, e^{t_1A}x\rangle - c_p\|\mathcal{L}_{t_1}^* y_\varepsilon\|_{L^q}^q.$$

Then, reasoning as in the above proof, we have

$$G(t_1,x)-G(t_2,x)\le \varepsilon+\langle y_\varepsilon,e^{t_1A}x\rangle-c_p\|\mathcal{L}^*_{t_1}y_\varepsilon\|^q_{L^q}-\langle y_\varepsilon,e^{t_2A}x\rangle+c_p\|\mathcal{L}^*_{t_2}y\|^q_{L^q}$$

$$=\ \langle e^{t_1A^*}y_\varepsilon,x-e^{(t_2-t_1)A}x\rangle+c_p\int_{t_1}^{t_2}|B^*e^{sA^*}y_\varepsilon|^q ds$$

$$\le C(t_1)\|\mathcal{L}^*_{t_1}y_\varepsilon\|_{L^q}|x-e^{(t_2-t_1)A}x|+c_p|B^*|^q\int_{t_1}^{t_2}\|\mathcal{L}^*_s y_\varepsilon\|^q_{L^q}ds.$$

Now, arguing as in (13), we get

$$|\mathcal{L}^*_{t_1}y_\varepsilon|^q_{L^q}\le\frac{2}{c_p}C_{c_p/2}C(t_1)^p|x|^p$$

and

$$\sup_{s\in[t_1,t_2]}\|\mathcal{L}^*_s y_\varepsilon\|^q_{L^q}\le\|\mathcal{L}^*_{t_2}y_\varepsilon\|^q_{L^q}\le\frac{2}{c_p}C_{c_p/2}C(t_2)^p|x|^p$$

so that

$$G(t_1,x)-G(t_2,x)\le\varepsilon+C_3(p)\left[C(t_1)^p|x|^{p-1}|x-e^{(t_2-t_1)A}x|+(t_2-t_1)C(t_2)^p|x|^p\right]$$

$$\le\varepsilon+C_3(p)C(t_1)^p|x|^{p-1}\left[|x-e^{(t_2-t_1)A}x|+|x|(t_2-t_1)\right],$$

where we have used that $C(t_2)\le C(t_1)$. Finally, reasoning as in (15) and using the analogue of (16),

$$\mathcal{E}^*_p(t_1,x)\ -\mathcal{E}^*_p(t_2,x)\le\mathcal{E}^*_p(t_1,x)^{1-p}\left[G(t_1,x)-G(t_2,x)\right]$$

$$=G(t_1,x)^{-1/q}\left[G(t_1,x)-G(t_2,x)\right]$$

$$\le\left[C_4(p)C(t_1)^p|x|^p\right]^{-1/q}C_3(p)C(t_1)^p|x|^{p-1}\left[|x-e^{(t_2-t_1)A}x|+|x|(t_2-t_1)\right]$$

$$\le C_5(p)C(t_1)\left[|x-e^{(t_2-t_1)A}x|+|x|(t_2-t_1)\right]$$

which gives the claim. □

Remark 3.10. From the proof of (ii) it follows that, for $t_1<t_2$, there exists $\varepsilon(t_2-t_1)>0$ such that

$$e^{(t_2-t_1)A}\mathcal{L}_{t_1}(\mathcal{U}_{p,\rho})\subset\mathcal{L}_{t_2}(\mathcal{U}_{p,\rho-\varepsilon}).$$

Moreover, it also follows that, for $(t,x)\in(0,+\infty)\times H$, the optimal trajectory $y(\cdot;x,u^*)$ corresponding to the minimum p-energy control u^* arrives at 0 only at the final time t. Finally, it is easy to find a one dimensional example ($A=0$, $B=1$), where the minimum 1-energy is constant with respect to t. So the case $p=1$ cannot satisfy our claims.

4. Main results. In this section we state and prove the following main results of the paper: Theorem 4.1, where we prove that the minimum time is implicitly defined by the minimum energy; Theorem 4.6, where we show that the growth rate of the minimum time $T_{p,\rho}(x)$ near $x = 0$ is connected to the explosion rate of the minimum energy $\mathcal{E}_p^*(t,x)$ as $t \searrow 0$; Theorem 4.8, where we prove the continuity of the minimum time and give an estimate of the modulus of local uniform continuity (and of the Hölder exponent); and Theorem 4.10, where we prove topological properties of reachable sets.

THEOREM 4.1. *Let $p \in (1, +\infty]$ and $\rho > 0$ be fixed. Assume that the system* (1) *satisfies Hypothesis* 2.1 *and is p-null controllable. Then the minimal time function $T_{p,\rho}$ is defined implicitly by the equation*

$$\mathcal{E}_p^*(T_p(x), x) = \rho. \tag{17}$$

For the proof of this theorem we need two lemmas that are interesting in themselves. The first one states existence of optimal controls for the minimum time problem while the second one states a "maximum principle" for the optimal control.

LEMMA 4.2. *Let $p \in (1, +\infty]$ and assume that the system* (1) *satisfies Hypothesis* 2.1*. Then there exists an optimal control for the minimum time problem.*

Proof. We give the proof only for the case $p \in (1, +\infty)$, as the case $p = +\infty$ can be found in [17]. For brevity we will set $\rho = 1$ along this proof. Then let $p \in (1, +\infty)$ and $x \in \mathcal{R}_p$. By definition of $T_p(x)$, there exists a sequence of controls $u_n \in \mathcal{U}_{p,1}$ and a sequence of times $T_n \downarrow T_p(x)$ such that $T_n \leq 2T_p(x)$ and

$$y(T_n; x, u_n) = 0.$$

The control strategy u_n can be taken equal to 0 out of $[0, T_n]$, and by passing to a subsequence (which we still denote by u_n) it converges weakly to an element $u_0 \in \mathcal{U}_{p,1}$. Now we can write, for every $h \in H$,

$$0 = \langle y(T_n; x, u_n), h\rangle$$

$$= \langle e^{T_n A}x, h\rangle + \int_0^{2T_p(x)} \langle u_n(s), \chi_{[0,T_n]}(s) B^* e^{(T_n - s)A^*} h\rangle ds.$$

By the strong continuity of $\{e^{tA^*},\ t \geq 0\}$, it follows that for every $h \in H$ the map $s \to \chi_{[0,T_n]}(s)B^* e^{(T_n-s)A^*} h$ is measurable and bounded and converges strongly in $L^q(0, +\infty; U)$ as $n \to +\infty$ for $q = \frac{p}{p-1}$. Then, by the weak convergence of u_n to u_0 we have, for every $h \in H$,

$$0 = \lim_{n \to +\infty} \langle y(T_n; x, u_n), h\rangle\ =\ \langle y(T_p(x); x, u_0), h\rangle.$$

This gives $y(T; x, u_0) = 0$ and so the claim follows. □

LEMMA 4.3. *Let $p \in (1, +\infty]$. Assume that the system* (1) *satisfies Hypothesis* 2.1 *and is p-null controllable. Then any optimal control u^* for the minimum time problem satisfies the following "maximum principle":*

$$\|u^*\|_p = \rho.$$

Proof. For brevity we will set $\rho = 1$ along this proof. Let $x \in H$ and assume by contradiction that $\|u^*\|_p = a < 1$. Then we prove that $T_p(x)$ is not optimal.

We observe first that, given $u \in \mathcal{U}_p$, $x \in H$, and $0 < s < t$, we can write

$$\mathcal{L}_t u = e^{sA}\mathcal{L}_{t-s}u + \mathcal{L}_s[u(t-s+\cdot)]$$

so that, setting $z(t-s) = y(t-s;x,u) - x$ and $\overline{u}(r) = u(t-s+r)$, we have

$$y(t;x,u) = e^{tA}x + \mathcal{L}_t u = e^{sA}x + e^{sA}[e^{(t-s)A}x - x] + e^{sA}\mathcal{L}_{t-s}u + \mathcal{L}_s[u(t-s+\cdot)]$$

$$= e^{sA}x + \mathcal{L}_s[u(t-s+\cdot)] + e^{sA}z(t-s) = y(s;x,\overline{u}) + e^{sA}z(t-s).$$

Then taking in the above equation $t = T_p(x)$ and u an optimal control for the minimum time problem we have that, for $0 < s < t$,

$$y(s;x,\overline{u}) + e^{sA}z(t-s) = 0,$$

where $\overline{u}$ and $z(t-s)$ are defined as above. Now, by the p-null controllability assumption we have $e^{sA}(H) \subset \mathcal{L}_s\mathcal{U}_p$, which implies (see, e.g., [11]) that, for every $s_0 > 0$, $\varepsilon > 0$, there exists $\delta > 0$ such that

$$e^{sA}\{|x| \le \delta\} \subset \mathcal{L}_s\mathcal{U}_{p,\varepsilon} \quad \forall s \ge s_0.$$

Now fix $\varepsilon = (1-a)/2$ and $s_0 = T_p(x)/2$. Take δ as above and $s_0 \le s < T_p(x)$ such that $|z(t-s)| \le \delta$. Then there exists a control $\widehat{u}$ such that $\|\widehat{u}\|_p \le \frac{1-a}{2}$ and $e^{sA}z(t-s) = \mathcal{L}_s\widehat{u}$ so that

$$y(s;x,\overline{u}+\widehat{u}) = e^{sA}x + \mathcal{L}_s[\overline{u}+\widehat{u}] = e^{sA}x + \mathcal{L}_s\overline{u} + e^{sA}z(t-s) = 0.$$

The claim follows by contradiction. □

Proof of Theorem 4.1. We argue by contradiction, setting for brevity $\rho = 1$. If $\mathcal{E}_p^*(T_p(x),x) = a > 1$, then every control driving x to 0 in time $T_p(x)$ has p-energy strictly greater than 1 and so it is not admissible. This is impossible due to the existence of the optimal control stated in Lemma 4.2.

Conversely, assume that $\mathcal{E}_p^*(T_p(x),x) = a < 1$. Then, by Theorem 3.3 there exists a control driving x to 0 in time $T_p(x)$ with energy $a < 1$. This contradicts the "maximum principle" stated above. Finally we observe that if $t \ne T(x)$, then $\mathcal{E}_p^*(t,x) \ne 1$, since the minimum energy is strictly decreasing in the variable t. □

Remark 4.4. Observe that, by definition of $T_p(x)$, $\mathcal{E}_p^*(T,x) \le 1$ when $T > T_p(x)$. Similarly we have that $\mathcal{E}_p^*(T,x) \ge 1$ when $T < T_p(x)$. Then, in the case $p \in (1,+\infty)$, the claim of Theorem 4.1 follows also by the continuity and the strict monotonicity of $\mathcal{E}_p^*(T,x)$ with respect to T (see Theorem 3.7).

Remark 4.5. The control u of minimal energy at time $T_p(x)$ satisfies also $\mathcal{E}_p(u) = 1$ and so it is optimal also for the minimum time problem. In particular, when $p \in (1,+\infty)$, this implies the uniqueness of the optimal control for the minimum time problem. Observe that if $p = 1$, we do not have existence of the minimum time optimal control; it is enough to take the one dimensional case with $A = 0$, $B = 1$.

The representation given above allows us to connect the behavior of T_p near 0 with L^p estimates of controls bringing a given state x to 0.

THEOREM 4.6. *Let* $p \in (1,+\infty]$. *Assume that the system* (1) *satisfies Hypothesis* 2.1 *and is p-null controllable. Let* $f : \mathbb{R}^+ \to \mathbb{R}^+$ *be a continuous strictly decreasing function such that* $\lim_{t\to 0^+} f(t) = +\infty$. *Then there exists* $T_0 > 0$ *such that*

$$\mathcal{E}_p^*(T,x) \le f(T)|x| \qquad \forall x \in H, \quad T \in]0,T_0] \tag{18}$$

if and only if*, taking as f^{-1} the inverse of f, we have*

$$T_{p,\rho}(x) \le f^{-1}(\rho/|x|)$$

in a suitable neighborhood of $x = 0$. The same holds with the reverse inequalities.

Proof. For brevity we will set $\rho = 1$ along this proof. We start by proving the "only if" part. Let $0 < T < T_0$ and $x \in H$. By Theorem 3.3, it follows that there exists a control $u_{T,x} \in \mathcal{M}(p, T, x)$ such that $\|u_{T,x}\|_p \le f(T)|x|$. If $f(T)|x| \le 1$, then the control $u_{T,x}$ belongs to $\mathcal{U}_{p,1}$ and is admissible for the minimum time problem. It follows that for every $x \in H$, such that $1/|x| \in \operatorname{Im} f$, the time $\overline{T}(x) = f^{-1}(1/|x|)$ satisfies $\overline{T}(x) \ge T_p(x)$ and the claim follows.

Conversely, assume that there exists $R_1 > 0$ such that $T_p(x) \le f^{-1}(1/|x|)$ $\forall |x| \le R_1$. This fact, together with the monotonicity of f^{-1}, implies that, given $t > 0$, all points x satisfying $|x| \le R_1$ and

$$f^{-1}(1/|x|) \le t \Longleftrightarrow |x| \le 1/f(t)$$

are controllable to 0 in time t with controls in $\mathcal{U}_{p,1}$. It follows that, taking t such that $1/f(t) \le R_1$,

$$e^{tA}\left\{|x| \le 1/f(t)\right\} \subset \mathcal{L}_t \left\{\|u\|_p \le 1\right\}$$

which is equivalent to (see [11])

$$e^{tA}\left\{|x| \le r\right\} \subset \mathcal{L}_t \left\{\|u\|_p \le f(t)r\right\} \qquad \forall r > 0$$

which gives the claim. The proof in the case of the reverse inequalities is completely similar and we omit it. □

Remark 4.7. From the above proof we can infer that

$$\mathcal{R}_{p,\rho} = \left\{x \in H : \lim_{t\to+\infty} \mathcal{E}_p^*(t,x) < \rho\right\}$$

and that, setting $g(t) = \mathcal{E}_p^*(t,x)$, g is invertible and $T_{p,\rho}(x) = g^{-1}\left(\rho/|x|\right)$.

Moreover, if we have $f(T) = CT^{-\beta}$ for suitable $C > 0$, then by the previous theorem we get that $\mathcal{E}_p^*(T,x) \le CT^{-\beta}|x|$ for T in a neighborhood of $t = 0$ and x in a neighborhood of $x = 0$ *if and only if* $T_p(x) \le C_0[|x|/\rho]^{1/\beta}$ in another neighborhood of $x = 0$. The same holds for the reverse inequalities. This means that the explosion rate of $\mathcal{E}_p^*(t,x)$ at $t = 0$ *characterizes* the Hölder continuity exponent of $T_{p,\rho}(x)$ at $x = 0$. So, in the finite dimensional case this exponent is completely determined thanks to the sharp result of [35, 37] (see Theorem 3.6).

The main consequence of Theorem 4.1 is the following result on the local uniform continuity (and the Hölder continuity) of $T_{p,\rho}$ in the case $p \in (1, +\infty)$ (the continuity of T_∞ is already known and is proved in [15]).

THEOREM 4.8. *Let $p \in (1, +\infty)$. Assume that the system* (1) *satisfies Hypothesis* 2.1 *and is p-null controllable. Then $T_p(\rho, \cdot)$ is locally uniformly continuous in $x \in \mathcal{R}_{p,\rho}$. Assume also that, for a given decreasing function $f : \mathbb{R}^+ \to \mathbb{R}^+$ with $\lim_{r\to0^+} f(r) = +\infty$, condition* (18) *is satisfied. Then a modulus of continuity of $T_p(\rho, \cdot)$ on $\mathcal{R}_{p,\rho}(t) \cap \{|x| \le K\}$ is given by $r \mapsto g\left(Me^{\omega t}[1 + K]r^{1-1/p}\right)$, where $g(r) = f^{-1}(\rho/r)$.*

More precisely, for every x_1, $x_2 \in \mathcal{R}_{p,\rho}(t) \cap \{|x| \le K\}$ we have

$$|T_p(\rho, x_1) - T_p(\rho, x_2)| \le g\left(Me^{\omega t}[1 + K]\left|(x_1 - x_2)\right|^{1-\frac{1}{p}}\right). \tag{19}$$

In particular, if we can choose $f(r) = C/r^{\beta}$ *for suitable* $C > 0$*,* $\beta \geq 1$*, then the minimum time function* T_p *is locally Hölder continuous with exponent* $(q\beta)^{-1}$*, where* $q^{-1} + p^{-1} = 1$*.*

We will give the proof of the above theorem in section 5.2.

Remark 4.9. A representation of the minimum time function such as (17) and a "maximum principle" such as the one of Lemma 4.3 can be found in [38] for the case $p = 2$ but no consequence of these results is given there. Moreover, Lemma 4.3 could be seen as a consequence of the following fact: fixed $x \in H$ and $p \in (1, +\infty]$, the set-valued map $t \to \mathcal{M}(p, t, x)$ is left lower semicontinuous (see [38] for the proof of this fact in the case $p = 2$).

Finally, the fact that $T_{p,\rho}$ is lower semicontinuous in H (also without assuming p-null controllability) has been given in [38] for the case $p = 2$ by a topological argument that can be generalized to the case $p \in (1, +\infty)$.

For the reachable sets we have the following theorem which generalizes to the case $p \in (1, +\infty)$ a result given in [12] for the case $p = \infty$.

THEOREM 4.10. *Assume that Hypothesis* 2.1 *holds and let* $p \in (1, +\infty)$*,* $\rho > 0$*.*

(i) *The system* (1) *is p-null controllable if and only if* $0 \in \operatorname{Int} \mathcal{R}_{p,\rho}(t)$ *for every* $t > 0$*.*

(ii) *The system* (1) *is p-null controllable if and only if* $\mathcal{R}_{p,\rho}(t_1) \subset \operatorname{Int} \mathcal{R}_{p,\rho}(t_2)$ *for every* $t_1 < t_2$*.*

(iii) *If the system* (1) *is p-null controllable, then* $\mathcal{R}_{p,\rho}$ *is open and* $\lim_{y \to x} T_{p,\rho}(y) = +\infty \ \forall x \in \operatorname{Fr} \mathcal{R}_{p,\rho}$*.*

(iv) *If the system* (1) *is p-null controllable, then* $\mathcal{R}_{p,\rho}(t) = \{x : T_{p,\rho}(x) \leq t\}$*,* $\operatorname{Int} \mathcal{R}_{p,\rho}(t) = \{x : T_{p,\rho}(x) < t\}$*, and* $\operatorname{Fr} \mathcal{R}_{p,\rho}(t) = \{x : T_{p,\rho}(x) = t\}$*.*

Before we prove the above theorem we give a lemma that is interesting in itself since it allows us to use the properties of the functions $\mathcal{E}_p^*$ and $T_{p,\rho}$ to study $\mathcal{R}_{p,\rho}(t)$.

LEMMA 4.11. *Assume that Hypothesis* 2.1 *holds and let* $p \in (1, +\infty)$*. If the system* (1) *is p-null controllable, then*

(i) $\mathcal{R}_{p,\rho}(t) = \{x : \mathcal{E}_p^*(t, x) \leq \rho\} = \{x : T_{p,\rho}(x) \leq t\}$,

(ii) $\operatorname{Int} \mathcal{R}_{p,\rho}(t) = \{x : \mathcal{E}_p^*(t, x) < \rho\} = \{x : T_{p,\rho}(x) < t\}$,

(iii) $\operatorname{Fr} \mathcal{R}_{p,\rho}(t) = \{x : \mathcal{E}_p^*(t, x) = \rho\} = \{x : T_{p,\rho}(x) = t\}$.

Proof of Lemma 4.11. For brevity we set $\rho = 1$ and drop the dependence on ρ writing $\mathcal{R}_p, T_p$ instead of $\mathcal{R}_{p,\rho}, T_{p,\rho}$. We prove only the identities $\mathcal{R}_p(t) = \{x : \mathcal{E}_p^*(t, x) \leq 1\}$, $\operatorname{Int} \mathcal{R}_p(t) = \{x : \mathcal{E}_p^*(t, x) < 1\}$, and $\operatorname{Fr} \mathcal{R}_p(t) = \{x : \mathcal{E}_p^*(t, x) = 1\}$ since the other ones, $\{x : \mathcal{E}_p^*(t, x) \leq 1\} = \{x : T_p(x) \leq t\}$, $\{x : \mathcal{E}_p^*(t, x) < 1\} = \{x : T_p(x) < t\}$, and $\{x : \mathcal{E}_p^*(t, x) = 1\} = \{x : T_p(x) = t\}$, easily follow from Theorem 4.1 and the strict decreasing property of the function $t \to \mathcal{E}_p^*(t, x)$ for every $x \in H$.

For the first identity we observe that, by definition of $\mathcal{R}_p(t)$, we have

$$\mathcal{R}_p(t) = \{x : \mathcal{M}(p, t, x) \cap \mathcal{U}_{p,1} \neq \emptyset\}.$$

Now, if $x \in \mathcal{R}_p(t)$ it is clear by the definition that $\mathcal{E}_p^*(t, x) \leq 1$. Conversely, if $\mathcal{E}_p^*(t, x) \leq 1$ then, by the existence of the optimal control for the minimum energy problem, we have $x \in \mathcal{R}_p(t)$.

For the second identity, first let $\mathcal{E}_p^*(t, x) < 1$. Then, by the continuity of the map $x \to \mathcal{E}_p^*(t, x)$ we have that, for y in a suitable neighborhood of x, $\mathcal{E}_p^*(t, y) < 1$ and so $x \in \operatorname{Int} \mathcal{R}_p(t)$. Conversely, let $x \in \operatorname{Int} \mathcal{R}_p(t)$. By the first identity we have $\mathcal{E}_p^*(t, x) \leq 1$. If by contradiction we have $\mathcal{E}_p^*(t, x) = 1$, then by the homogeneity of degree 1 of $\mathcal{E}_p^*(t, \cdot)$ (see Theorem 3.7) it follows that for every $\lambda > 1$ we have $\mathcal{E}_p^*(t, \lambda x) > 1$ and so $\lambda x \notin \mathcal{R}_p(t)$, which is a contradiction.

For the third identity it is enough to recall that $\mathcal{R}_p(t)$ is closed and so

$$\mathrm{Fr}\,\mathcal{R}_p(t) = \mathcal{R}_p(t) - \mathrm{Int}\,\mathcal{R}_p(t) = \{x : \mathcal{E}_p^*(t,x) = 1\}. \qquad \square$$

Proof of Theorem 4.10. As in the proof above we set $\rho = 1$ and drop the dependence on ρ. We first observe that from the existence of optimal control we have, also without assuming p-null controllability, $\mathcal{R}_p(t) = \{x : T_p(x) \le t\}$ $\forall t > 0$. Then to prove (i) it is enough to observe that, by Theorem 2.5 (see [15]), the p-null controllability of the system (1) is equivalent to the continuity of T_p at $x = 0$ and thus equivalent to the fact that the set $\{T_p \le t\}$ is a neighborhood of $x = 0$ for every $t > 0$.

We now prove (ii). First, if $\mathcal{R}_p(t_1) \subset \mathrm{Int}\,\mathcal{R}_p(t_2)$ for every $t_1 < t_2$, then given any $t > 0$ we have $0 \in \mathcal{R}_p(\frac{t}{2}) \subset \mathrm{Int}\,\mathcal{R}_p(t)$ and thus have p-null controllability by part (i). Conversely, if p-null controllability holds true, then, by Lemma 4.11, for every $t_1 < t_2$ we have

$$\mathcal{R}_p(t_1) = \{x : T_p(x) \le t_1\} \subset \{x : T_p(x) < t_2\} = \mathrm{Int}\,\mathcal{R}_p(t_2).$$

We prove (iii). First we observe that $\mathcal{R}_p = \cup_{t>0}\{T_p < t\}$ and so it is open. Moreover, take $x \in \mathrm{Fr}\,\mathcal{R}_p$ and a sequence $(x_n) \subset \mathrm{Int}\,\mathcal{R}_p$ such that $x_n \to x$. From (ii) it follows that, for every $t > 0$, $\mathcal{R}_p(t) \subset \mathrm{Int}\,\mathcal{R}_p$. So, since $\mathcal{R}_p(t)$ is closed, we have definitively $x_n \notin \mathcal{R}_p(t)$ which means $T_p(x_n) > t$. The claim follows since t is arbitrary.

Statement (iv) follows from Lemma 4.11. $\square$

5. Uniform continuity and Hölder continuity.

In this section we are concerned with Hölder continuity properties of the Bellman function T_p. In the above section we obtained an estimate of the growth rate of T_p in a neighborhood of $x = 0$ in connection with the explosion rate of the minimum energy at $t = 0$. Here we discuss the possibility of extending such an estimate out of $x = 0$.

In the case $p = +\infty$ we use the results of the previous section and the dynamic programming principle, recalled below, to derive local Hölder continuity of the function T_∞ in some cases by extending results of [15].

In the case $p \in (1, +\infty)$ (where no results are available in the literature even in finite dimension) we observe, giving also an example, that the dynamic programming principle does not hold in general. Then, letting ρ be a new state variable of the problem, we prove a version of the dynamic programming principle which allows us again to transfer (with a different method) the estimate of Theorem 4.6 to the whole reachable sets obtaining local uniform continuity, an estimate of the local modulus and, in some cases, of the local Hölder exponent.

We then discuss a finite dimensional example where, for a given $T_0 > 0$, we have $\|\Gamma(t)\| \le Ct^{-3/2}$, $0 < t \le T_0$, and $T_{2,1}$ is locally Hölder continuous with exponent $1/3$ and without higher exponents.

5.1. The case $p = +\infty$.

For simplicity we set $\rho = 1$ and drop the dependence on ρ writing $\mathcal{R}_p, T_p$ instead of $\mathcal{R}_{p,\rho}, T_{p,\rho}$.

We start by recalling the dynamic programming principle for the minimum time problem if $p = +\infty$ (proved, e.g., in [15]): *For every $t \le T_\infty(x)$, for all $x \in \mathcal{R}_\infty$ we have*

$$T_\infty(x) = \inf\{t + T_\infty(y(t; x, u));\ u \in \mathcal{U}_{\infty,1}\}.$$

We observe that the dynamic programming principle is very important for studying the regularity of T_∞. In fact it allows us to estimate the continuity modulus of T_∞ by its behavior in a neighborhood of 0 (see [15] for the proof).

The above result, together with the result of the previous section, gives as a consequence the following corollary (see [15] for the proof).

COROLLARY 5.1. *Assume that the system* (1) *satisfies Hypothesis* 2.1 *and is* ∞*-null controllable. Then the minimum time function* T_∞ *is locally uniformly continuous. Assume also that, for a given decreasing function* $f : \mathbb{R}^+ \to \mathbb{R}^+$ *with* $\lim_{r\to 0^+} f(r) = +\infty$, *condition* (18) *is satisfied. Then a modulus of continuity of* T_∞ *is given by*

$$|T_\infty(x_1) - T_\infty(x_2)| \leq g\left(Me^{\omega t}\left|(x_1 - x_2)\right|\right),$$

where $g(r) = f^{-1}(1/r)$. *In particular, if we can choose* $f(r) = C/r^\beta$ *for suitable* $C > 0$, $\beta \geq 1$, *then the minimum time function* T_∞ *is locally Hölder continuous with exponent* $1/\beta$.

Remark 5.2. The claim of Corollary 5.1 is a consequence of the dynamic programming principle. This renders the case $p = +\infty$ easier to treat when we want to prove continuity properties of the minimum time function. We will see that, even for simple cases, for $p < +\infty$ the dynamic programming principle does not hold (see Example 5.14).

At this point, to prove local Hölder continuity of T_∞ we can simplify the problem to find good L^∞ estimates on the controls that bring to 0 the points x belonging to a neighborhood of 0. We remark that the best thing to do would be to calculate the minimal ∞-energy needed to bring x to 0. We now show some situations where this estimate can be given. Some of them (as we will specify below) are already given in the literature, but our results allow us to see them in a unified approach.

We start by estimating the ∞-norm of the control of minimal 2-energy $u^*_{T,x}$ defined in Theorem 3.3, since this norm is surely greater than the minimal ∞-energy needed to bring x to 0.

PROPOSITION 5.3. *Assume that Hypothesis* 2.1 *holds, that the system* (1) *is* 2*-null controllable, and that* A *generates a group. Then the system* (1) *is also* ∞*-null controllable. Moreover, if there exists* $C > 0$ *such that* $\|\Gamma(t)\| \leq Ct^{-\beta/2}$ *as* $t \to 0^+$, *then* T_∞ *is locally Hölder continuous of exponent* $\frac{1}{\beta}$. *If the dimension of* H *is finite, then* T_∞ *is locally Hölder continuous of exponent* $2/(1+\beta)$.

Proof. We first remark that by Theorem 3.3 the control $u^*_{T,x}(r) = \Gamma_r(T)^*\Gamma(T)x$, where $\Gamma_r(T) = Q_T^{-\frac{1}{2}}S(T-r)B$ transfer x to 0 in time T. We now show that $u^*_{T,x}$ is bounded and we estimate $\|u^*_{T,x}\|_\infty$. Since A generates a group, then we can write $\Gamma_r(T) = \Gamma(T)S(-r)B$, which gives

$$\|u^*_{T,x}\|_\infty \leq \sup_{r\in[0,T]} \|\Gamma(T)\|^2 \|S(-r)B\| |x|$$

and the first part of the claim easily follows.

When the dimension of H is finite it is enough to use the results of Seidman and Yong [37] contained in Theorem 3.6 for the case $p = +\infty$.

Remark 5.4. In the case of finite dimension the results of Gyurkovics [19] state that T_∞ is locally Hölder continuous exactly of exponent $1/(k+1)$, where k is the first integer such that rank $[B, AB, \ldots, A^kB] = n$. By Theorem 3.6 it then follows that the relationship between the singularity of $\|\Gamma(T)\|$ as $T \to 0$ and the continuity properties of T_∞ when $H = \mathbb{R}^n$ is the following: If $\|\Gamma(t)\| \leq Ct^{-\beta/2}$, then T_∞ is locally Hölder continuous of exponent $2/(1+\beta)$. This result confirms the one of Proposition 5.3. The same result can be deduced by the estimate of the ∞-minimum energy given in Theorem 3.6 (see [37]).

Remark 5.5. By the proof of Proposition 5.3, it follows that in the general case an estimate of the ∞-minimum energy needed to steer a state x to 0 (and so of the modulus of continuity of T_∞) is given by an estimate of the norm of the operator $\Gamma_r(T)^*\Gamma(T)$. In Proposition 5.3 such an estimate is given for the case when A is a group or H is finite dimensional. In fact, in some special cases, also when A does not generate a group, an estimate of this kind can be proved (see [16, Appendix B]). In particular, when B is onto it can be easily proved that $\sup_{r\in[0,T]}\|\Gamma_r(T)^*\Gamma(T)\| \leq C/T$ for a suitable constant C and so that $\|u^*_{T,x}\|_\infty \leq C|x|/T$. This follows also from Proposition 3.5 and proves local Lipschitz continuity of T_∞ when B is onto.

Remark 5.6. If we have for a given decreasing function $f : \mathbb{R}^+ \to \mathbb{R}^+$ $\|\Gamma(T)\| \leq f(T)$, then we obtain an estimate of the continuity modulus of T_∞. For example, for known boundary control problems (see, e.g., [36]) the best estimate is $f(T) = e^{1/T}$. This case is not treated in this paper. However, such estimates suggest that for boundary control problems the function T_∞ should not be Hölder continuous.

Remark 5.7. The result of the latter proposition is useful in the infinite dimensional case when it is possible to give an estimate of $\|\Gamma(t)\|$ as $t \searrow 0$. This is not a simple task but can be done, for example, when the system (1) is an abstract wave equation. This case has already been treated by Krabs [26] who found that $\|\Gamma(t)\| \sim t^{-3/2}$ as $t \searrow 0$ and by Carja [13] who applied this estimate to prove that T_∞ is locally Hölder continuous of exponent 1/3.

We now consider a diagonal case where the relation between the Hölder exponent of T_∞ and the singularity of $\|\Gamma(t)\|$ as $t \searrow 0$ is the same as the finite dimensional case. Assume the following.

HYPOTHESIS 5.8. *Let H be a separable Hilbert space and let $\{e_k\}$ be a complete orthonormal system in H. We assume here that A and B are of the following form:*

$$Ae_k = -\alpha_k e_k, \quad Be_k = b_k e_k, \quad k \in \mathbb{N}, \tag{20}$$

where $\{\alpha_k\}$ is an increasing (to $+\infty$) sequence of positive numbers and $\{b_k\}$ is a bounded sequence of positive numbers.

Under Hypothesis 5.8 we have, for $k \in \mathbb{N}$, $Q_t e_k = (1 - e^{-2\alpha_k t})b_k^2/(2\alpha_k)$ and, for every $t > 0$,

$$\|\Gamma(t)\|^2 = \sup_{k\in\mathbb{N}} \frac{2\alpha_k e^{-2\alpha_k t}}{b_k^2(1 - e^{-2\alpha_k t})} = \sup_{k\in\mathbb{N}} \frac{2\alpha_k}{b_k^2(e^{2\alpha_k t} - 1)}. \tag{21}$$

By Theorem 3.1 the system is 2-null controllable if and only if

$$\|\Gamma(t)\|^2 = \sup_{k\in\mathbb{N}} \frac{2\alpha_k}{b_k^2(e^{2\alpha_k t} - 1)} < +\infty. \tag{22}$$

In this case the control of minimum energy $u^*_{T,x}$ is given by

$$u^*_{T,x}(s) = \sum_{k=1}^{+\infty} \frac{2\alpha_k}{b_k} \frac{e^{-(2T-s)\alpha_k}}{1 - e^{-2T\alpha_k}} x_k e_k.$$

This control is bounded if and only if

$$\sup_{k\in\mathbb{N}} \frac{2\alpha_k}{b_k} \frac{1}{e^{2T\alpha_k} - 1} < +\infty$$

and in this case

$$\|u^*_{T,x}\|_\infty \leq \sup_{k\in\mathbb{N}} \frac{2\alpha_k}{b_k}\frac{1}{e^{2T\alpha_k}-1}|x|.$$

The estimates above depend on the behavior at infinity of the sequence $(b_k)_{k\in\mathbb{N}}$. Let us consider the particular case when $b_k = \alpha_k^{-\frac{r}{2}}$ for some $r \geq 0$ (the case $b_k \sim \alpha_k^{-\frac{r}{2}}$ would give the same results). Then (22) is fulfilled since

$$\|\Gamma(t)\|^2 = \sup_{n\in\mathbf{N}} \frac{2\alpha_k^{1+r}}{e^{2\alpha_k t}-1} \leq \frac{1}{t^{1+r}} \sup_{s>0} \frac{2s^{1+r}}{e^{2s}-1} \tag{23}$$

so that $\|\Gamma(t)\| \leq \sqrt{C(r)}t^{-(1+r)/2}$, where $C(r) = \sup_{s>0} 2s^{1+r}/(e^{2s}-1)$. Similarly we can prove that, for a suitable positive constant $c(r)$, $\|\Gamma(t)\| \geq \sqrt{c(r)}t^{-(1+r)/2}$ for t sufficiently small so that $\|\Gamma(t)\| \sim t^{-(1+r)/2}$ as $t \searrow 0$. Using the same arguments we can also show that $\|u^*_{T,x}\|_\infty \sim T^{-1-r/2}$. This allows us to extend to a diagonal infinite dimensional case the finite dimensional results of Gyurkovics [19] (see Proposition 5.3 and Remark 5.4).

Example 5.9. An example in which Hypothesis 5.8 is satisfied is the following. Let $C_N = [0,\pi]^N$ and $X = L^2(C_N)$, $N \leq 3$, and take the Laplace operator with Dirichlet conditions at the boundary defined as

$$D(A) = H^2(C_N) \cap H^1_0(C_N), Ax = \Delta x \text{ for } x \in D(A).$$

The operator A satisfies Hypothesis 2.1 and it generates an analytic semigroup of compact operators. Moreover, A satisfies Hypothesis 5.8 by taking, for $(n_1,\dots,n_N) \in \mathbb{N}^N$,

$$e_{n_1,\dots,n_N}(\xi) = \left(\frac{2}{\pi}\right)^{\frac{N}{2}} \sin n_1\xi_1 \cdots \sin n_N\xi_N$$

and $\alpha_{n_1,\dots,n_N}(\xi) = n_1^2 + \cdots + n_N^2$ so that, by ordering the eigenvalues, we obtain $\alpha_k \approx k^{\frac{2}{N}}$ as $k \to +\infty$. If we take $B = A^{-\frac{r}{2}}$, $r \geq 0$, then we have exactly the case introduced above.

We now proceed by estimating the sup norm of a different control bringing x to 0 (a similar idea, yielding different results, is used in [15]).

HYPOTHESIS 5.10.

(i) *For every* $t > 0$, $e^{tA}(H) \subset B(H)$.

(ii) *For some* $C, \alpha > 0$ $|B^{-1}e^{sA}x| \leq Cs^{-\alpha}|x|$, *where* B^{-1} *denotes the pseudo-inverse of* B *(see, e.g.,* [16, p. 407]*).*

Remark 5.11.

(i) Under Hypothesis 5.10(i), given $x \neq 0$ such that $e^{tA}x \neq 0$ for all $t > 0$, the map $t \to 1/|B^{-1}e^{tA}x|$ is locally integrable on $(0,+\infty)$ since $|B^{-1}e^{tA}x| > |e^{tA}x|/\|B\|$ and $\{e^{tA},\ t \geq 0\}$ is strongly continuous.

(ii) Hypothesis 5.10(i) implies that Ker $B^* = 0$ and so $B(H)$ is dense.

Let $\overline{u}_x(s) = -B^{-1}e^{sA}x/|B^{-1}e^{sA}x|$ (we take $\overline{u}_x(s) = 0$ if $e^{sA}x = 0$). In this case we have that $\overline{u}_x \in \mathcal{U}_{\infty,1}$ and

$$y(t;x,\overline{u}_x) = e^{tA}x\left[1 - \int_0^t \frac{1}{|B^{-1}e^{sA}x|}ds\right]$$

so, if $T > 0$ is such that

$$\int_0^T \frac{1}{|B^{-1}e^{sA}x|} ds = 1,$$

then $T_\infty(x) \le T$. Now thanks to Hypothesis 5.10(ii) we have, for $t > 0$,

$$\int_0^t \frac{1}{|B^{-1}e^{sA}x|} ds \ge \int_0^t \frac{s^\alpha}{C}|x| ds.$$

Since

$$\int_0^t \frac{s^\alpha}{C}|x| ds = 1 \Longleftrightarrow t = [C(\alpha+1)]^{\frac{1}{\alpha+1}} |x|^{\frac{1}{\alpha+1}}$$

we have, by Hypothesis 5.10(ii),

$$T_\infty(x) \le C_1 |x|^{\frac{1}{\alpha+1}} \quad \text{for some } C_1 > 0$$

which gives, by Theorem 4.6, that T_∞ is locally Hölder continuous of exponent $\frac{1}{\alpha+1}$. We have thus proved the following.

PROPOSITION 5.12. *Assume that Hypotheses* 2.1 *and* 5.10 *hold true. Then the system* (1) *is* ∞*-null controllable and* T_∞ *is locally Hölder continuous with exponent* $1/(\alpha+1)$.

Remark 5.13. If the system (1) is 2-null controllable and Im $\{Q_t^{\frac{1}{2}}\} \subset$ Im $\{B\}$ $\forall t > 0$, then we can write

$$|B^{-1}e^{sA}x| = |B^{-1}Q_s^{\frac{1}{2}}\Gamma(s)x| \le |B^{-1}Q_s^{\frac{1}{2}}||\Gamma(s)x|.$$

So, if we assume that, for a given constant $C > 0$ and for s in a neighborhood of 0 we have $|\Gamma(s)x| \le Cs^{-\beta/2}|x|$, then, to apply Proposition 5.12 we have only to estimate $|B^{-1}Q_s^{\frac{1}{2}}|$. This can be done, for example, in the diagonal case described above yielding the same results obtained there.

A framework where Proposition 5.12 can be used is the following. Let A be the generator of an analytic semigroup on H such that $A^{-1} \in \mathcal{L}(H)$. Assume also that B can be written as $A^{-r}B_0$, where B_0 is continuous and onto (so that B_0^{-1} is also continuous).

Then we clearly have, by the analyticity of the semigroup, $e^{tA}(H) \subset D(A^r) = B(H)$ and, for every $t > 0$, $x \in H$, and a suitable constant $C > 0$,

$$|B^{-1}e^{tA}x| = |B_0^{-1}A^r e^{tA}x| \le \|B_0^{-1}\| \, \|A^r e^{tA}\| \, |x| \le Ct^{-r}|x|.$$

This implies that T_∞ is locally Hölder continuous with exponent $1/(1+r)$.

5.2. The case $p \in (1, +\infty)$. In the following we construct an example for the case $p = 2$, showing that the dynamic programming principle does not apply. This difference is crucial when studying regularity properties of the minimum time function. Then we will consider the minimum time problem where the radius ρ of the set of admissible controls is variable. For this problem we prove the dynamic programming principle and then we use it to get local estimates of the modulus of uniform continuity of the Bellman function.

Applying the same argument used in the proof of Theorem 4.6 we obtain that, in the case $p = 2$,

$$\begin{gathered} \|\Gamma(t)\| \leq C_1 t^{-\beta/2}|x| \text{ in a right neighborhood of } t = 0 \\ \Downarrow \\ T_{2,\rho}(x) \leq C_2[|x|/\rho]^{2/\beta} \text{ in a neighborhood of } x = 0 \end{gathered}$$

and, by the dynamic programming principle,

$$\begin{gathered} T_{2,\rho}(x) \leq C_2[|x|/\rho]^{\frac{2}{\beta}} \text{ in a neighborhood of } x = 0 \\ \Updownarrow \\ T_{2,\rho}(x_1) - T_{2,\rho}(x_2) \leq [C_3/\rho^{2/\beta}]|x_1 - x_2|^{\frac{2}{\beta}} \quad \forall x_1, x_2 \in \mathcal{R}_2, |x_1 - x_2| \text{ sufficiently small.} \end{gathered}$$

In particular this would imply that, when $B = I$, the function T_2 is constant, which is not true as can be seen by the following example.

Example 5.14. Let $H = \mathbb{R}$ and the state equation be

$$y' = y + u; \qquad y(0) = x.$$

Then $Q_t = \int_0^t e^{2s} ds = \frac{e^{2t}-1}{2}$ and $\Gamma(t) = Q_t^{-\frac{1}{2}} e^{tA} = \frac{\sqrt{2}}{\sqrt{e^{2t}-1}} e^t$ so that

$$|\Gamma(t)x| = \frac{\sqrt{2}}{\sqrt{1 - e^{-2t}}}|x| \text{ and, in particular, } |\Gamma(t)x| \approx \frac{C}{t^{\frac{1}{2}}}.$$

By Theorem 4.1 we have $|\Gamma(T_2(x))x| = \rho$, so we obtain by easy calculations that $\mathcal{R}_{2,\rho} = (-\frac{\rho}{\sqrt{2}}, \frac{\rho}{\sqrt{2}})$ and $T_{2,\rho}(x) = -\frac{1}{2}\log(1 - 2x^2/\rho^2)$.

Remark 5.15. This example is in contrast with the result given in [15, Lemma 2.1]. In fact the proof of this lemma seems to be incorrect in the case $p < +\infty$. From the proof of the dynamic programming principle it can be seen that for $p \in (1, +\infty)$ only the following inequality holds: $T_p(x) \geq \inf\{t + T_p(y(t; x, u)),\ u \in \mathcal{U}_{p,1}\}$.

Consider now, for $p \in (1, +\infty]$, the minimum time problem, where the radius ρ of the set of admissible controls is a state variable that represents the amount of p-energy left to consume. Then ρ clearly follows the equation

$$\frac{d}{dt}\rho^p(t) = |u(t)|^p; \qquad \rho(0) = \rho_0$$

which can be written in integral form as (denoting as $\rho(t; \rho_0, u)$ the unique solution of the above equation)

$$\rho^p(t; \rho_0, u) = \rho_0^p - \int_0^t |u(s)|^p ds.$$

The latter equation, coupled with (1), gives the state equation of the new problem (with state variables ρ and y); the Bellman function will depend now on (ρ, x) and we will denote it by $T_p(\rho, x)$. We have the following result (which is the main advantage of taking the ρ variable).

THEOREM 5.16. *Let $p \in (1, +\infty)$ and $(\rho_0, x_0) \in (0, +\infty) \times R_{p,\rho_0}$. For every $t \leq T_p(\rho_0, x_0)$ we have*

$$T_p(x) = \inf\left\{t + T_p(\rho(t; \rho_0, u), y(t; x_0, u));\ u \in \mathcal{U}_{\infty,\rho_0}\right\}.$$

Proof. The proof follows along the same line of the proof given in [15] after taking account of the variability of ρ. Fix $x \in H$. The first step of the proof is to show

$$T_p(\rho_0, x_0) \le t + T_p(\rho(t;\rho_0,u), y(t;x_0,u)) \quad \forall u \in \mathcal{U}_{\infty,\rho_0}.$$

If $y(t;x_0,u)$ is not in $\mathcal{R}_{p,\rho_0}$, then $T_p(\rho(t;\rho_0,u), y(t;x_0,u)) = +\infty$ and there is nothing to prove. Hence, we can assume that $y(t;x_0,u) \in \mathcal{R}_{p,\rho(t;\rho_0,u)}$ and this is equivalent to saying that $T_p(\rho(t;\rho_0,u), y(t;x_0,u)) < +\infty$. Let $s \ge T_p(\rho(t;\rho_0,u), y(t;x_0,u))$. Then $y(t;x_0,u) \in \mathcal{R}_{p,\rho(t;\rho_0,u)}(s)$, i.e., $y(s, y(t;x_0,u), \hat{u}) = 0$ for some $\hat{u} \in \mathcal{U}_{\infty,\rho(t;\rho_0,u)}$. We then have

$$e^{sA}\left[e^{tA}x_0 + e^{sA}\int_0^t e^{(t-s)A}Bu(s)ds\right] + \int_0^s e^{(s-r)A}B\hat{u}(r)dr = 0$$

which means

$$e^{(s+t)A}x_0 + \int_0^{t+s} e^{(t+s-r)A}Bu^*(r)dr = 0,$$

where $u^* \in \mathcal{U}_{\infty,\rho_0}$ is defined as $u^*(r) = u(r)$ for $0 \le r \le t$ and $u^*(r) = \hat{u}(r-t)$ for $t \le r \le s+t$. This implies that $x_0 \in \mathcal{R}_{p,\rho_0}(s+t)$ and then

$$T_p(\rho_0, x_0) \le t + s \qquad \forall s \ge T_p(\rho(t;\rho_0,u), y(t;x_0,u)).$$

Hence

$$T_p(\rho_0, x_0) \le t + T_p(\rho(t;\rho_0,u), y(t;x_0,u)) \qquad \forall u \in \mathcal{U}_{\infty,\rho_0}.$$

This ends the first step. To prove the claim it remains to show that for every $\varepsilon > 0$ there exists $u \in \mathcal{U}_{\infty,\rho_0}$ such that

$$T_p(\rho_0, x_0) \ge t + T_p(\rho(t;\rho_0,u), y(t;x_0,u)) - \varepsilon \qquad \text{for } t \le T_p(\rho_0, x_0).$$

Fix $\varepsilon > 0$. By definition of T_p there exists $s < T_p(\rho_0, x_0) + \varepsilon$ such that $x_0 \in \mathcal{R}_{p,\rho_0}(s)$. Since $x_0 \in \mathcal{R}_{p,\rho_0}(s)$, we have

$$e^{sA}x_0 + \int_0^s e^{(s-r)A}Bv(r)dr = 0 \tag{24}$$

for some $v \in \mathcal{U}_{p,\rho_0}$. By assumption we have $t \le T_p(\rho_0, x_0)$. This implies $t \le s$. We consider $t < s$ (otherwise everything is clear) and we rewrite (24) as

$$e^{(s-t)A}e^{tA}x_0 + e^{(s-t)A}\int_0^t e^{(t-r)A}Bv(r)dr + \int_t^s e^{(s-r)A}Bv(r)dr = 0$$

which yields

$$e^{(s-t)A}y(t;x_0,v) + \int_0^{s-t} e^{(s-t-r)A}Bv(t+r)dr = 0.$$

Since we have $v(t+\cdot) \in \mathcal{U}_{p,\rho(t;\rho_0,v)}$ the latter implies that $y(t;x_0,v) \in \mathcal{R}_{p,\rho(t;\rho_0,v)}(s-t)$ and so that $T_p(\rho(t;\rho_0,v), y(t;x_0,v)) \le s - t$, i.e.,

$$t + T_p(\rho(t;\rho_0,v), y(t;x_0,v)) \le s \le T_p(\rho_0, x_0) + \varepsilon.$$

This ends the proof. $\square$

The key consequence of the above theorem is Theorem 4.8 which we prove below.

Proof of Theorem 4.8. Let $\rho_0 > 0$, $x_1, x_2 \in H$, and $T_p(\rho_0, x_1) > T_p(\rho_0, x_2)$. For simplicity we will prove the claim when $|x_1 - x_2| \leq 1/2$, as the estimate (19) for the case $|x_1 - x_2| > 1/2$ follows by simply using the local boundedness of $T_p(\rho_0, \cdot)$ on $\mathcal{R}_{p,\rho_0} \cap \{|x| \leq K\}$. Let t be such that $T_p(\rho_0, x_1) > t$. By the dynamic programming principle

$$T_p(\rho_0, x_1) \leq t + T_p(\rho(t; \rho_0, u), y(t; x_1, u)) \qquad \forall u \in \mathcal{U}_{p,\rho_0}. \tag{25}$$

Taking $t = T_p(\rho_0, x_2)$ in (25) we find

$$T_p(\rho_0, x_1) - T_p(\rho_0, x_2) \leq T_p\left(\rho(T_p(\rho_0, x_2); \rho_0, u), y\left(T_p(\rho_0, x_2); x_1, u\right)\right) \qquad \forall u \in \mathcal{U}_{p,\rho_0}$$

so that, using Theorem 4.6 and calling $g(r) = f^{-1}(\rho_0/r)$ (which is a strictly increasing function),

$$T_p(\rho_0, x_1) - T_p(\rho_0, x_2) \leq g\left(\frac{\rho_0 y\left(T_p(\rho_0, x_2); x_1, u\right)}{\rho(T_p(\rho_0, x_2); \rho_0, u)}\right) \qquad \forall u \in \mathcal{U}_{p,\rho_0}. \tag{26}$$

Now let $\lambda \in (0, 1)$, let u_2 be the optimal control for $T_p(\rho_0, x_2)$, and choose $u_\lambda = \lambda u_2$ in the above inequality. We then have, by easy calculations,

$$\rho(T_p(\rho_0, x_2); \rho_0, u_\lambda) = \rho_0[1 - \lambda^p]^{1/p} > \rho_0[1 - \lambda]^{1/p}. \tag{27}$$

Moreover, since $y(T_p(\rho_0, x_2); x_2, u_2) = 0$, we have

$$\int_0^{T_p(\rho_0,x_2)} e^{(T_p(\rho_0,x_2)-s)A} B u_2(s) ds = -e^{T_p(\rho_0,x_2)A} x_2$$

and so

$$\begin{aligned} y(T_p(\rho_0, x_2); x_1, u_\lambda) &= e^{T_p(\rho_0,x_2)A} x_1 - \lambda e^{T_p(\rho_0,x_2)A} x_2 \\ &\leq M e^{\omega T_p(\rho_0,x_2)} |x_1 - x_2| + (1 - \lambda) M e^{\omega T_p(\rho_0,x_2)} |x_2|. \end{aligned} \tag{28}$$

Hence, putting (27) and (28) into (26) we can write

$$T_p(\rho_0, x_1) - T_p(\rho_0, x_2) \leq g\left(\frac{M e^{\omega T_p(\rho_0,x_2)}[|x_1 - x_2| + (1 - \lambda)|x_2|]}{[1 - \lambda]^{1/p}}\right).$$

Setting $\lambda = 1 - |x_1 - x_2|$ we then have

$$T_p(\rho_0, x_1) - T_p(\rho_0, x_2) \leq g\left(M e^{\omega T_p(\rho_0,x_2)} [1 + |x_2|] |x_1 - x_2|^{1-1/p}\right)$$

which gives the claim by interchanging the roles of x_1 and x_2. $\square$

In the finite dimensional case an answer about the local Hölder continuity of T_p follows immediately from the following corollary and the result of Seidman and Yong [35, 37] recalled in Theorem 3.6.

Corollary 5.17. *Assume that the system* (1) *is p-null controllable* ($p \in (1, +\infty)$) *and that $H = \mathbb{R}^n$. Let k be the minimum integer such that the matrix $\left[B, AB, ..., A^k B\right]$ is full rank. Then the minimum time function $T_p(\rho, \cdot)$ is Hölder continuous of exponent $[1 + kq]^{-1}$, where $q^{-1} + p^{-1} = 1$.*

We now give a finite dimensional example where the exponent $[1+kq]^{-1}$ stated above is sharp.

Example 5.18. Let $\mathrm{n}=2$, $\mathcal{U}=\mathcal{U}_{2,1}$, and take the system (1) with

$$A=\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix}\ B=\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}.$$

We have rank $[B, AB]=2$ so 2-null controllability holds. By easy calculations we get that the control of minimum energy $u^*_{t,x}(r)=-B^*e^{(t-r)A^*}Q_t^{-1}e^{tA}x$ is given by

$$u^*_{t,x}(r)=-\frac{2}{t}\left[\left(2-\frac{3r}{t}\right)x_1+\frac{3}{t}\left(1-\frac{2r}{t}\right)x_2\right]$$

and the minimum energy by

$$|\Gamma(t)x|^2=\frac{4}{t^2}\left[tx_1^2+\frac{3}{t}x_2^2+3x_1x_2\right].$$

Now to compute $T_2(x)$ we solve $|\Gamma(T_2(x))x|^2=1$ which gives

$$T_2(x)=\frac{2^{\frac{2}{3}}}{3}\left[2^{\frac{4}{3}}x_1^2+2^{\frac{2}{3}}x_1(4x_1^3+9x_2)^{\frac{1}{3}}+(4x_1^3+9x_2)^{\frac{2}{3}}\right].$$

Now it is not hard to show that $T_2(x)$ is locally Hölder continuous with exponent 1/3. In fact, it is the sum of three functions: the first is locally Lipschitz continuous, the second is locally Hölder continuous with exponent 1/3, and the third is locally Hölder continuous with exponent 2/3. Moreover, $T_2(x)$ is not locally Hölder continuous with higher exponents because it can be easily verified that the second function is not locally Hölder continuous with exponents greater than 1/3.

Acknowledgments. The authors wish to thank P. Cannarsa for suggesting the idea of adding the new variable ρ to prove the uniform and Hölder continuities in the case $p\in(1,\infty)$; I. C. Dolcetta for suggesting the problem; A. Cutrí, V. Komornik, and J. Zabczyk for discussions and comments; and the referees for suggestions which improved the presentation of the paper.

REFERENCES

[1] A. Bacciotti, *Fondamenti Geometrici della Teoria della Controllabilitá*, Quaderno UMI n. 31, Pitagora, Bologna, 1986.

[2] V. Barbu, *The dynamic programming equation for the time-optimal control problem in infinite dimensions*, SIAM J. Control Optim., 29 (1991), pp. 445–456.

[3] V. Barbu, *The time optimal control of Navier-Stokes equations*, Systems Control Lett., 30 (1997), pp. 93–100.

[4] M. Bardi, *A boundary value problem for the minimum-time function*, SIAM J. Control Optim., 27 (1989), pp. 776–785.

[5] M. Bardi and P. Soravia, *Time-Optimal Control, Lie Brackets, and Hamilton-Jacobi Equations*, preprint.

[6] R. M. Bianchini and G. Stefani, *Time-optimal problem and time-optimal map*, Rend. Sem. Mat. Univ. Politec. Torino, 48 (1990), pp. 401–429.

[7] H. Brèzis, L. Nirenberg, and G. Stampacchia, *A remark on Ky Fan's minimax principle,* Boll. Un. Mat Ital. (4), 6 (1972), pp. 293–300.

[8] P. Cannarsa and C. Sinestrari, *Convexity properties of the minimum time function*, Calc. Var. Partial Differential Equations, 3 (1995), pp. 273–298.

[9] P. Cannarsa and C. Sinestrari, *On a class of minimum time problems*, Discrete Contin. Dynam. Systems, 1 (1995), pp. 285–300.

[10] R. Conti, *Controlli in tempo minimo*, Boll. Un. Mat. Ital. A (6), 2 (1983), pp. 271–288.

[11] O. Carja, *On constraint controllability of linear systems in Banach spaces*, J. Optim. Theory Appl., 56 (1988), pp. 215–225.

[12] O. Carja, *On continuity of the minimal time function for distributed control system*, Boll. Un. Mat. Ital. A (6), 4 (1985), pp. 293–302.

[13] O. Carja, *The minimal time function for vibrating systems*, in Differential Equations and Control Theory, V. Barbu, ed., Longman Scientific and Technical, Harlow, UK, 1991, pp. 58–62.

[14] O. Carja, *On the minimal time function for distributed control systems in Banach spaces*, J. Optim. Theory Appl., 44 (1984), pp. 397–406.

[15] O. Carja, *The minimal time function in infinite dimensions*, SIAM J. Control Optim., 31 (1993), pp. 1103–1114.

[16] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions,* Encyclopedia of Mathematics and Its Applications 44, Cambridge University Press, Cambridge, 1992.

[17] H. O. Fattorini, *The time optimal control problem in Banach spaces*, Appl. Math. Optim., 1 (1974), pp. 163–188.

[18] C. L. Fefferman and D. H. Phong, *Subelliptic eigenvalue problems*, in Proceedings of the Conference on Harmonic Analysis in Honour of A. Zygmund, Wadsworth Math. Ser., Wadsworth, Belmont, CA, 1983, pp. 590–606.

[19] E. Gyurkovics, *Hölder condition for the minimum time function of linear systems*, in System Modelling and Optimization, Lecture Notes in Control and Inform. Sci. 59, P. Thoft-Chistensen, ed., Springer-Verlag, Berlin, pp. 382–392.

[20] O. Hajek, *Geometric theory of time-optimal control*, SIAM J. Control, 9 (1971), pp. 339–350.

[21] H. Hermes and J. P. Lasalle, *Functional Analysis and Time Optimal Control,* Academic Press, New York, 1969.

[22] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory,* John Wiley, New York, 1967.

[23] A. A. Liverovskii, *A Hölder's condition for Bellman's functions*, Differensial'nye Uravneniya, 13 (1977), pp. 2180–2187 (in Russian; English translation in Differential Equations, 13 (1977), pp. 1521–1526).

[24] A. A. Liverovskii, *Some properties of the Bellman function for linear and symmetric polysystems*, Differensial'nye Uravneniya, 16 (1980), pp. 414–423 (in Russian; English translation in Differential Equations, 16 (1980), pp. 255–261).

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

[26] W. Krabs, *On time-minimal distributed control of vibrating systems governed by an abstract wave equation*, Appl. Math. Optim., 13 (1985), pp. 137–149.

[27] W. Krabs, *On time-minimal distributed control of vibrations*, Appl. Math. Optim., 19 (1989), pp. 65–73.

[28] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.

[29] N. V. Petrov, *The Bellman problem for a time-optimality problem*, Prikl. Mat. Meh., 34 (1970), pp. 820–826.

[30] T. Yu. Petrenko, *The Properties of the Bellman function*, Differensial'nye Uravneniya, 9 (1973), pp. 1244–1255 (in Russian).

[31] F. Rampazzo and C. Sartori, *The minimum time function with unbounded controls*, J. Math. Systems Estim. Control, 8 (1998), pp. 185–188.

[32] M. Ranguin, *Propriété höldérienne de la function temps minimal d'un systéme linéaire autonome*, RAIRO Automat., 16 (1982), pp. 329–340.

[33] M. Ranguin, *A propos de la propriété höldérienne de la function temps minimal d'un systéme linéaire autonome*, RAIRO Automat., 17 (1983), pp. 99–100.

[34] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[35] T. I. Seidman, *How violent are fast controls?*, Math. Control Signals Systems, 1 (1988), pp. 89–95.

[36] T. I. Seidman, *Two results on exact boundary control of parabolic equations*, Appl. Math. Optim., 11 (1984), pp. 145–152.

[37] T. I. Seidman and J. Yong, *How violent are fast controls?,* II, Math. Control Signals Systems, 9 (1996), pp 327–340.

[38] R. Tauraso, *Tempo Ottimale per Sistemi Lineari Autonomi in Spazi di Hilbert*, Tesi di Laurea, Università di Pisa, A.A. 1990–1991.

[39] J. Zabczyk, *Mathematical Control Theory: An Introduction*, Birkhäuser, Boston, 1992.

SIAM J. Control Optim.
Vol. 37, No. 4, pp. 1222–1250

ON THE CONVERGENCE AND APPLICATIONS OF GENERALIZED SIMULATED ANNEALING[*]

P. DEL MORAL[†] AND L. MICLO[†]

Abstract. The convergence of the generalized simulated annealing with time-inhomogeneous communication cost functions is discussed. This study is based on the use of log-Sobolev inequalities and semigroup techniques in the spirit of a previous article by one of the authors. We also propose a natural test set approach to study the global minima of the virtual energy. The second part of the paper is devoted to the application of these results. We propose two general Markovian models of genetic algorithms and we give a simple proof of the convergence toward the global minima of the fitness function. Finally we introduce a stochastic algorithm that converges to the set of the global minima of a given mean cost optimization problem.

Key words. simulated annealing, genetic algorithms, stochastic optimization

AMS subject classifications. 60J05, 92D15

PII. S0363012996313987

Introduction. Let E be a finite state space and q an irreducible Markov kernel. The main purpose of this paper is to study the limiting behavior of a large class of time-inhomogeneous Markov processes controlled by two parameters $(\gamma,\beta)\in\mathbb{R}_+^2$ and associated with a family of Markov kernels $Q_{\gamma,\beta}(x,y)$ having the following property:

$$\exists k>0\ :\quad k^{-1}\, q(x,y)e^{-\beta V_\gamma(x,y)} \leq Q_{\gamma,\beta}(x,y) \leq k\ \ q(x,y)e^{-\beta V_\gamma(x,y)}, \tag{1}$$

where $V:\mathbb{R}_+\times E^2\to\mathbb{R}_+\cup\{\infty\}$, $V_\gamma(x,y)<+\infty \iff q(x,y)>0$, and for any $x,y\in E$, $(\gamma,\beta)\to Q_{\gamma,\beta}(x,y)\in C^1$.

For a discussion on the origins of this problem the reader is referred to the introduction of Trouvé [12], who studies the asymptotic behavior of such chains, with time-homogeneous function $V(x,y)$, using large deviation techniques. The fundamental notions here are those of the log-Sobolev constant $a(\gamma,\beta)$ of $Q_{\gamma,\beta}$ and the relative entropy of a measure with respect to another measure. Other complementary results relating to time-inhomogeneous communication cost function can be found in Frigerio–Grillo [7], Younes [13], and more recently in Löwe [9], where Sobolev inequalities rather than log-Sobolev inequalities are used for classical models where q is assumed to be reversible and V_γ is associated with an a priori potential depending on γ.

For a probability measure m on E, inverse-freezing schedule $\beta\in C^1(\mathbb{R}_+,\mathbb{R}_+)$ and $\gamma\in C^1(\mathbb{R}_+,\mathbb{R}_+)$, we denote (Ω,P,F_t,X_t) as the canonical process associated with the family of generators $(L_{\gamma_t,\beta_t})_{t\geq 0}=(Q_{\gamma_t,\beta_t}-I)_{t\geq 0}$ whose initial condition is $m_0=m$, and we denote m_t as the distribution of X_t.

The aim of section 1.1 is to give several conditions on the rate of increase of $\gamma_t,\beta_t\to+\infty$ to ensure the entropy of m_t with respect to π_{γ_t,β_t} converges to 0. We shall examine as much of the theory as possible in a form applicable to general optimization problems and applicable in particular to mean cost optimization problems.

[*] Received by the editors December 20, 1996; accepted for publication (in revised form) July 24, 1998; published electronically May 26, 1999. This research was partly supported by INTAS-RFBR 95-0091 and EU TMR programme project ERBF MRX CT 960075A.

http://www.siam.org/journals/sicon/37-4/31398.html

[†] Laboratoire de Statistique et Probabilités, CNRS-UMR C5583, Bat 1R1, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse Cedex, France (delmoral@cict.fr, miclo@cict.fr).

To illustrate our results we will restrict attention to various special classes of generalized simulated annealing. We will commence with a detailed analysis of general Markov kernels of the form

$$Q_\beta(x,y) = \sum_{u \in U} \overline{q}_\beta(x,u,y) e^{-\beta \overline{V}(x,u,y)}, \tag{2}$$

where U is a given finite set, $\overline{V} : E \times U \times E \to \mathbb{R}_+$, and $\overline{q}_\beta : E \times U \times E \to \mathbb{R}_+$, $\beta \in \mathbb{R}_+$, is a family of functions satisfying some continuity and irreducibility conditions. This situation can be formulated in the general form (1). We will settle this question and provide the explicit computation of the corresponding communication cost function V. In a final stage we will give several conditions on the rate of decrease of the cooling schedule to ensure the convergence in probability of the corresponding canonical process X_t, as $t \to +\infty$, to the set of global minima of the virtual energy associated with V.

Another application is the situation in which the Markov kernel $Q_{\gamma,\beta}$ has the form

$$Q_{\gamma,\beta}(x,y) = q_\beta(x,y)\ e^{\beta V_\gamma(x,y)}$$

with

$$\lim_{\gamma \to +\infty} V_\gamma(x,y) = V(x,y), \qquad \lim_{\beta \to +\infty} q_\beta(x,y) = q(x,y)$$

for some Markov kernel q and some function $V : E \times E \to \mathbb{R}^+$. In this situation, let π_β be the unique invariant probability measure of the Markov generator $L_\beta = Q_\beta - I$, where

$$Q_\beta(x,y) = q(x,y)\ e^{-\beta\, V(x,y)}.$$

We will give several conditions on the rate of decrease of the cooling schedule and on the rate of convergence $\lim_{t \to +\infty} V_{\gamma_t} = V$ to ensure the entropy of m_t with respect to π_{β_t} converges to 0.

The above results imply that the canonical process X_t converges in law to the set of the global minima $V^\star$ of a virtual energy V. This leads us to investigate more closely the properties of such function. Section 1.2 introduces a natural test set approach to study $V^\star$. Specifically, we will give a condition for a given subset $H \subset E$ to contain $V^\star$.

Section 2 is devoted to application of these results, an area of which is the situation in which Q_β is the transition probability kernel of a genetic algorithm. Such algorithms can be formulated by a Markov process with state space $E = S^N$ ($N > 1$ and S a finite set) and whose transition probabilities Q_β includes a mutation transition $Q_\beta^{(1)}$ and a selection mechanism $Q_\beta^{(2)}$. More precisely the mutation transition is modeled by independent motion of each particle and the selection mechanism chooses randomly in the previous population according to a given fitness function. The first convergence result was obtained by Cerf [2] in the case in which $Q_\beta = Q_\beta^{(1)} Q_\beta^{(2)}$ and the mutations vanish, that is, $\lim_{\beta \to +\infty} Q_\beta^{(1)}(x,y) = 1_x(y)$.

In section 2.1 we will use the results of section 1.1 and the test set approach introduced in section 1.2 to derive a new and simple proof of the convergence in probability of such algorithms to the set of the global minima of the fitness function in the following situations:

$$Q_\beta = Q_\beta^{(1)} Q_\beta^{(2)} \quad \text{and} \quad Q_\beta = \alpha\, Q_\beta^{(1)} + (1-\alpha)\, Q_\beta^{(2)}, \quad 0 < \alpha < 1.$$

Finally, in subsection 2.2 we will apply the results of the first section to mean cost optimization problems. However, here we touch upon a slightly different aspect of the theory. Namely, the object will be to find the global minima of a function $U : E \to \mathbb{R}_+$ defined by

$$U(x) = E(L(Z, x)),$$

where Z is a random variable taking values in a finite set F and $L : F \times E \to \mathbb{R}_+$. We will solve this optimization problem by an original method based on the use of Monte Carlo simulations coupled with simulated annealing. This special case will require specific developments because the corresponding function V_γ will necessarily behave as a random process. We will present a time-inhomogeneous Markov process which converges to the global minima of U.

1. General results.

The purpose of this section is to study the limiting behavior of time-inhomogeneous Markov chains controlled by two parameters $(\gamma, \beta) \in \mathbb{R}_+^2$ and associated with a family of Markov kernels $Q_{\gamma,\beta}(x, y)$ having the property (1), with the assumptions given in the introduction. This is in keeping with our second objective, which is to introduce some areas in which such results are useful.

The reader who is especially interested in genetic algorithms has to consult Corollary 1 and Propositions 3 and 4. Finally, the numerical solving of mean cost optimization problems requires only the use of Theorem 2 or Corollary 4.

1.1. Relative entropy convergence.

Our analysis will be based entirely on considerations of the time-continuous semigroup associated with the Markov kernels $Q_{\gamma,\beta}(x, y)$ introduced in (1). Namely, define for $f : E \to \mathbb{R}$

$$L_{\gamma,\beta} f(x) = \sum_{y \in E} (f(y) - f(x))\, Q_{\gamma,\beta}(x, y).$$

For a probability measure m on E, an inverse-freezing schedule $\beta \in C^1(\mathbb{R}_+, \mathbb{R}_+)$, and $\gamma \in C^1(\mathbb{R}_+, \mathbb{R}_+)$, we denote (Ω, P, F_t, X_t) as the canonical process associated with the family of generators $(L_{\gamma_t,\beta_t})_{t \geq 0} = (Q_{\gamma_t,\beta_t} - I)_{t \geq 0}$ whose initial condition is $m_0 = m$, and we write m_t the distribution of X_t.

Whenever X is time-homogeneous (i.e., $\beta_t = \beta$ and $\gamma_t = \gamma$) it is well known that $L_{\gamma,\beta}$ has a unique invariant probability measure $\pi_{\gamma,\beta}$ so that

$$\forall f : E \to \mathbb{R} \qquad \pi_{\gamma,\beta}(L_{\gamma,\beta} f) = 0$$

and $\pi_{\gamma,\beta}$ charges all the points. It is also convenient to recall the notion of log-Sobolev constant of $Q_{\gamma,\beta}$. Namely,

$$a(\gamma, \beta) \overset{\text{def}}{=} \min \left\{ \mathcal{E}_{\gamma,\beta}(f, f) / \mathcal{L}_{\gamma,\beta}(f),\ \mathcal{L}_{\gamma,\beta}(f) \neq 0 \right\},$$

where the Dirichlet form $\mathcal{E}_{\gamma,\beta}$ and $\mathcal{L}_{\gamma,\beta}$ are defined by

$$\mathcal{E}_{\gamma,\beta}(f, g) = -\langle L_{\gamma,\beta} f, g \rangle_{\pi_{\gamma,\beta}} = -\sum_{x \in E} L_{\gamma,\beta} f(x)\, g(x)\, \pi_{\gamma,\beta}(x),$$

$$\mathcal{L}_{\gamma,\beta}(f) = \sum_{x \in E} f(x)^2 \log \left(f(x)^2 / \|f\|_{2,\pi_{\gamma,\beta}}^2 \right) \pi_{\gamma,\beta}(x).$$

Let us recall the notion of relative entropy of a measure m with respect to a measure π charging all the points

$$\mathrm{Ent}_{\pi}(m) = \sum_{x\in E} m(x)\ \log\left(m(x)/\pi(x)\right).$$

Using this notation, whenever X is time-homogeneous, one has the following basic inequality (for instance, see Miclo [10]):

$$\frac{d}{dt}\mathrm{Ent}_{\pi_{\gamma,\beta}}(m_t) \leq -2\ a(\gamma,\beta)\ \mathrm{Ent}_{\pi_{\gamma,\beta}}(m_t). \tag{3}$$

For an expository paper on log-Sobolev constants for the general Markov chain on finite spaces the reader is referred to Diaconis–Saloff-Costes [5]. Holley and Stroock [8] use Sobolev and log-Sobolev inequalities to study the standard simulated annealing. For another approach using only spectral gap estimates the reader should consult [4]. Using log-Sobolev inequalities, one of the authors addressed the convergence of a simulated annealing associated with a Markov transition kernel of the form using the entropy distance to stationarity (see [10]). The purpose of this section is to extend these results to general Markov transition kernels of form (1).

What follows is an exposition of some basic results regarding the description of $\pi_{\gamma,\beta}$, by Bott–Mayberry [1] and also exposed in Freidlin–Wentzell [6]. For $x \in E$ we denote $G_E(x)$, or $G(x)$ when there are no possible confusions, as the set of x-graphs. We shall also use the following notations for $x \in E$ and $g \in G(x)$:

$$R_{\gamma,\beta}(x) = \sum_{g\in G(x)} Q_{\gamma,\beta}(g), \qquad Q_{\gamma,\beta}(g) = \prod_{(y\to z)\in g} Q_{\gamma,\beta}(y,z),$$
$$V_{\gamma}(g) = \sum_{(y\to z)\in g} V_{\gamma}(y,z), \qquad \mathbf{Q}_{\gamma,\beta}(x,y) = q(x,y)\ e^{-\beta\, V_{\gamma}(x,y).}$$

Then, whenever X is time-homogeneous, its invariant distribution $\pi_{\gamma,\beta}$ is given by

$$\pi_{\gamma,\beta}(x) = R_{\gamma,\beta}(x)/\sum_{z\in E} R_{\gamma,\beta}(z).$$

Similarly, let $\mu_{\gamma,\beta}$ be the invariant probability measure of

$$\mathbf{L}_{\gamma,\beta} f(x) = \sum_{y\in E} (f(y) - f(x))\ \mathbf{Q}_{\gamma,\beta}(x,y).$$

If $\mathbf{a}(\gamma,\beta)$ is the log-Sobolev constant of $\mathbf{Q}_{\gamma,\beta}$, then under assumption (1) there exists some constant $k_1 > 0$ such that

$$k_1^{-1}\mu_{\gamma,\beta}(x) \leq \pi_{\gamma,\beta}(x) \leq k_1 \mu_{\gamma,\beta}(x). \tag{4}$$

Now, from the very definition of $\mu_{\gamma,\beta}$ and (4), we have the estimate

$$-\beta^{-1}\ \log \pi_{\gamma,\beta}(x) \xrightarrow[\beta\,\to\,+\infty]{} V_{\gamma}(x) - \min_{z\in E} V_{\gamma}(z) \quad \text{with} \quad V_{\gamma}(x) \overset{\text{def}}{=} \min_{g\in G(x)} V_{\gamma}(g).$$

As a direct consequence of Lemma 3.3, Diaconis–Saloff-Costes [5], and the inequalities (1) and (4) there exists some constant $B > 0$ such that

$$B^{-1}\ \mathbf{a}(\gamma,\beta) \leq a(\gamma,\beta) \leq B\ \mathbf{a}(\gamma,\beta).$$

Finally, by Theorem 3.23 of Holley–Stroock [8] and the inequalities stated in Miclo [10], we have the following proposition.

PROPOSITION 1. *There exists some constant $A > 0$ such that $a(\gamma,\beta) \geq A\,\frac{e^{-\beta\,c(\gamma)}}{1+\beta}$, where $c(\gamma)$ is the critical height associated with the communication cost V_γ given by*

$$c(\gamma) = \max_{x,y\in E}\left(\min_{p\in\mathcal{S}_{x,y}} e_\gamma(p) - V_\gamma(x) - V_\gamma(y)\right) + \min_{z\in E} V_\gamma(z),$$
$$e_\gamma(p) = \max_{1\leq i\leq n} \min\left(V_\gamma(p_{i-1}) + V_\gamma(p_{i-1},p_i), V_\gamma(p_i) + V_\gamma(p_i,p_{i-1})\right),$$

where $\mathcal{S}_{x,y}$ is the set of all the finite sequences from x to y and $e_\gamma(p)$ denotes elevation of a path. It can also be shown that

$$c(\gamma) = \max_{x,y\in E}\left(\min_{p\in C_{x,y}} \tilde{e}_\gamma(p) - V_\gamma(x) - V_\gamma(y)\right) + \min_{z\in E} V_\gamma(z),$$

where

$$\tilde{e}_\gamma(p) = \max_{1\leq i\leq n} V_\gamma(p_{i-1}) + V_\gamma(p_{i-1},p_i)$$

and $C_{x,y}$ is the set of all paths (admissible for q) from x to y.

By choosing $t \to (\gamma_t,\beta_t)$ to go to infinity and using (3) we arrive at

$$\frac{d}{dt}\mathrm{Ent}_{\pi_{\gamma_t,\beta_t}}(m_t) \leq -2\,A\,\frac{e^{-\beta_t\,c(\gamma_t)}}{1+\beta_t}\mathrm{Ent}_{\pi_{\gamma_t,\beta_t}}(m_t) - \sum_{x\in E} m_t(x)\frac{d}{dt}\log\pi_{\gamma_t,\beta_t}(x). \tag{5}$$

Therefore it remains to estimate the derivatives $d\pi_{\gamma,\beta}/d\beta$ and $d\pi_{\gamma,\beta}/d\gamma$. For this purpose, write

$$\overline{Q}_{\gamma,\beta}(g) = Q_{\gamma,\beta}(g)/\sum_{h\in I_x} Q_{\gamma,\beta}(h), \qquad I_x = \left\{g\in G(x)\ :\ \prod_{(y\to z)\in g} q(y,z) > 0\right\}.$$

By a simple analysis it is easily checked that

$$\frac{d}{d\beta}\log R_{\gamma,\beta}(x) = \sum_{g\in I_x} \overline{Q}_{\gamma,\beta}(g)\frac{d}{d\beta}\log Q_{\gamma,\beta}(g),$$
$$\frac{d}{d\beta}\log \pi_{\gamma,\beta}(x) = \sum_{z\in E}\left(\frac{d}{d\beta}\log R_{\gamma,\beta}(x) - \frac{d}{d\beta}\log R_{\gamma,\beta}(z)\right)\pi_{\gamma,\beta}(z).$$

In order to derive a useful inequality we assume there exist two functions $d_1,\ d_2 : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\sup_{g\in\cup_x I_x}\frac{d}{d\beta}\log Q_{\gamma,\beta}(g) - \inf_{g\in\cup_x I_x}\frac{d}{d\beta}\log Q_{\gamma,\beta}(g) \leq d_1(\gamma), \tag{6}$$

$$\sup_{g\in\cup_x I_x}\frac{d}{d\gamma}\log Q_{\gamma,\beta}(g) - \inf_{g\in\cup_x I_x}\frac{d}{d\gamma}\log Q_{\gamma,\beta}(g) \leq d_2(\beta). \tag{7}$$

Note that by the very definitions of the sets I_x it clearly suffices to have

$$\left|\frac{d}{d\beta}\log Q_{\gamma,\beta}(x,y)\right| \leq \tilde{d}_1(\gamma), \qquad \left|\frac{d}{d\gamma}\log Q_{\gamma,\beta}(x,y)\right| \leq \tilde{d}_2(\beta)$$

for every $q(x,y) > 0$ and two functions $\tilde{d}_1,\ \tilde{d}_2 : \mathbb{R}_+ \to \mathbb{R}_+$. Now we are ready to apply the following technical lemma.

LEMMA 1 (Stroock [11]). *If μ is a probability measure and $\theta \in L^1(\mu)^+$ satisfies $\int \theta d\mu = 1$, then for every $\phi \in L^\infty(\mu)$ satisfying $\int \phi d\mu = 0$ one has*

$$\left| \int \phi\, \theta\, d\mu \right| \leq \sqrt{2}\, \|\phi\|_{L^\infty(\mu)} \left(\int \theta\, \log \theta\, d\mu \right)^{1/2}.$$

Using this lemma with $\mu = \pi_{\gamma,\beta}$, $\phi = d \log \pi_{\gamma,\beta}/d\beta$ (resp., $\phi = d \log \pi_{\gamma,\beta}/d\gamma$), and $\theta = m_t/\pi_{\gamma_t,\beta_t}$ and writing

$$I_t \overset{\mathrm{def}}{=} (\mathrm{Ent}_{\pi_{\gamma_t,\beta_t}}(m_t))^{1/2}$$

we see from (5) that

$$\frac{dI_t}{dt} \leq -A_0\, \frac{e^{-\beta_t\, c(\gamma_t)}}{1+\beta_t} I_t + A_1\, d_1(\gamma_t) \left| \frac{d\beta_t}{dt} \right| + A_2\, d_2(\beta_t) \left| \frac{d\gamma_t}{dt} \right| \tag{8}$$

for some constants $A_0, A_1, A_2 > 0$. Hence by taking, for t sufficiently large, an inverse-temperature schedule of the form $\beta_t = K^{-1} \log t$ we obtain

$$\frac{dI_t}{dt} \leq -A_t\, I_t + B_t$$

with

$$A_t = A_0\, t^{-c(\gamma_t)/K} (1 + K^{-1} \log t)^{-1} \quad \text{and} \quad B_t = A_1 \frac{d_1(\gamma_t)}{Kt} + A_2\, d_2(K^{-1} \log t) \left| \frac{d\gamma_t}{dt} \right|.$$

Now, it is well known that

$$\exists t_0 \in \mathbb{R}_+ \quad \int_{t_0}^{+\infty} A_s\, ds = +\infty, \qquad \lim_{t \to +\infty} \frac{B_t}{A_t} = 0 \Longrightarrow \lim_{t \to +\infty} I_t = 0.$$

We can now summarize the entire consideration in the following way.

THEOREM 1. *Assume that $c = \limsup_{\gamma \to +\infty} c(\gamma) < +\infty$ and the conditions*

$$\sup_{x \in E} \left| \frac{d}{d\beta} \mu_{\gamma,\beta}(x) \right| \leq d_1(\gamma), \qquad \sup_{x \in E} \left| \frac{d}{d\gamma} \mu_{\gamma,\beta}(x) \right| \leq d_2(\beta)$$

are satisfied for two nonnegative functions $d_1,\ d_2$.

When the inverse-freezing schedule has parametric form $\beta_t = K^{-1} \log t$, for t sufficiently large and $K > c$, we have

$$\lim_{t \to +\infty} \mathrm{Ent}_{\pi_{\gamma_t,\beta_t}}(m_t) = 0 \tag{9}$$

whenever

$$\frac{d\gamma_t}{dt} = o\left(1/(d_2(\log t^{1/K}) t^{c/K} \log t) \right), \qquad d_1(\gamma_t)/t = o\left(1/(t^{c/K} \log t) \right). \tag{10}$$

Remark. The hypothesis above might seem difficult to check; we shall see in fact that in many cases it is indeed fulfilled. In practice the Markov kernels $Q_{\gamma,\beta}$ are known and it clearly suffices to check that

$$\sup_{g\in\cup_x I_x}\frac{d}{d\beta}\log Q_{\gamma,\beta}(g)-\inf_{g\in\cup_x I_x}\frac{d}{d\beta}\log Q_{\gamma,\beta}(g)\leq d_1(\gamma),$$

$$\sup_{g\in\cup_x I_x}\frac{d}{d\gamma}\log Q_{\gamma,\beta}(g)-\inf_{g\in\cup_x I_x}\frac{d}{d\gamma}\log Q_{\gamma,\beta}(g)\leq d_2(\beta).$$

Because of the general orientation provided in this section we can proceed immediately to review the most important special class of generalized simulated annealing, which we shall study later. Let us consider the transition probability kernels

$$Q_\beta(x,y)=\sum_{u\in U}\overline{q}_\beta(x,u,y)e^{-\beta\overline{V}(x,u,y)}, \tag{11}$$

where U is a given finite set, $\overline{V}:U\times E^2\to\mathbb{R}_+$, and $\overline{q}_\beta(x,u,y)\geq 0$. The proof of Theorem 1 shows it is important that the Markov kernel q is irreducible. For this purpose we will assume the existence of nonnegative functions $\overline{q}(x,u,y)$ so that

$$(\mathcal{C})\qquad \overline{q}_\beta(x,u,y)>0\Longleftrightarrow\overline{q}(x,u,y)>0,\qquad \lim_{\beta\to+\infty}\overline{q}_\beta(x,u,y)=\overline{q}(x,u,y),$$

and we will work with the following irreducibility condition:

$$(\mathcal{I})\,\forall x,y\in E\quad \exists(p_k,u_k)_{1\leq k\leq r}:\ \ p_0=x,\quad p_k\in E\ u_k\in U\ p_r=y\ \overline{q}(p_k,u_k,p_{k+1})>0.$$

Our immediate goal is to prove the following consequence of Theorem 1 which gives conditions assuring the convergence stated in Theorem 1 for a time-inhomogeneous Markov chain with transitions (11). This corollary is applied in section 2 to the convergence of genetic algorithms

COROLLARY 1. *Assume $\overline{q}_\beta$ and $\overline{q}$ satisfy the continuity and irreducibility conditions* ($\mathcal{C}$) *and* ($\mathcal{I}$). *Then, the transition probabilities*

$$Q_\beta(x,y)=\sum_{u\in U}\overline{q}_\beta(x,u,y)e^{-\beta\overline{V}(x,u,y)} \tag{12}$$

satisfy the inequalities (1) *with*

$$V(x,y)=\min_{u\in U(x,y)}\overline{V}(x,u,y),\qquad U(x,y)=\{u\in U\ :\ \overline{q}(x,u,y)>0\},$$

$$q(x,y)=\sum_{u\in U^\star(x,y)}\overline{q}(x,u,y),\qquad U^\star(x,y)=\{u\in U(x,y)\ :\ \overline{V}(x,u,y)=V(x,y)\}.$$

Let V be the virtual energy function corresponding to the above communication cost function

$$V(x)=\min_{g\in G(x)}\sum_{(y\to z)}V(y,z)$$

and let c be the corresponding critical height. In addition, suppose that for every $\overline{q}(x,u,y)>0$ and for some $\beta_0\geq 0$

$$\sup_{\beta\geq\beta_0}\left|\frac{d\log\overline{q}_\beta}{d\beta}(x,u,y)\right|<+\infty. \tag{13}$$

In this case, if m is a probability measure on E and if β_t assumes the parametric form $\beta_t = K^{-1}\log t$, for sufficiently large t, with $K > c$, then

$$\lim_{t\to+\infty} Ent_{\pi_{\beta_t}}(m_t) = 0 \qquad \textit{and} \qquad \lim_{t\to+\infty} P(X_t \in V^\star) = 1,$$

where (Ω, P, F_t, X_t) is the canonical process associated with the family of generators $(L_{\beta_t})_{t\geq 0} = (Q_{\beta_t} - I)_{t\geq 0}$ whose initial condition is $m_0 = m$, m_t is the distribution of X_t, π_β is the unique invariant probability measure of L_β, and

$$V^\star = \{x \in E \ : \ V(x) = \min_E V\}.$$

Proof. Using the form of $\overline{q}_\beta$ we have for some suitable function $\epsilon(\beta) \to 0$, as $\beta \to +\infty$,

$$(1 - \epsilon(\beta))\, K_\beta(x,y) \leq Q_\beta(x,y) \leq (1 + \epsilon(\beta))\, K_\beta(x,y), \tag{14}$$

where

$$K_\beta(x,y) = \sum_{u\in U(x,y)} \overline{q}(x,u,y)e^{-\beta\overline{V}(x,u,y)}.$$

But now we have that

$$\begin{aligned} K_\beta(x,y) = &\sum_{u\in U^\star(x,y)} \overline{q}(x,u,y)\; e^{-\beta V(x,y)} \\ &+ e^{-\beta V(x,y)} \sum_{u\in U(x,y)-U^\star(x,y)} \overline{q}(x,u,y)\; e^{-\beta(\overline{V}(x,u,y)-V(x,y))} \\ = q(x,y)&e^{-\beta V(x,y)} + e^{-\beta V(x,y)} \sum_{u\in U(x,y)-U^\star(x,y)} \overline{q}(x,u,y)\; e^{-\beta(\overline{V}(x,u,y)-V(x,y))}. \end{aligned}$$

Thus, condition $(\mathcal{I})$ implies that q is irreducible. Furthermore, if we write

$$\begin{aligned} I &= \{(x,y) \in E^2 \ : \ U(x,y) \neq \emptyset\}, \\ J &= \{(x,u,y) \in E \times U \times E \ : \ (x,y) \in I \ \ u \in U(x,y)\} \end{aligned}$$

and

$$\begin{aligned} h_1 &= \min_{(x,y)\in I} \sum_{u\in U(x,y)-U^\star(x,y)} \overline{q}(x,u,y)/q(x,y), \\ h_2 &= \min_{(x,u,y)\,:\, u\notin U^\star(x,y)} \overline{V}(x,u,y) - V(x,y), \end{aligned}$$

then using (14) we get the system of inequalities

$$(1-\epsilon(\beta))\; q(x,y)\; e^{\beta V(x,y)} \leq Q_\beta(x,y) \leq (1+\epsilon(\beta))\; (1 + h_1\; e^{-\beta h_2})q(x,y)\; e^{\beta V(x,y)}.$$

To end the proof it remains to check condition (6). Choose $q(x,y) > 0$; after some computations we find

$$\left|\frac{d\log Q_\beta(x,y)}{d\beta}\right| \leq \sup_{u\in U(x,y)} \left|\frac{d\log \overline{q}_\beta}{d\beta}(x,u,y)\right| + \sup_{u\in U(x,y)} \overline{V}(x,u,y).$$

Thus (13) implies (6), and using Theorem 1 the proof of the first assertion is complete. Examination of the invariant distribution of L_β soon yields that $\forall x \notin V^\star$ we have $\lim_{\beta\to+\infty}\pi_\beta(x) = 0$. Then, to prove the last assertion it is enough to recall the basic inequality

$$\|m_t - \pi_{\beta_t}\|_{TV}^2 \leq 2\,\mathrm{Ent}_{\pi_{\beta_t}}(m_t),$$

where $\|.\|_{TV}$ is the distance in total variation given by

$$\|\mu - \nu\|_{TV} = 2\sup_{A\subset E}|\mu(A) - \nu(A)|. \qquad \square$$

It is quite clear from definition (12) that the following situations are covered:

$$Q_\beta = Q_\beta^{(1)}Q_\beta^{(2)} \quad\text{and}\quad Q_\beta = \alpha\,Q_\beta^{(1)} + (1-\alpha)\,Q_\beta^{(2)}, \quad 0<\alpha<1.$$

In section 2 we will develop properties of both class of chains which we shall find includes, as a special case, the evolutionary processes studied by Cerf in [2]. For the sake of unity and to highlight issues specific to evolutionary processes, we give some examples to suggest how these results translate in this special situation.

Examples.

1. If $d_1(\gamma) = \gamma^p$ and $d_2(\beta) = \beta^q$ for some $p \geq 0$ and $q > 0$, it clearly suffices to choose $\gamma_t = \log t$.
2. Let us study a way to combine the transitions $Q_\beta^{(1)}, \dots, Q_\beta^{(r)}$ given by

$$Q_\beta^{(k)}(x,y) = q_\beta^{(k)}(x,y)\,e^{-\beta V^{(k)}(x,y)}, \qquad V^{(k)} : E^2 \to \mathbb{R}_+, \qquad q_\beta^{(k)}(x,y) \geq 0.$$

It is clear from (11) that the following situation is covered:

$$\text{(a)}\ \ Q_\beta(x,y) = Q_\beta^{(1)}\dots Q_\beta^{(r)}(x,y) = \sum_{z_1,\dots,z_{r-1}\in E} Q_\beta^{(1)}(x,z_1)\dots Q_\beta^{(r)}(z_{r-1},y).$$

This situation can be formulated in the form (11) with $U = E^{r-1}$ and

$$\overline{q}_\beta(x,u,y) = q_\beta^{(1)}(x,u_1)\dots q_\beta^{(r)}(u_{r-1},y),$$
$$\overline{V}(x,u,y) = V^{(1)}(x,u_1) + \cdots + V^{(r)}(u_{r-1},y).$$

Also from (11), the following situation is covered:

$$\text{(b)}\ \ \tilde{Q}_\beta(x,y) = \sum_{k=1}^r \alpha_k\,Q_\beta^{(k)}(x,y) \qquad\text{with}\qquad \sum_{k=1}^r \alpha_k = 1.$$

This situation can be formulated in the form (11) with $E = \{1,\dots,r\}$ and

$$\overline{q}_\beta(x,u,y) = \alpha_u\,q_\beta^{(u)}(x,y), \qquad \overline{V}(x,u,y) = V^{(u)}(x,y).$$

Probabilistically and in precise language 1(a) has the interpretation of being the transition of a chain obtained through overlapping $r-1$ other chains, and 1(b) has the interpretation of being the transition of a chain obtained through choosing randomly at each step among r chains. Let us remark, by way of illustration, that it is also possible to consider a way of combining 1(a) and 1(b) that subsumes such parallel and series combinations. For instance, each transition probability $Q_\beta^{(k)}$ in the expression (a) may be of type (b) and conversely. As a result one has a great freedom in the design and the physical construction of the transition probabilities $Q_{\gamma,\beta}$, and they appear ideally suited to describe a large class of processes encountered in applications.

3. Let us examine the above example when $r = 2$. In this situation we introduce the irreducibility condition $(\mathcal{I})'$

$$q^{(1)} \quad \text{irreducible} \quad \text{and} \quad \forall x \in E \qquad q^{(2)}(x,x) > 0 \tag{15}$$

and the continuity condition $(\mathcal{C})'$

$$\lim_{\beta \to +\infty} q_\beta^{(k)}(x,y) = q^{(k)}(x,y), \quad q_\beta^{(k)}(x,y) = 0 \Longleftrightarrow q^{(k)}(x,y) = 0 \quad \forall k = 1, 2, \tag{16}$$

where $q^{(k)}(x,y)$ are transition probability kernels such that

$$q^{(k)}(x,y) = 0 \Longleftrightarrow V^{(k)}(x,y) = +\infty \qquad \forall k = 1, 2.$$

In this situation the conditions $(\mathcal{C})$ and $(\mathcal{I})$ of Corollary 1 are satisfied. In addition, if we assume that for every $k = 1, 2$ and $q^{(k)}(x,y)$

$$\left| \frac{d \log q_\beta^{(2)}}{d\beta}(x,y) \right| < +\infty,$$

then the last condition (13) introduced in Corollary 1 is satisfied.

COROLLARY 2. *Suppose the Markov kernel of the chain has the form*

$$Q_{\gamma,\beta} = q_\beta(x,y) e^{-\beta V_\gamma(x,y)}$$

with

$$q_\beta(x,y) = 0 \Longleftrightarrow q(x,y) = 0, \qquad \lim_{\beta \to +\infty} q_\beta(x,y) = q(x,y)$$

and assume the following conditions are satisfied for every $\beta, \gamma \in \mathbb{R}_+$ and some constant $d > 0$:

$$c = \sup_\gamma c(\gamma) < +\infty, \qquad \sup_{g \in \cup_x I_x} V_\gamma(g) - \inf_{g \in \cup_x I_x} V_\gamma(g) \le d,$$

$$\sup_{g \in \cup_x I_x} \frac{d}{d\beta} \log q_\beta(g) - \inf_{g \in \cup_x I_x} \frac{d}{d\beta} \log q_\beta(g) \le d \sup_{g \in \cup_x I_x} \frac{d}{d\gamma} V_\gamma(g) - \inf_{g \in \cup_x I_x} \frac{d}{d\gamma} V_\gamma(g) \le d.$$

When the inverse freezing schedule has parametric form $\beta_t = K^{-1} \log t$, for t sufficiently large and $K > c$, we have

$$\lim_{t \to +\infty} \mathrm{Ent}_{\pi_{\gamma_t, \beta_t}}(m_t) = 0 \qquad \textit{whenever} \quad \frac{d\gamma_t}{dt} = o(1/(t^{c/K} \log^2 t)).$$

As a consequence we have the well-known corollary that follows.

COROLLARY 3 (Miclo [10]). *Suppose the Markov kernel of the chain has the form*

$$Q_{\gamma,\beta} = q(x,y) e^{-\beta V(x,y)} \overset{\text{def}}{=} Q_\beta(x,y).$$

When the inverse freezing schedule has parametric form $\beta_t = K^{-1} \log t$, for t sufficiently large and $K > c$, we have

$$\lim_{t \to +\infty} \mathrm{Ent}_{\pi_{\beta_t}}(m_t) = 0,$$

where π_β is the unique invariant probability measure of the Markov generator $Q_\beta - I$.

The usefulness of Theorem 1 will now be illustrated in the case where the transition probabilities $Q_{\gamma,\beta}$ converge to a transition probability kernel Q_β as $\gamma \to +\infty$.

THEOREM 2. *Let $Q_\beta(x,y)$ be a Markov kernel such that $Q_\beta(x,y) = 0 \iff q(x,y) = 0$. Suppose the assumptions of Theorem 1 are satisfied and for every $q(x,y) > 0$*

$$\lim_{t\to+\infty} |\log Q_{\gamma_t,\beta_t}(x,y) - \log Q_{\beta_t}(x,y)| = 0. \tag{17}$$

Then

$$\lim_{t\to+\infty} \mathrm{Ent}_{\pi_{\beta_t}}(m_t) = 0, \tag{18}$$

where π_β is the unique invariant probability measure of the Markov generator $Q_\beta - I$.

Proof. By the same line of argument as before π_β may be described as follows:

$$\pi_\beta(x) = \frac{R_\beta(x)}{\sum_{z\in E} R_\beta(z)} \quad \text{with} \quad R_\beta(x) = \sum_{g\in I_x} Q_\beta(g) \quad \text{and} \quad Q_\beta(g) = \prod_{(y\to z)\in g} Q_\beta(y,z). \tag{19}$$

It follows that

$$\begin{aligned}\log\frac{\pi_{\gamma,\beta}}{\pi_\beta}(x) &= -\log\left(\sum_{g\in I_x}\prod_{(y\to z)\in g}(Q_\beta Q_{\gamma,\beta}^{-1})(y,z)\,\overline{Q}_{\gamma,\beta}(g)\right)\\ &\quad -\log\left(\sum_{x'\in E, g\in I_{x'}}\prod_{(y\to z)\in g}(Q_{\gamma,\beta}Q_\beta^{-1})(y,z)\,\tilde{Q}_\beta(g)\right),\end{aligned}$$

where

$$\overline{Q}_{\gamma,\beta}(g) = Q_{\gamma,\beta}(g)/\sum_{h\in I_x} Q_{\gamma,\beta}(h) \qquad \text{and} \qquad \tilde{Q}_\beta(g) = Q_\beta(g)/\sum_{z\in E, h\in I_z} Q_\beta(h).$$

By Jensen's inequality, we have

$$\log\frac{\pi_{\gamma,\beta}}{\pi_\beta}(x) \le 2\sup_{g\in\cup_z I_z} |\log Q_{\gamma,\beta}(g) - \log Q_\beta(g)|.$$

Finally, we obtain

$$\mathrm{Ent}_{\pi_{\beta_t}}(m_t) \le \mathrm{Ent}_{\pi_{\gamma_t,\beta_t}}(m_t) + 2\sup_{g\in\cup_x I_x} |\log Q_{\gamma_t,\beta_t}(g) - \log Q_{\beta_t}(g)|. \tag{20}$$

Using (9) the proof is complete. □

We now make the above observations precise by considering more specific, although general, transitions $Q_{\gamma,\beta}$.

COROLLARY 4. *Let V be a nonnegative function $V: E\times E \to \mathbb{R}_+$ and $Q_{\gamma,\beta}(x,y) = q_\beta(x,y)\, e^{\beta V_\gamma(x,y)}$. Suppose the assumptions of Corollary 2 are satisfied and, for every $q(x,y) > 0$,*

$$|V_{\gamma_t}(x,y) - V(x,y)| = o(1/\log t). \tag{21}$$

When the inverse-freezing schedule has parametric form $\beta_t = K^{-1}\log t$, for t sufficiently large and $K > c$, we have

$$\lim_{t\to+\infty} Ent_{\pi_{\beta_t}}(m_t) = 0 \qquad \textit{and} \qquad \lim_{t\to+\infty} P(X_t \in V^\star) = 1, \tag{22}$$

where π_β is the unique invariant probability measure of the Markov generator $Q_\beta - I$ with

$$Q_\beta(x,y) = q(x,y)\, e^{-\beta V(x,y)}, \qquad V^\star = \{x \in E :\ V(x) = \min_E V\},$$

and

$$V(x) = \min_{g\in G(x)} \sum_{(y\to z)\in g} V(y,z).$$

If $V_{\gamma_t}(x,y) = (U_t(y) - U_t(x))^+$ with $U_t : E \to \mathbb{R}_+$ and $\lim_{t\to+\infty} U_t(x) = U(x)$, condition (21) takes the form

$$\lim_{t\to+\infty} \log t\ |U_t(x) - U(x)| = 0.$$

The following examples illustrate the results and the conditions stated in the above theorems.

Examples.

1. Let us now turn our attention to the transition probability kernel

$$Q_{\gamma,\beta}(x,y) = q(x,y)\, e^{-\beta V_\gamma(x,y)},$$

where

$$V_\gamma(x,y) = (U_\gamma(y) - U_\gamma(x))^+ \ \text{ if } q(x,y) > 0 \text{ and } \ +\infty \ \text{ otherwise.}$$

In this case conditions (6) and (7) take the form

$$\sup_{x\in E} U_\gamma(x) - \inf_{x\in E} U_\gamma(x) \le d_1(\gamma), \qquad \sup_{x\in E} \frac{d}{d\gamma} U_\gamma(x) - \inf_{x\in E} \frac{d}{d\gamma} U_\gamma(x) \le d_2(\beta)/\beta.$$

In addition, if

$$\sup_{x,\gamma} U_\gamma(x) < +\infty, \qquad \sup_{x,\gamma} \frac{d}{d\gamma} U_\gamma(x) < +\ infty,$$

then we can choose $d_2(\beta) = \beta\, d_2$ and $d_1(\gamma) = d_1 < +\infty$ $(d_1, d_2 > 0)$ and the condition (10) takes the form

$$\frac{d\gamma_t}{dt} = o(1/(t^{c/K}(\log t)^2)).$$

2. Let us examine the above example with time-inhomogeneous potential given by

$$U_\gamma(x) = \theta(\gamma)^{-1} \int_0^{\theta(\gamma)} C(s,x)\, ds$$

with $C : \mathbb{R}_+ \times E \to \mathbb{R}_+$ bounded and $\frac{d}{d\gamma} \log \theta(\gamma) < +\infty$.
From the boundedness of C we can choose constants $d_1, d_2 > 0$ so that $d_1(\gamma) = d_1$ and $d_2(\beta) = \beta\ d_2$ satisfy the required conditions. Finally, if $\gamma_t = A\ \log t$ and $\theta(\gamma) = e^\gamma$, then we have

$$\frac{d\gamma_t}{dt} = A/t = o\left(1/(t^{c/K}(\log t)^2)\right), \qquad \frac{d}{d\gamma} \log \theta(\gamma) = 1,$$

and

$$U_{\gamma_t}(x) = \frac{1}{t^A} \int_0^{t^A} C(s, x)\ ds.$$

1.2. General results on the virtual energy. In section 1 we proved the convergence in probability of a class of stochastic algorithms to the set of the global minima $V^\star$ of a virtual energy function V. The crucial need of course is to estimate $V^\star$. In the case where q is symmetric and $V(x, y) = (U(y) - U(x))^+$ with $U : E \to \mathbb{R}_+$ it is well known that $V^\star = U^\star$. For a generalization of this see Trouvé [12]. The situation becomes considerably more involved when the above assumptions are dispensed with. The purpose of this section is to introduce a natural test set approach to study $V^\star$. More precisely, we will give several conditions for a given subset $H \subset E$ to contain $V^\star$.

Let us recall some basic definitions. Let E be a finite set and q an irreducible Markov kernel. Assume that a given function $V : E \times E \to \overline{\mathbb{R}}_+$ satisfies

$$V(x, y) < +\infty \Longleftrightarrow q(x, y) > 0.$$

Let us write $C_{x,y}$ the paths p in E joining x and y, that is,

$$\forall k \in \{0, \ldots, |p| - 1\} \quad q(p_k, p_{k+1}) > 0 \quad \text{and} \quad p_0 = x, p_{|p|} = y,$$

where $|p|$ is the length of p. For $x, y \in E$, $p \in C_{x,y}$, and $g \in G(x)$ we note

$$V(p) = \sum_{k=0}^{|p|-1} V(p_k, p_{k+1}), \qquad V(g) = \sum_{(y \to z) \in g} V(y, z), \qquad V(x) = \min_{g \in G(x)} V(g).$$

For $H \subset E$ and g an x-graph over H (that is, $g \in G_H(x)$) it is convenient to define a new communication cost V_H by making the set H a *taboo* set. Namely, for every $x, y \in H$

$$\begin{aligned} V_H(x, y) &= \min\{V(p)\ :\ p \in C_{x,y} \text{ with } \forall k \in \{1, \ldots, |p| - 1\}\ p_k \notin H\}, \\ V_H(g) &= \sum_{(y \to z) \in g} V_H(y, z). \end{aligned}$$

It is also convenient to define the virtual energy function associated with V_H:

$$\forall x \in H \qquad V_H(x) = \min_{g \in G_H(x)} V_H(g) - \min_{y \in H, h \in G_H(y)} V_H(h).$$

Finally, let us write

$$V(H) = \min_{x \in H} V(x) \qquad \text{and} \qquad V_H^\star = \left\{x \in H\ :\ V_H(x) = \min_H V_H\right\}.$$

LEMMA 2. *$\forall x \in H,\ V_H(x) = V(x) - V(H)$.*

Lemma 2 is an easy consequence of the following lemma.

LEMMA 3. *Let Q be an irreducible transition probability over E with invariant measure μ. Let X be a Markov chain with transition probability Q and initial measure m such that $m(x) > 0\ \forall x \in E$. Given a subset $H \subset E$ define*

$$T_1 = \inf\{n \geq 1\ :\ X_n \in H\} \qquad \tilde{Q} = P(X_{T_1} = y/X_0 = x).$$

Then $\tilde{Q}$ is an irreducible Markov kernel over H and its invariant probability measure is given by

$$\tilde{\mu}(x) = \mu(x)/\mu(H).$$

There are several ways to prove Lemmas 2 and 3. The following may be the shortest in this context.

Proof of Lemma 3. The proof is a consequence of the law of large numbers. Let us set by induction on the parameter $n \geq 1$

$$T_{n+1} = \inf\{k > T_n\ :\ X_k \in H\}, \qquad n \geq 1.$$

Now, under the above conditions, the random process $Y = (Y_n)_{n\geq 0}$ defined by

$$Y_0 = X_0, \qquad Y_n = X_{T_n}, \qquad n \geq 1,$$

is an irreducible Markov chain over H with transition probability kernel $\tilde{Q}$. Let $\tilde{\mu}$ be its invariant probability measure. First we note that P-almost surely

$$\forall x \in E \qquad \frac{1}{n}\sum_{i=1}^{n} 1_x(Y_i) \underset{n \to +\infty}{\longrightarrow} \tilde{\mu}(x)$$

On the other hand we have

$$\frac{1}{n}\sum_{i=1}^{n} 1_x(Y_i) = \frac{T_n}{n}\left(\frac{1}{T_n}\sum_{i=1}^{T_n} 1_x(X_i)\right), \qquad \frac{n}{T_n} = \frac{1}{T_n}\sum_{i=1}^{T_n} 1_H(X_i),$$

and

$$\frac{1}{T_n}\sum_{i=1}^{T_n} 1_x(X_i) \underset{n \to +\infty}{\longrightarrow} \mu(x), \qquad \frac{1}{T_n}\sum_{i=1}^{T_n} 1_H(X_i) \underset{n \to +\infty}{\longrightarrow} \mu(H)$$

P-almost everywhere (P.a.e.). The lemma follows immediately. □

We come to the proof of Lemma 2.

Proof of Lemma 2. We shall give a sketch of the proof. Let us denote by Q_β the Markov kernel over E given by

$$Q_\beta(x,y) = \begin{cases} |E|^{-1}\ \exp -\beta V(x,y) & \text{if } x \neq y, \\ 1 - |E|^{-1}\ \sum_{z\in E\,:z\neq x} \exp -\beta V(x,z) & \text{otherwise.} \end{cases}$$

Let μ_β be the invariant measure of Q_β. From the description of μ_β in terms of x-graphs over E it is clear that

$$\mu_\beta(x) \underset{\beta \to +\infty}{\sim} C(x)\ \exp -\beta V(x)$$

for some nonnegative function $C : E \to \mathbb{R}_+^*$. If one now defines $\tilde{Q}_\beta$ as in Lemma 3, by elementary large deviation arguments one sees the equivalence

$$\tilde{Q}_\beta(x,y) \underset{\beta \to +\infty}{\sim} \tilde{q}(x,y)\ \exp -\beta V_H(x,y) \qquad \forall x,y \in H$$

for some irreducible Markov kernel $\tilde{q} : E \times E \to \mathbb{R}_+$. Therefore if $\tilde{\mu}_\beta$ is the invariant measure of $\tilde{Q}_\beta$ one has for some nonnegative function $\tilde{C} : E \to \mathbb{R}_+^*$

$$\tilde{\mu}_\beta(x) \underset{\beta \to +\infty}{\sim} \tilde{C}(x)\ \exp -\beta V_H(x),$$

from which our claim follows easily. □

We now give a definition that we will use in our formulation of our test set approach to understand the limiting behavior of a generalized simulated annealing.

DEFINITION 1. *Let H be a subset of E. We say that a partition $\mathcal{H} = \{H_1, \ldots, H_n\}$ of H is a V-partition if for every $1 \le i \le n$ and $x, y \in H_i$, $x \ne y$, there exists a path $p \in C_{x,y}$ such that*

$$\forall 0 \le k < |p| \qquad p_k \in H_i \quad \textit{and} \quad V(p_k, p_{k+1}) = 0.$$

We observe that for a given subset $H \subset E$, one can always obtain a V-partition. For instance

$$\mathcal{H} = \{\{x\} \ :\ x \in H\}$$

is a V-partition. Given a V-partition $\mathcal{H} = \{H_1, \ldots, H_n\}$ of $H \subset E$, it is convenient to define a new communication cost function $V_{\mathcal{H}}$ by setting for every $x \in H_i$ and $y \in H_j$, $1 \le i, j \le n$

$$\begin{aligned} V_{\mathcal{H}}(x,y) = \{V(p) \ :\ p \in C_{x,y} \ \ &\exists 0 \le n_1 < n_2 \le |p| \\ &\forall 0 \le k \le n_1, p_k \in H_i, \quad \forall n_1 < k < n_2, p_k \notin H, \quad \forall n_2 \le k \le |p| p_k \in H_j\}. \end{aligned}$$

It is easily seen that $V_{\mathcal{H}}(x,y)$ does not depend on the choice of $x \in H_i$ and $y \in H_j$. Moreover we note that if $\mathcal{H} = \{\{x\} \ :\ x \in H\}$, then $V_{\mathcal{H}} = V_H$.

Let $V_{\mathcal{H}}$ be the virtual energy function associated with the communication cost function $V_{\mathcal{H}}$, namely,

$$\forall x \in H \qquad V_{\mathcal{H}}(x) = \min_{g \in G_H(x)} V_{\mathcal{H}}(g) \qquad \text{with} \quad V_{\mathcal{H}}(g) = \sum_{(y \to z) \in g} V_{\mathcal{H}}(y,z).$$

As usual, we also put

$$V_{\mathcal{H}}^\star = \left\{ x \in H \ :\ V_{\mathcal{H}}(x) = \min_H V_{\mathcal{H}} \right\}.$$

PROPOSITION 2. *If $\mathcal{H}$ is a V-partition of a subset $H \subset E$, then we have $V_{\mathcal{H}} = V_H$.*

Proof. Let us prove that for every $x \in H$

$$\min_{g \in G_H(x)} V_H(g) = \min_{g \in G_H(x)} V_{\mathcal{H}}(g).$$

Because $\mathcal{H} = \{H_1, \ldots, H_n\}$ is a V-partition it is clear that $V_{\mathcal{H}} \leq V_H$. So our claim will follow provided that for every $g \in G_H(x)$ we can build a new $\tilde{g} \in G_H(x)$ such that $V_{\mathcal{H}}(g) \geq V_H(\tilde{g})$. For this purpose let $g \in G_H(x)$ for some $x \in H$ and let $i \in \{1, \ldots, n\}$ such that $x \in H_i$. We will construct an x-graph $\tilde{g} \in G_H(x)$ such that $V_{\mathcal{H}}(g) \geq V_H(\tilde{g})$. For this we introduce the set

$$\Gamma = \{(j \to k) \ : \ \exists (y, z) \in g \text{ such that } \ y \in H_j \text{ and } z \in H_k\}.$$

Obviously Γ is not an i-graph over $\{1, \ldots, n\}$ but examination of Γ soon yields that it contains at least one i-graph G_i. Now for $j \neq i$ and $(j \to k) \in G_i$ (unique) there exists an arrow $(y_j \to z_k) \in g$ such that $y_j \in H_j$ and $z_k \in H_k$.

On the other hand, from the definition of $V_{\mathcal{H}}$ there exists a path $p \in C_{y_j, z_k}$ and $0 \leq n_1 < n_2 \leq |p|$ such that

$$\forall 0 \leq l \leq n_1 \quad p_l \in H_j, \qquad \forall n_1 < l < n_2 \quad p_l \notin H, \qquad \forall n_2 \leq l \leq |p| \quad p_l \in H_k,$$

and $V(p) = V_{\mathcal{H}}(y_j, z_k)$. Given such a path $p \in C_{y_j, z_k}$ let us set

$$\tilde{y}_j = p_{n_1}, \qquad \tilde{z}_k = p_{n_2}.$$

Finally, because $\mathcal{H}$ is a V-partition there exists for every $1 \leq i \leq n$ a $\tilde{y}_j$-graph $\tilde{g}_j \in G_{H_j}(\tilde{y}_j)$ such that $V(\tilde{g}_j) = 0$, with the convention $\tilde{y}_i = x$.

Using the above construction it is easily seen that the set of arrows

$$\tilde{g} = \bigcup_{i=1}^{n} \tilde{g}_i \ \bigcup \ \bigcup_{(j \to k) \in G_i} \{(\tilde{y}_j \to \tilde{z}_k)\}$$

is an x-graph over H and, from the construction of g, it follows that

$$V_{\mathcal{H}}(g) \geq V_{\mathcal{H}}(\tilde{g}) = \sum_{(j \to k) \in G_i} V_{\mathcal{H}}(\tilde{y}_j, \tilde{z}_k) = \sum_{(j \to k) \in G_i} V_H(\tilde{y}_j, \tilde{z}_k) = V_H(\tilde{g}).$$

This ends the proof. □

The following concept of λ-stability leads to a natural test set approach to study $V^\star$.

Definition 2. *Let λ be a nonnegative real number. A subset $H \subset E$ is called λ-stable with respect to a communication cost function V when the following conditions are satisfied:*

1. $\forall x \in H \ \ \forall y \notin H, \qquad V(x, y) > \lambda,$
2. $\forall x \notin H \ \ \exists y \in H, \qquad V(x, y) \leq \lambda.$

The importance of the notion of λ-stability resides in the following result, which extends Lemma 4.1 of Freidlin–Wentzell [6].

Proposition 3. *Let λ be a nonnegative real number and $H \subset E$. Any λ-stable subset H with respect to V contains $V^\star$ and $V_H = V_{/H}$*

Proof. Let H be a λ-stable subset of E. Let $x \notin H$ and let g be an element of $G(x)$ such that $V(g) = V(x)$. There exists almost one $y \in H$ such that $V(x, y) \leq \lambda$. Now we note p the exit path from y to x extracted from g, that is,

$$p_0 = y, p_{|p|} = x \quad \text{and} \quad \forall k = 0, \ldots, |p| - 1, \quad (p_k \to p_{k+1}) \in g.$$

Write $k_0 = \inf\{k = 0, \ldots, |p| - 1 \quad : \quad p_k \notin H\}$. Let $\overline{g}$ be the graph obtained from g by replacing the arrow $(p_{k_0 - 1} \to p_{k_0})$ by $(x \to y)$. Then we have $\overline{g} \in G(p_{k_0 - 1})$ and

$$V(x) = V(g) > V(\overline{g}) \geq V(p_{k_0 - 1}), \qquad p_{k_0 - 1} \in H.$$

This completes the proof of the first assertion. The last assertion is a clear consequence of Lemma 2. □

Let us now reduce all of the results of this section to a proposition which we shall use for later reference.

PROPOSITION 4. *Let A be a 0-stable subset of E with respect to V and let $\mathcal{A}$ be a V-partition of A. If H is a λ-stable subset of A with respect to $V_{\mathcal{A}}$ for some $\lambda \geq 0$, then $V^\star \subset H$.*

Proof. Using Proposition 3 it is easily seen that $V^\star \subset A$, $V_A = V_{/A} = V_{\mathcal{A}}$, and $V_{\mathcal{A}}^\star \subset H$. Thus one gets $V^\star = V_{\mathcal{A}}^\star = V_{\mathcal{A}}^\star \subset H$, and the proof of Proposition 4 is completed. □

2. Applications.

In this section we examine two applications. In section 2.1 we use the results of the first section to derive a new and simple proof of the convergence of the genetic algorithms. We shall prove this important result in a way different from the original proof of Cerf [2]. The proof splits quite naturally into two distinct parts:

1. We use the relative entropy convergence results stated in the first section to prove the convergence of the algorithm.
2. Then we investigate the test set approach, introduced in the second part of section 1, to prove that the set of the global minima of the virtual energy is contained in the product of the set of the global minima of the fitness function.

In section 2.2 we apply Corollary 4 to construct a stochastic algorithm for the numerical solving of a general mean cost optimization problem.

2.1. Genetic algorithms.

A genetic algorithm is a discrete time Markov process $\widehat{x} = (\widehat{x}_n)_n$ with state space $E = S^N$ ($N > 1$ and S a finite set) and whose transition probabilities G_n include a mutation M_n and a selection S_n mechanism. The N-tuple of elements of S, i.e., the points of the set E, are called particle systems and most will be denoted by the letters x, y, z. In what follows, we shall distinguish two kinds of combinations, namely,

$$\text{(a)} \qquad P\left(\widehat{x}_n \in dx \,/\widehat{x}_{n-1} = z\,\right) = \int_E M_n(z, dy)\; S_n(y, dx),$$

$$\text{(b)} \qquad P\left(\widehat{x}_n \in dx \,/\widehat{x}_{n-1} = z\,\right) = \alpha\; M_n(z, dx) + (1-\alpha)\;\; S_n(z, dx), \qquad 0 < \alpha < 1.$$

Mutations. The mutation transition is modeled by independent motion of each particle, that is,

$$M_n(z, dy) = \prod_{p=1}^{N} \; K_n(z^p, dy^p),$$

where K_n is a Markov kernel over S, $z = (z^p)_{1 \leq p \leq N}$, and $y = (y^p)_{1 \leq p \leq N}$.

Selection. In the selection transition the particles are chosen randomly and independently in the previous population according to a given selection function $F_n : S \to \mathbb{R}^+$, namely,

$$S_n(y, dx) = \prod_{p=1}^{N} \sum_{i=1}^{N} \frac{F_n(y^i)}{\sum_{j=1}^{N} F_n(y^j)} \;\; 1_{y^i}(dx^p).$$

The study of the convergence, as $N \to +\infty$ or $n \to +\infty$, of such algorithms requires specific developments because each individual particle is no longer Markovian and it is difficult to produce mean error estimates. In [3] one of the authors applied and extended such algorithms to nonlinear filtering problems. An apparent difficulty in establishing a convergence result as $n \to +\infty$ is finding a candidate invariant measure that enables us to describe some aspects of the limiting behavior of the algorithm. To our knowledge, Cerf gives in his Ph.D. dissertation [2] the first convergence result $n \to +\infty$ for a genetic algorithm to converge in probability to the global minima of a given fitness function. More precisely, he studies the following situation:

1. The state space S is finite and $G_n = M_n\, S_n$.
2. The mutation Markov transition kernels $K_n(x^1, y^1)$ are governed by a parameter a and a cooling schedule $\beta(n) : \mathbb{N} \to \mathbb{R}^+$, namely,

$$M_n(x,y) = Q^{(1)}_{\beta(n)}(x,y) \overset{\text{def}}{=} \prod_{p=1}^{N} k_{\beta(n)}(x^p, y^p) \tag{23}$$

with

$$k_\beta(x^1,y^1) \overset{\text{def}}{=} \begin{cases} k(x^1,y^1)\, e^{-a\,\beta} & \text{if} \quad x^1 \neq y^1, \\ 1 - \sum\limits_{z^1 \neq x^1} k(x^1,z^1)\, e^{-a\,\beta} & \text{if} \quad x^1 = y^1, \end{cases} \tag{24}$$

where k is a given irreducible Markov kernel on the space S.

3. The selection operator is built with a fitness function $f : S \to \mathbb{R}^+$ and a cooling schedule $\beta(n) : \mathbb{N} \to \mathbb{R}^+$, namely,

$$S_n(x,y) = Q^{(2)}_{\beta(n)}(x,y) \overset{\text{def}}{=} \prod_{p=1}^{N} \sum_{i=1}^{N} \frac{e^{-\,\beta(n)\, f(x^i)}}{\sum\limits_{j=1}^{N} e^{-\,\beta(n)\, f(x^j)}}\, 1_{x^i}(y^p). \tag{25}$$

Cerf gives several conditions on the rate of decrease of the cooling schedule $\beta(n) \to +\infty$ to ensure all the particles visit the set of global minima, as times goes on, when the number of particles is greater than a critical value. He carries out in a discrete time setting a precise study using large deviation techniques and the powerful tools developed by Trouvé [12]. Simplifying and extending techniques of Cerf and Trouvé, our results are obtained by using the relative entropy convergence result stated in Corollary 1 and by investigating the test set approach introduced in section 1.2.

2.1.1. General results and notations. In this section we will consider genetic algorithms described by the transition probability kernel

$$Q_\beta = Q^{(1)}_\beta Q^{(2)}_\beta \quad \text{or} \quad \tilde{Q}_\beta = \alpha_1\, Q^{(1)}_\beta + \alpha_2\, Q^{(2)}_\beta \quad \text{with} \quad \alpha_1 + \alpha_2 = 1 \quad \text{and} \quad 0 < \alpha_1 < 1 \tag{26}$$

and nonnecessarily vanishing mutations. More precisely, we assume that the mutation transition kernels k_β in (24) have the property

$$\exists b > 0, \qquad b^{-1}\;\; k(x_1,x_2)\, e^{-\beta\, a(x_1,x_2)} \leq k_\beta(x_1,x_2) \leq b\;\; k(x_1,x_2)\, e^{-\beta\, a(x_1,x_2)}, \tag{27}$$

where $a : S^2 \to \overline{\mathbb{R}}_+$, $a(x,y) < +\infty \iff k(x,y) > 0$, k is an irreducible Markov kernel, and the relation on S defined by

$$x_1 \sim x_2 \iff a(x_1,x_2) = 0$$

is an equivalence relation. This leads us naturally to consider the partition $S_1, \ldots, S_{n(a)}$, $n(a) \geq 1$, induced by $\sim$. If x_1 is a typical element of S, then the equivalence class of x_1 will be denoted by $S(x_1)$:

$$S(x_1) = \{x_2 \in S \;:\; x_1 \sim x_2\}.$$

We further require that

$$a(x_1, x_2) = 0 \Longrightarrow f(x_1) = f(x_2)$$

and for some $\beta_0 > 0$

$$\forall x_1, x_2 \in S \qquad \sup_{\beta \geq \beta_0} |d \log k_\beta(x_1, x_2)/d\beta| < +\infty. \tag{28}$$

To our knowledge the models of evolutionary processes (26) have not been covered by the literature of genetic algorithms.

Examples. The following mutation transition kernels have the properties (27) and (28):

1. $k_\beta(x_1, x_2) = \begin{cases} k(x_1, x_2)\, e^{-\beta\, a(x_1, x_2)} & \text{if} \quad a(x_1, x_2) > 0, \\ |S(x_1)|^{-1} \left(1 - \sum_{y_1 \notin S(x_1)} k(x_1, y_1)\, e^{-\beta\, a(x_1, y_1)}\right) & \text{otherwise.} \end{cases}$
2. $k_\beta(x_1, x_2) = \dfrac{e^{-\beta\, a(x_1, x_2)}\; k(x_1, x_2)}{\sum_{y_1 \in S} e^{-\beta\, a(x_1, y_1)}\; k(x_1, y_1)} \qquad \forall (x_1, x_2) \in S^2.$

Finally, let us note that if a is given by

$$a(x_1, x_2) = a\,(1 - 1_{x_1}(x_2)) \qquad \forall (x_1, x_2) \in S^2 \;:\; k(x_1, x_2) > 0,$$

then the first transition probability kernel is clearly the same as the mutation transition probability kernel (24) studied by Cerf.

In this special situation, the first model $Q_\beta = Q_\beta^{(1)} Q_\beta^{(2)}$ is of course identical to Cerf's model of a genetic algorithm.

Let us recall some terminology introduced by Cerf in [2]. The cardinality of a set will be denoted by $|.|$. If x and y belong to $E = S^N$ and $f : S \to \mathbb{R}^+$, we write

$$[x] = \{x_k \;:\; 1 \leq k \leq N\}, \qquad f^\star = \{x_1 \in S \;:\; f(x_1) = \min_S f\},$$

$$\widehat{x} = \{k \;:\; 1 \leq k \leq N\,,\; f(x_k) = \widehat{f}(x)\}, \qquad \widehat{f}(x) = \min_{1 \leq k \leq N} f(x_k),$$

$$x(y_1) = \text{Card}\{k \;:\; 1 \leq k \leq N\,,\; x_k = y_1\}.$$

Using these notations, an easy calculation shows that for $k = 1$ or $k = 2$

$$Q_\beta^{(k)}(x, y) = q_\beta^{(k)}(x, y)\; e^{-\beta V^{(k)}(x, y)}, \qquad q_\beta^{(k)}(x, y) = q^{(k)}(x, y)\; \theta_\beta^{(k)}(x, y),$$

where

$$q^{(1)}(x, y) = \prod_{i : a(x_i, y_j) > 0} k(x_i, y_i), \qquad q^{(2)}(x, y) = \prod_{i=1}^N \frac{x(y_i)}{|\widehat{x}|},$$

$$V^{(1)}(x, y) = \sum_{i=1}^N a(x_i, y_i) \quad \text{if} \quad q^{(1)}(x, y) > 0,$$

$$V^{(2)}(x,y) = \sum_{i=1}^{N} \left(f(y_i) - \widehat{f}(x) \right) \quad \text{if} \quad q^{(2)}(x,y) > 0,$$
$$\theta_\beta^{(1)}(x,y) = \prod_{i:a(x_i,y_i)=0} k_\beta(x_i,y_i),$$
$$\theta_\beta^{(2)}(x,y) = \left(1 + |\widehat{x}|^{-1} \sum_{k \notin \widehat{x}} \exp -\beta(f(x_k) - \widehat{f}(x)) \right)^{-N} .$$

As usual, we will use the convention

$$\forall k \in \{1,2\} \qquad V^{(k)}(x,y) = +\infty \Longleftrightarrow q^{(k)}(x,y) = 0.$$

The asymptotic mutation dynamics of the genetic algorithms is governed by the kernel k and the function a. The irreducibility condition on the kernel k and the fact that a is an equivalence relation are sufficient conditions to allow the system of particles to visit all the state space E. Thus, using the above notations, it is easily checked that

- $q^{(1)}$ is irreducible and $q^{(2)}(x,x) > 0$ for every $x \in E$;
- for every $k = 1,2$ and $q^{(k)}(x,y) > 0$ we have $\sup_{\beta \geq \beta_0} |d \log \theta_\beta^k(x,y)/d\beta| < +\infty$ for some $\beta_0 \geq 0$.

Then, returning to our general model (12), the conditions introduced in Corollary 1 are satisfied in both situations:

1. $$Q_\beta(x,y) \overset{\text{def}}{=} Q_\beta^{(1)} \, Q_\beta^{(2)}(x,y) = \sum_{u \in U} \overline{q}_\beta(x,u,y) e^{-\beta \overline{V}(x,u,y)}$$
   with $U = E, \overline{V}(x,u,y) = V^{(1)}(x,u) + V^{(2)}(u,y), \overline{q}(x,u,y) = q^{(1)}(x,u) q^{(2)}(u,y)$, and $\theta_\beta(x,u,y) = \theta_\beta^{(1)}(x,u) \, \theta_\beta^{(2)}(u,y), \overline{q}_\beta(x,u,y) = \theta_\beta(x,u,y)\overline{q}(x,u,y)$.
2. $$\tilde{Q}_\beta(x,y) \overset{\text{def}}{=} \alpha_1 \, Q_\beta^{(1)}(x,y) + \alpha_2 \, Q_\beta^{(2)}(x,y) = \sum_{u \in U} \overline{q}_\beta(x,u,y) e^{-\beta \overline{V}(x,u,y)}$$
   with $U = \{1,2\}, \overline{V}(x,u,y) = V^{(u)}(x,y), \overline{q}(x,u,y) = \alpha_u \; q^{(u)}(x,y)$, and $\theta_\beta(x,u,y) = \theta_\beta^{(u)}(x,y), \overline{q}_\beta(x,u,y) = \theta_\beta(x,u,y)\overline{q}(x,u,y)$.

2.1.2. A convergence theorem. To clarify the notations, in the remainder of section 2 we will use the diacritic $(\tilde{\cdot})$ to distinguish the communication cost functions, the virtual energy function, and the critical height associated with the transition probability kernels Q_β from those associated with $\tilde{Q}_\beta$.

From the above observations and Corollary 1, choosing β of the form

$$\beta_t = K^{-1} \, \log t, \quad \text{where} \quad K > c \quad (\text{resp., } K > \tilde{c}),$$

for t sufficiently large yields that the canonical process (Ω, P, F_t, X_t) associated with the family of generators $(L_{\beta_t})_{t \geq 0} = (Q_{\beta_t} - I)_{t \geq 0}$ (resp., $(\tilde{Q}_{\beta_t} - I)_{t \geq 0}$) converges in probability to the set of the global minima $V^\star$ (resp., $\tilde{V}^\star$) of the virtual energy V (resp., $\tilde{V}$) associated with Q_β (resp., $\tilde{Q}_\beta$) and described in Corollary 1. One open problem is to compare c and $\tilde{c}$. Let us remark that $\tilde{c}$ does not depend on the choice of the parameter $\alpha \in]0,1[$. In view of these observations the bulk of the proof rests on showing that $V^\star$ and $\tilde{V}^\star$ are subsets of $(f^\star)$, where $(f^\star)$ is the set in E defined by

$$(f^\star) = \left\{ x \in E \ : \ \widehat{f}(x) = \min_E f \right\}.$$

The main purpose of this section is to prove a more general result. We will prove that $V^\star$ and $\tilde{V}^\star$ are subsets of $(f^\star) \cap A$, where A is the set in E defined by

$$A = \{x \in E \ : \ x_k \sim x_l \quad \forall 1 \leq k, l \leq N\}.$$

By $\mathcal{A}$ we will denote the partition of A induced by the equivalence relation $\sim$

$$\mathcal{A} = \{A_1, \ldots, A_{n(a)}\}, \qquad A_i = \{x \in E \ : \ [x] \subset S_i\} \qquad \forall 1 \leq i \leq n(a).$$

As usual we associate with each typical element $x = (x_1, \ldots, x_N) \in A$ the subset

$$A(x) = \{y \in E \ : \ [y] \subset S(x_1)\}.$$

Note that A is 0-stable with respect to V and $\tilde{V}$. Moreover, from our constructions, a routine proof yields that $\mathcal{A}$ is a V and $\tilde{V}$-partition of A. In view of Propositions 2 and 3 it follows that $V^\star \subset A$, $\tilde{V}^\star \subset A$, and

$$\forall x \in A \qquad V(x) = \min_{g \in G_A(x)} V_{\mathcal{A}}(g), \qquad \tilde{V}(x) = \min_{g \in G_A(x)} \tilde{V}_{\mathcal{A}}(g).$$

Now, from Proposition 4, to prove that $V^\star$ and $\tilde{V}^\star$ are subsets of $(f^\star)$ it clearly suffices to find a constant λ such that the subset $(f^\star) \cap A$ is λ-stable with respect to $V_{\mathcal{A}}$ and $\tilde{V}_{\mathcal{A}}$.

As we will see such results hold when the size N of the particle systems is greater that a critical value which depends on the functions a and f. We shall study this critical size now, beginning with two important lemmas.

Before proceeding we need to introduce some additional notations.

By Γ_{x_1,x_2}, $x_1, x_2 \in S$, we denote the paths q in S joining x_1 and x_2, that is,

$$\forall 0 \leq l < |q| \qquad k(x_l, x_{l+1}) > 0, \qquad q_0 = x_1, \quad q_{|q|} = x_2.$$

We will also denote $R(a)$ as the smallest integer such that for every $x_1, x_2 \in S$ in two different classes there exists a path joining x_1 and x_2 with length $|q| \leq R(a)$. More precisely,

$$R(a) = \max_{1 \leq i,j \leq n(a)} \min_{(x_i,x_j) \in S_i \times S_j} \min_{q \in \Gamma_{x_i,x_j}} |q|.$$

It also will be convenient to use the following definitions:

$$\begin{array}{ll} \triangle a = \min\{a(x_1,x_2) \ : \ a(x_1,x_2) \neq 0\}, & \triangle f = \min\{|f(x_1) - f(x_2)| \ : \ f(x_1) \neq f(x_2)\}, \\ \delta(a) = \sup\{a(x_1,x_2) \ : \ x_1, x_2 \in S\}, & \delta(f) = \sup\{|f(x_1) - f(x_2)| \ : \ x_1, x_2 \in S\}. \end{array}$$

LEMMA 4. *For every $x, y \in A$ such that $\widehat{f}(x) \geq \widehat{f}(y)$ we have*

$$\tilde{V}_{\mathcal{A}}(x,y) \leq \delta(a)\ R(a). \tag{29}$$

Moreover, for every $x \in A$ there exists a state $y \in (f^\star) \cap A$ such that

$$V_{\mathcal{A}}(x,y) \leq (\delta(a) + \delta(f))\ R(a). \tag{30}$$

LEMMA 5. *For every $x, y \in A$ such that $\widehat{f}(x) < \widehat{f}(y)$ we have*

$$V_{\mathcal{A}}(x,y) \geq \min(\Delta a, \Delta f)\ N \quad \text{and} \quad \tilde{V}_{\mathcal{A}}(x,y) \geq \min(\Delta a, \Delta f)\ N. \tag{31}$$

Let us write

$$\lambda(a,f) = (\delta(a)+\delta(f))\,R(a), \qquad \tilde{\lambda}(a,f) = \delta(a)\,R(a),$$
$$N(a,f) = \lambda(a,f)/\min(\Delta a, \Delta f), \qquad \tilde{N}(a,f) = \tilde{\lambda}(a,f)/\min(\Delta a, \Delta f).$$

Combining Lemmas 4 and 5 one easily gets

$$N > N(a,f) \Longrightarrow (f^\star)\cap A \text{ is } \lambda(a,f)\text{-stable with respect to } V_{\mathcal{A}},$$
$$N > \tilde{N}(a,f) \Longrightarrow (f^\star)\cap A \text{ is } \tilde{\lambda}(a,f)\text{-stable with respect to } \tilde{V}_{\mathcal{A}}.$$

These results and Proposition 4 combine to yield the following theorem.

THEOREM 3. *We denote (Ω, P, F_t, X_t) as the canonical process associated with the family of generators $(L_{\beta_t})_{t\geq 0} = (Q_{\beta_t} - I)_{t\geq 0}$ (resp., $(\tilde{Q}_{\beta_t} - I)_{t\geq 0}$), c (resp., $\tilde{c}$) the critical height associated with the communication cost function V (resp., $\tilde{V}$), and m_t the distribution of X_t. If β_t assumes the parametric form $\beta_t = K^{-1}\log t$, for sufficiently large t, with $K > c$ and if $N > N(a,f)$ (resp., $\tilde{N}(a,f)$), then we have*

$$\lim_{t\to+\infty} \mathrm{Ent}_{\pi_{\beta_t}}(m_t) = 0 \qquad \text{and} \qquad \lim_{t\to+\infty} P(X_t \in (f^\star)\cap A) = 1,$$

where π_β is the invariant probability measure of $L_\beta = Q_\beta - I$ (resp., $\tilde{Q}_\beta - I$).

We come to the proof of Lemmas 4 and 5.

Proof of Lemma 4. Let $x = (x_1, \dots, x_N)$ and $y = (y_1, \dots, y_N)$ be two elements of A such that $\widehat{f}(x) \geq \widehat{f}(y)$. First let us remark that

$$x_1 \sim y_1 \Longrightarrow \forall 1 \leq k \leq N, \quad a(x_k, y_k) = 0 \Longrightarrow \forall 1 \leq k \leq N, \quad f(x_k) = f(y_k).$$

In this situation, a routine proof yields

$$V_{\mathcal{A}}(x,y) = 0 \qquad \text{and} \qquad \tilde{V}_{\mathcal{A}}(x,y) = 0.$$

If $a(x_1, y_1) > 0$ then the irreducibility condition implies the existence of a path $q \in \Gamma_{x_1,y_1}$ and a pair of integers $0 \leq n_1 < n_2 \leq |q|$ such that

$$\forall 0 \leq k \leq n_1,\ q_k \in S(x_1), \quad \forall n_1 < k < n_2,\ q_k \notin S(x_1), \quad \forall n_2 \leq k \leq |q|,\ q_k \in S(y_1). \tag{32}$$

Let us prove (29). For this, let $p \in \tilde{C}_{x,y}$ be the path defined by

$$\forall 0 \leq k \leq |q|, \qquad p_k = (q_k, x_2, \dots, x_N), \qquad p_{|q|+1} = (y_1), \qquad p_{|q|+2} = y.$$

From the definition of q we have

$$\forall 0 \leq k \leq n_1,\ p_k \in A(x), \quad \forall n_1 < k \leq |q|,\ p_k \notin A, \quad \forall k \in \{|q|+1, |q|+2\},\ p_k \in A(y). \tag{33}$$

Moreover, it follows that

$$0 \leq \sum_{k=0}^{n_1-1} \tilde{V}(p_k, p_{k+1}) \leq \sum_{k=0}^{n_1-1} V^{(1)}(p_k, p_{k+1}) = 0,$$
$$0 \leq \tilde{V}(p_{|q|}, p_{|q|+1}) \leq V^{(2)}(p_{|q|}, p_{|q|+1}) = 0, \quad \text{and}$$
$$\tilde{V}(p_{|q|+1}, p_{|q|+2}) = V^{(1)}(p_{|q|+1}, p_{|q|+2}) = 0.$$

Now, it appears from the proceeding that

$$\begin{aligned}\tilde{V}(p) &\leq \sum_{k=0}^{n_1-1} \tilde{V}(p_k, p_{k+1}) + \sum_{k=n_1}^{|q|-1} \tilde{V}(p_k, p_{k+1}) + \tilde{V}(p_{|q|}, p_{|q|+1}) + \tilde{V}(p_{|q|+1}, p_{|q|+2}) \\ &= \sum_{k=n_1}^{|q|-1} \tilde{V}(p_k, p_{k+1}) \leq \delta(a)|q|.\end{aligned}$$

Therefore $\tilde{V}(x,y) \leq \delta(a)\ R(a)$ and the proof of (29) is completed.

The proof of (30) is just a little more complicated.

Suppose $x \in A$ and y' is an element of A such that $\widehat{f}(x) \geq \widehat{f}(y') = \min_S f$ and $a(x_1, y'_1) > 0$. Let q be the path joining x_1 and y'_1 and defined as in (32). Using the above notations, let $(t_m)_m$ be the sequence of integers defined by

$$t_0 = n_1, \qquad t_m = \inf\left\{k > t_{m-1}\ :\ f(q_k) < f(q_{t_{m-1}})\right\} \qquad \forall m \geq 1.$$

Using the assumption $\widehat{f}(x) \geq \widehat{f}(y') = \min_S f$, examination of q soon yields that there exists an $m_0 \geq 1$ such that $t_{m_0} \leq n_2$ and $f(q_{t_{m_0}}) = \min_S f$. Consequently, we have constructed a sequence of states $(q_{t_m})_{0 \leq m \leq m_0}$ such that

$$\widehat{f}(x) = f(q_{t_0}) > f(q_{t_1}) > \cdots > f(q_{t_m}) > \cdots > f(q_{t_{m_0}}) = \widehat{f}(y), \tag{34}$$

where $y = (q_{t_{m_0}}, \ldots, q_{t_{m_0}}) \in A \cap (f^\star)$. With each state q_{t_m} we associate a state $p_m \in E$, $0 \leq m \leq m_0$, by setting

$$p_{t_0} = (q_{t_0}, x_2, \ldots, x_N), \qquad p_{t_m} = (q_{t_m}, \ldots, q_{t_m}, x_N) \quad \forall 1 \leq m \leq m_0.$$

First we note that $p_{t_m} \notin A\ \forall 1 \leq m \leq m_0$ and

$$\widehat{f}(x) = \widehat{f}(p_{t_0}) > \widehat{f}(p_{t_1}) > \cdots > \widehat{f}(p_{t_m}) > \cdots > \widehat{f}(p_{t_{m_0}}) = \min_S f = \widehat{f}(y).$$

Our next task is to construct a sequence of paths $(p^{(m)})_{0 \leq m \leq m_0+1}$ such that

$$p^{(0)} \in C_{x, p_{t_0}}, \qquad p^{(m)} \in C_{p_{t_{m-1}}, p_{t_m}}, \quad \forall 1 \leq m \leq m_0 \qquad p^{(m_0+1)} \in C_{p_{t_{m_0}}, y},$$

and

- the path $p^{(0)}$ has length $|p^{(0)}| = t_0$ and for every $0 \leq k \leq t_0$ the states $p_k^{(0)}$ belong to $A(x)$;
- for every $1 \leq m \leq m_0$, $p^{(m)}$ is a path joining $p_{t_{m-1}}$ and p_{t_m} in time

  $$|p^{(m)}| = t_m - t_{m-1},$$

  and for every $0 \leq k \leq t_m - t_{m-1}$ the states $p_k^{(m)}$ do not belong to A except the first initial state $p_0^{(1)} = p_{t_0} \in A(x)$;
- $p^{(m_0+1)} = (p_{t_{m_0}}, y)$.

It is straightforward to see that any path $p^{(0)}$ satisfying the above conditions has null cost, that is, $V(p^{(0)}) = 0$. Then, to obtain the desired upper bound it clearly suffices to have

$$\forall 1 \leq m \leq m_0 \qquad V(p^{(m)}) \leq (t_m - t_{m-1})\ (\delta(a) + \delta(f)).$$

We proceed to define $(p^{(m)})_{0 \leq m \leq m_0+1}$ as follows:

1. In view of (32) and (34) it is natural to define the initial path $p^{(0)} \overset{\text{def}}{=} (p_0, \dots, p_{t_0})$ by setting

$$\forall 0 \leq k \leq t_0 = n_1 \qquad p_k = (q_k, x_2, \dots, x_N) \in A(x).$$

As has already been noted, a simple calculation shows that

$$V(p^{(0)}) = \sum_{k=0}^{t_0-1} V^{(1)}(p_k, p_{k+1}) + V^{(2)}(p_{k+1}, p_{k+1}) = 0.$$

2. Taking into account that t_1 is the first time k such that $f(q_k) < \widehat{f}(x)$ we are lead to define $p^{(1)} \overset{\text{def}}{=} (p_{t_0}, p_{t_0+1}, \dots, p_{t_1})$ by setting

$$\forall t_0 \leq k < t_1 \qquad p_k = (q_k, x_2, \dots, x_N), \qquad p_{t_1} = (q_{t_1}, \dots, q_{t_1}, x_N).$$

Let us write $\overline{p}_{t_1} = (q_{t_1}, x_2, \dots, x_N)$. In this situation $p_k \notin A \; \forall t_0 < k \leq t_1$ and it is easy to verify that

$$\begin{aligned} V(p^{(1)}) \leq & \sum_{k=t_0}^{t_1-2} V^{(1)}(p_k, p_{k+1}) + V^{(2)}(p_{k+1}, p_{k+1}) + V^{(1)}(p_{t_1-1}, \overline{p}_{t_1}) \\ & + V^{(2)}(\overline{p}_{t_1}, p_{t_1}) \\ \leq & \, (t_1 - t_0)(\delta(a) + \delta(f)). \end{aligned}$$

3. As for item 2, we define the paths $p^{(m)} \overset{\text{def}}{=} (p_{t_{m-1}}, p_{t_{m-1}+1}, \dots, p_{t_m})$ for $2 \leq m \leq m_0$ by setting

$$\forall t_{m-1} \leq k < t_m p_k = (q_k, q_{t_{m-1}}, \dots, q_{t_{m-1}}, x_N), \quad p_{t_m} = (q_{t_m}, \dots, q_{t_m}, x_N).$$

Here again we have $p_k \notin A \; \forall t_{m-1} \leq k \leq t_m$. Let us introduce a new state $\overline{p}_{t_m} = (q_{t_m}, q_{t_{m-1}}, \dots, q_{t_{m-1}}, x_N)$. It is then an elementary matter to prove the inequalities

$$\begin{aligned} V(p^{(m)}) \leq & \sum_{k=t_{m-1}}^{t_m-2} \Big(V^{(1)}(p_k, p_{k+1}) + V^{(2)}(p_{k+1}, p_{k+1}) \Big) \\ & + V^{(1)}(p_{t_m-1}, \overline{p}_{t_m}) + V^{(2)}(\overline{p}_{t_m}, p_{t_m}) \\ \leq & \, (t_m - t_{m-1} - 1) \; (\delta(a) + 2\,\delta(f)) + \delta(a) + \delta(f) \\ \leq & \, (t_m - t_{m-1}) \; (\delta(a) + \delta(f)). \end{aligned}$$

4. Finally, let us note that

$$0 \leq V(p^{(m_0+1)}) \leq V^{(1)}(p_{t_{m_0}}, p_{t_{m_0}}) + V^{(2)}(p_{t_{m_0}}, y) = 0.$$

Consider the path $p = (p^{(0)}, \dots, p^{(m_0+1)}) \in C_{x,y}$ obtained by joining end to end all these paths. From the above inequalities it follows easily that $V(p) \leq |q| \; (\delta(a) + \delta(f))$. As a clear consequence one gets

$$V_{\mathcal{A}}(x, y) \leq (\delta(a) + \delta(f)) \; R(a).$$

This ends the proof of Lemma 4. □

Much more is true. In view of our assumptions on the function a and the constructions given in the proof of Lemma 4 we observe easily that

$$\forall x \in A \qquad \forall y \in A \cap (f^{\star}) \qquad V_{\mathcal{A}}(x,y) \leq (\delta(a) + \delta(f))\, R(a).$$

Proof of Lemma 5. Let (x,y) be a pair of points of A such that $\widehat{f}(x) < \widehat{f}(y)$. Now, let p belong to $\tilde{C}_{x,y}$. Note that since $\widehat{f}(x) < \widehat{f}(y)$ there exists a real number λ such that $\widehat{f}(x) < \lambda < \widehat{f}(y)$. Let us set

$$\forall 0 \leq l \leq |q| \qquad I_l = \left\{ i \in \{1, \dots, N\} \;:\; f(p_l^i) > \lambda \right\} \quad \text{and} \quad n_l = |I_l|.$$

It follows easily that $n_0 = 0$ and $n_{|p|} = N$.

Now, let T_k, $0 \leq k \leq N$, be the first time $l \geq 0$ such that $n_l \geq k$. More precisely,

$$T_k = \inf \{ l \in \{0, \dots, |p|\} \;:\; n_l \geq k \} \qquad \forall 0 \leq k \leq N.$$

Clearly it appears from the above that

$$T_0 = 0, \qquad T_N \leq |p|, \qquad n_{T_N} = N, \qquad n_{T_0} = n_0 = 0. \tag{35}$$

By definition of the communication cost function $V^{(1)}$ we can see that

$$V^{(1)}(p_{T_k-1}, p_{T_k}) \geq \sum_{i \in I_{T_k} - I_{T_{k-1}}} a(p^i_{T_k-1}, p^i_{T_k}) \geq (n_{T_k} - n_{T_k-1})\, \Delta a \qquad \forall 1 \leq k \leq N. \tag{36}$$

More precisely, p_{T_k-1} contains n_{T_k-1} individuals $p^i_{T_k-1}$ such that $f(p^i_{T_k-1}) > \lambda$ and p_{T_k} contains n_{T_k} individuals $p^j_{T_k}$ such that $f(p^j_{T_k}) > \lambda$. Therefore, if $V^{(1)}(p_{T_k-1}, p_{T_k}) < +\infty$, then the transition $p_{T_k-1} \to p_{T_k}$ necessarily involves at least $(n_{T_k} - n_{T_k-1})$ individual mutations.

Similarly, if $V^{(2)}(p_{T_k-1}, p_{T_k}) < +\infty$, then the system p_{T_k} contains at least $(n_{T_k} - n_{T_k-1})$ new individuals $p^i_{T_k} \in [p_{T_{k-1}}]$ such that $f(p^i_{T_k}) > \lambda$. Thus a discussion similar to that above leads to

$$V^{(2)}(p_{T_k-1}, p_{T_k}) \geq (n_{T_k} - n_{T_k-1})\, \Delta f. \tag{37}$$

Finally, by definition of $\tilde{V}$, we have

$$\tilde{V}(q) \geq \tilde{V}(p_{T_1-1}, p_{T_1}) + \cdots + \tilde{V}(p_{T_N-1}, p_{T_N}).$$

Let us remark that

$$n_{T_k-1} \leq k - 1 \leq n_{T_{k-1}} \quad \forall 1 \leq k \leq N.$$

Thus, combining (36) and (37) , we arrive at

$$\tilde{V}(q) \geq N \, \min(\Delta a, \Delta f).$$

Taking the minimum of all $p \in \tilde{C}_{x,y}$ and taking into account that $V \geq \tilde{V}$ we obtain

$$V(x,y) \geq \tilde{V}(x,y) \geq N \, \min(\Delta a, \Delta f).$$

Finally we have

$$\tilde{V}_{\mathcal{A}}(x,y) \geq \tilde{V}(x,y) \geq N \, \min(\Delta a, \Delta f) \quad \text{and} \quad V_{\mathcal{A}}(x,y) \geq V(x,y) \geq N \, \min(\Delta a, \Delta f).$$

This ends the proof of Lemma 5. $\square$

Remark. Lemmas 4 and 5 show that the costs of good transitions are bounded whereas the costs of the bad ones increase at least linearly with the size of the system. On the basis of the definition of V and $\tilde{V}$ and in view of the proof of these lemmas it is clear that the above result is easier to establish for the cost function $\tilde{V}$. It also turns out that the estimate of the cost of bad transitions with respect to $\tilde{V}$ provides a quick and natural way to estimate their costs with respect to V.

In [2] an inductive proof of this result is presented for the genetic algorithm associated with V and without the equivalence relation considered here. The main contribution here is the extension of the results presented in [2] to any equivalence relation a and to the genetic algorithm associated with the cost function $\tilde{V}$.

On the other hand and in contrast to the inductive proof presented in [2], the approach described here is based on a precise study of the cost of bad or good paths.

The constants $\lambda(a,f), \tilde{\lambda}(a,f)$ represent the difficulty for a population to travel from an equivalence class to better ones. In connection with this remark it is interesting to note that $\tilde{\lambda}(a,f)$ does not depend on the fitness function f and

$$\lambda(a,f) > \tilde{\lambda}(a,f).$$

In other words, it is more difficult for the genetic algorithm associated with V to move from one configuration to a better one. The above observations also imply that the critical value $N(a,f)$ is greater than $\tilde{N}(a,f)$.

Examples. Let us see what happens when our second general model (26) is specialized for the case where the state is

$$S = \{-1,+1\}^{\mathcal{S}}, \qquad \mathcal{S} = [-n,n]^p, \qquad p \geq 1,$$

and the fitness function $f : S \to \mathbb{R}$ is given by

$$f(x) = \frac{1}{2} \sum_{s\in\mathcal{S}} \sum_{s'\in V_s} x(s)\, x(s') + \frac{1}{2} \sum_{s\in\mathcal{S}} x(s),$$

where

$$\forall s \in \mathcal{S} \qquad V_s = \{s' \in \mathcal{S} \ : \ |s_k - s'_k| \leq 1, \ \ 1 \leq k \leq p\}.$$

Let k be the Markovian mutation kernel on S given by

$$k(x_1,x_2) = \frac{1}{|\mathcal{V}(x_1)|}\, 1_{\mathcal{V}(x_1)}(x_2),$$

$$\mathcal{V}(x_1) \overset{\text{def}}{=} \{x_2 \in S \ : \ \text{Card}\{s \in \mathcal{S} \ : \ x_1(s) \neq x_2(s)\} \leq 1\}.$$

Suppose that the function a is given by

$$a(x_1,x_2) = (1 - 1_{x_1}(x_2)) \qquad \forall (x_1,x_2) \in S^2 \ : \ k(x_1,x_2) > 0.$$

Then, one can check that

$$R(a) \leq \max_{x,y} \min_{q\in C_{x,y}} |q| = \text{card}(\mathcal{S}) = (2n+1)^p \qquad \text{and} \qquad \delta(a) = \Delta(a) = 1.$$

Let N be an integer that $N > (2n+1)^p$. The above theorem shows that N individuals will solve the optimization problem when using the genetic algorithm associated with $\tilde{Q}_\beta$.

2.2. Mean cost optimization. In this section we discuss the ways in which the results of section 1 are applied in mean cost optimization problems. Namely, the object will be to find the global minima of a function $V : E \to \mathbb{R}_+$ given by

$$V(x) = E(\mathcal{U}(Z, x)) \qquad \text{or} \qquad V(x) = \min_{g \in G(x)} \sum_{(y,z) \in g} E(\mathcal{V}(Z; y, z)),$$

where

- E is a finite set and $G(x)$ is the set of x-graphs over E,
- Z is a random variable taking value in a finite set F (we denote μ its distribution),
- $U : F \times E \to \mathbb{R}_+$, and $\mathcal{V} : F \times E \times E \times \to \mathbb{R}_+$.

We have seen how to construct a stochastic algorithm converging in probability to global minima of the virtual energy associated with a communication cost function. It is clear from the description above that the appropriate communication cost function is given by

$$V(x, y) = (E(\mathcal{U}(Z; y)) - E(\mathcal{U}(Z; x)))^+ \qquad \text{or} \qquad V(x, y) = E(\mathcal{V}(Z; y, z)).$$

Unfortunately the huge size of the set F often precludes the use of such an algorithm, and the essential problem is to compute a mean cost function at each step. Therefore it is natural to choose, for instance, a Markovian kernel K which ensures that

$$V_{\gamma_t}(x, y) = \frac{1}{t^A} \int_0^{t^A} \mathcal{V}(Z_s; x, y)\, ds \xrightarrow[t \to +\infty]{} V(x, y) = E(\mathcal{V}(Z; x, y)) \qquad \text{P.a.e.}$$

$$\text{or } V_{\gamma_t}(x, y) = \left(\frac{1}{t^A} \int_0^{t^A} \mathcal{U}(Z_s; y)\, ds - \frac{1}{t^A} \int_0^{t^A} \mathcal{U}(Z_s; x)\, ds \right)^+$$

$$\xrightarrow[t \to +\infty]{} V(x, y) = (E(\mathcal{U}(Z; y)) - E(\mathcal{U}(Z; x)))^+ \qquad \text{P.a.e.},$$

where

- $V_\gamma \overset{\text{def}}{=} 1/e^\gamma \int_0^{e^\gamma} \mathcal{V}(Z_s; x, y)\, ds$ or
  $V_\gamma \overset{\text{def}}{=} \left(\frac{1}{e^\gamma} \int_0^{e^\gamma} \mathcal{U}(Z_s; y)\, ds - \frac{1}{e^\gamma} \int_0^{e^\gamma} \mathcal{U}(Z_s; x)\, ds \right)^+$,
- $\gamma_t = A \log t$,
- Z_s is a time-homogeneous Markov process associated with the generator $\mathcal{L} = K - I$,
- μ is an invariant measure of $\mathcal{L}$.

Before starting the description of our stochastic algorithm, we give some details about the above convergences.

LEMMA 6. *Let K be a an irreducible transition kernel with unique invariant measure μ. For every $x, y \in E$ and $A > 0$ we have*

$$\lim_{t \to +\infty} \sqrt{\frac{t^A}{\log t}}\, |V_{\gamma_t}(x, y) - V(x, y)| = 0 \qquad \textit{P.a.e.}$$

Proof. In this situation it is well known that for every $x, y \in E$ there exists a bounded function $F(.; x, y)$ such that

$$\mathcal{V}(.; x, y) - \mu(\mathcal{V}(.; x, y)) = \mathcal{L}(\mathrm{F}(.; x, y)) \qquad \text{and} \qquad \mu(\mathrm{F}(.; x, y) = 0.$$

This equation is the Poisson equation associated with $\mathcal{V}(.;x,y)$ and $\mathcal{L}$. Thus, one gets the decomposition

$$\frac{1}{t}\int_0^t (\mathcal{V}(Z_s;x,y)-\mu(\mathcal{V}(.;x,y)))\,ds=\frac{1}{t}(F(Z_t;x,y)-F(Z_0;x,y)-M(x,y)_t),$$

where $M(x,y)$ is a martingale with angle bracket $\langle M(x,y)\rangle$ given by

$$\langle M(x,y)\rangle_t=\sum_{z\in F}\int_0^t (F(z;x,y)-F(Z_s;x,y))^2\,K(Z_s,z)\,ds.$$

Since the function F is bounded we have $\langle M(x,y)\rangle_t \le c\ t$ for some nonnegative constant. Finally, by the standard iterated-log law and since jumps are bounded, it follows that

$$\frac{1}{t}M(x,y)_t\le\frac{\sqrt{2\,c\,t\log\log(c\,t)}}{t}$$

and thus $\lim_{t\to+\infty}\sqrt{\frac{t}{\log t}}\left|\frac{1}{t}M(x,y)_t\right|=0$. This ends the proof. □

Remark. The Poisson equation is a standard tool in the study of Markov processes. For instance it was also used by Younes [13] to study the convergence of a stochastic gradient algorithm to a maximum likelihood estimator. The context of Younes is more complex than those considered here, and the speed of convergence cannot be obtained by a mere application of the iterated logarithm law as before. But Younes also noticed that if the convergence is fast enough (in a negative power in time), then one can couple the estimation procedure to a simulated annealing algorithm (with the classical reversibility conditions) to get the global minima of a function depending on the parameter to be estimated. To do this, Younes uses the Dobrushin coefficients, but the entropy approach enables one to get more precise results on the admissible logarithmic schedules of temperature (the constant c given below).

Let us fix some terminology.

- Let $(\Omega^{(Z)},P^{(Z)},F_t^{(Z)},Z_t)$ be the canonical process associated with the generator $\mathcal{L}$.
- For a given probability measure m on E, $\beta,\gamma\in C^1(\mathbb{R}_+,\mathbb{R}_+)$ and given the Markov process Z we note $(\Omega_{(Z)},P_{(Z)},F_{(Z),t},X_t)$ the canonical process associated with the family of generators $(L_{\gamma_t,\beta_t})_{t\ge0}=(Q_{\gamma_t,\beta_t}-I)_{t\ge0}$ whose initial condition is $m_0=m$, and we note m_t the distribution of X_t, where

$$Q_{\gamma,\beta}(x,y)=q(x,y)\,e^{-\beta\,V_\gamma(x,y)}\qquad\text{with}\quad q\ \text{irreducible.}$$

- To capture all randomness we note $\Omega=\Omega^{(Z)}\times\Omega_{(Z)}$, $F_t=F_t^{(Z)}\times F_{(Z),t}$, and we define P as follows:

$$\forall A\in F_{(Z),t}\in\qquad\forall B\in F_t^{(Z)}\qquad P(A\times B)=\int_B P_{(Z)}(A)\,dP^{(Z)}.$$

The above lemma and Corollary 4 lead us to the following proposition.

PROPOSITION 5. *Let us set* $c=\limsup_{\gamma\to+\infty}c(\gamma)<+\infty$ *P.a.e., where* $c(\gamma)$ *is the critical height associated with the communication cost* V_γ.

When the inverse-freezing schedule has parametric form $\beta_t=K^{-1}\log t$, *for* t *sufficiently large and* $K>c$, *we have*

$$\lim_{t\to+\infty}Ent_{\pi_{\beta_t}}(m_t)=0\quad P.a.e.\qquad\text{and}\qquad\lim_{t\to+\infty}P(X_t\in V^\star)=1,\tag{38}$$

where π_β is the unique invariant probability of $L_\beta = Q_\beta - I$ with

$$Q_\beta(x,y) = q(x,y)\, e^{-\beta V(x,y)}.$$

In many practical situations we also want a quantitative measure of the convergence (38). Unfortunately our method of proof is not suitable for estimating such quantitative behavior. In our settings a natural alternative approach is to look at the convergence of the mean value of the process $t \to \mathrm{Ent}_{\pi_{\beta_t}}(m_t)$ with respect to the random media (given by Z). In view of the inequality (20) we immediately observe that the speed of convergence of the mean value is related to the speed of convergence of the mean values of $\mathrm{Ent}_{\pi_{\gamma_t,\beta_t}}(m_t)$ and $|V_{\gamma_t}(x,y) - V(x,y)|$. The first term, linked to the critical height $c(\gamma_t)$ and to the derivative of γ_t, depends in a complicated way on the constant A, but we know that it is a nondecreasing function of the parameter A. On the other hand, the second term is a nonincreasing function of the parameter A. If we know how these quantities are linked to A a good adjustment of this parameter is then related to a classical minimization problem. We will examine this quantitative behavior in a forthcoming paper.

REFERENCES

[1] R. Bott and J.P. Mayberry, *Matrices and trees*, in Economics Activity Analysis, John Wiley, New York, 1954.

[2] R. Cerf, *Une théorie asymptotique des algorithmes génétiques*, Thése de Doctorat, Université Montpellier II, Montpellier, France, 1994.

[3] P. del Moral, *Nonlinear filtering: Interacting particle resolution*, Markov Process. Related Fields, 2 (1996), pp. 555–579.

[4] J.D. Deuschel and C. Mazza, *L^2-Convergence of time nonhomogeneous Markov processes* I: *Spectral estimates*, Ann. Appl. Probab., 4 (1994), pp. 1012–1056.

[5] P. Diaconis and L. Saloff-Costes, *Logarithmic Sobolev inequalities for finite Markov chains*, Ann. Appl. Probab., 6 (1996), pp. 695–750.

[6] M.I. Freidlin and A.D. Wentzell, *Random Perturbations of Dynamical Systems*, Springer-Verlag, Berlin, New York, 1984.

[7] A. Frigerio and G. Grillo, *Simulated annealing with time-dependent energy function*, Math. Z., 213 (1993), pp. 97–116.

[8] R. Holley and D. Stroock, *Annealing via Sobolev inequalities*, Commun. Math. Phy., 115 (1988), pp. 553–569.

[9] M. Löwe, *Simulated annealing with time-independent energy function via Sobolev inequalities*, Stochastic Process. Appl., 63 (1996), pp. 221–233.

[10] L. Miclo, *Recuit simulé sans potentiel sur un ensemble fini*, in Séminaire de Probabilités, Lecture Notes in Math. 1526, Springer-Verlag, Berlin, New York, 1992, pp. 47–60.

[11] D. Stroock, *Logarithmic Sobolev inequalities for Gibbs states*, in Dirichlet Forms, G. Dell'Antonio and U. Mosco, eds., Lecture Notes in Math. 1563, Springer-Verlag, Berlin, New York, 1993, pp. 194–228.

[12] A. Trouvé, *Parallélisation massive du recuit simulé*, Thèse de Doctorat, Université Paris XI, Paris, France, 1993.

[13] L. Younes, *Estimation and annealing for Gibbsian fields*, Ann. Inst. H. Poincaré Probab. Statist., 24 (1988), pp. 269–294.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 4, pp. 1251–1272

ON THE TOPOLOGICAL DERIVATIVE IN SHAPE OPTIMIZATION*

J. SOKOŁOWSKI† AND A. ŻOCHOWSKI‡

Abstract. In this paper the topological derivative for an arbitrary shape functional is defined. Examples are provided for elliptic equations and the elasticity system in the plane. The topological derivative can be used for solving shape optimization problems in structural mechanics.

Key words. shape optimization, shape derivative, elasticity system, topological derivative, asymptotic expansion

AMS subject classifications. Primary, 49Q10, 49Q12; Secondary, 35J05, 35J50, 35B37

PII. S0363012997323230

1. Introduction.

The topological derivative for a shape functional is defined in the following way [10].

Assume that $\Omega \subset \mathbb{R}^N$ is an open set and that there is given a shape functional

$$\mathcal{J} \ : \ \Omega \setminus K \to \mathbb{R}$$

for any compact subset $K \subset \overline{\Omega}$. We denote by $B_\rho(x), x \in \Omega$, the ball of radius $\rho > 0$, $B_\rho(x) = \{y \in \mathbb{R}^N | \|y - x\| < \rho\}$, $\overline{B_\rho(x)}$ is the closure of $B_\rho(x)$, and assume that there exists the following limit

$$\mathfrak{T}(x) = \lim_{\rho \downarrow 0} \frac{\mathcal{J}(\Omega \setminus \overline{B_\rho(x)}) - \mathcal{J}(\Omega)}{|\overline{B_\rho(x)}|},$$

which can be defined in an equivalent way by

$$\tilde{\mathfrak{T}}(x) = \lim_{\rho \downarrow 0} \frac{\mathcal{J}(\Omega \setminus \overline{B_\rho(x)}) - \mathcal{J}(\Omega)}{\rho^N}.$$

The function $\mathfrak{T}(x), x \in \Omega$ is called the topological derivative of $\mathcal{J}(\Omega)$ and provides the information on the infinitesimal variation of the shape functional $\mathcal{J}$ if a small hole is created at $x \in \Omega$. We shall show in the sequel that the method is constructive; i.e., the topological derivative can be evaluated for shape functionals depending on solutions of elliptic equations defined in Ω.

The partial differential equation for $u_\rho = u_{\Omega_\rho}$ is called the state equation for the shape optimization problems under consideration. We show that for a class of shape functionals it is sufficient to solve in the unperturbed domain Ω the state equation as well as the appropriate adjoint state equation in order to evaluate the topological

*Received by the editors June 20, 1997; accepted for publication (in revised form) September 14, 1998; published electronically May 28, 1999. Partially supported by grant 8T1100815 of the State Committee for the Scientific Research of the Republic of Poland and by the French–Polish research programme POLONIUM between INRIA–Lorraine and Systems Research Institute of the Polish Academy of Sciences.

http://www.siam.org/journals/sicon/37-4/32323.html

†Institut Elie Cartan, Laboratoire de Mathématiques, Université Henri Poincaré Nancy I, B.P. 239, 54506 Vandoeuvre-lès-Nancy Cedex, France, and Systems Research Institute of the Polish Academy of Sciences, ul. Newelska 6, 01-447 Warszawa, Poland (sokolows@iecn.u-nancy.fr).

‡Systems Research Institute of the Polish Academy of Sciences, ul. Newelska 6, 01-447 Warszawa, Poland (zochowsk@ibspan.waw.pl).

derivative $\mathfrak{T}(x), x \in \Omega$. This means that the derivative can be used in shape optimization for broad classes of shape functionals and partial differential equations. Some examples, where the derivative is explicitly given for model problems, are provided.

Our results can be described in the following way. For the shape functional $\mathcal{J}(\Omega \setminus \overline{B_\rho(x)})$ we introduce the function of the small parameter $\rho \geq 0$ of the form $J(\rho) = \mathcal{J}(\Omega \setminus \overline{B_\rho(x)})$ and determine for $N = 2$ the second-order derivative $J''(0^+)$. Therefore, the following expansion is obtained:

$$\mathcal{J}(\Omega_\rho) = \mathcal{J}(\Omega) + \frac{\rho^2}{2} J''(0^+) + o(\rho^2) \ .$$

In the very special case of the energy functional, the so-called compliance functional in linear elasticity, the topological derivative is in fact considered in [8]. The derivative is used in numerical methods of optimal design for the specific choice of shape functional [8]. In order to differentiate the energy functional with respect to the variations of the boundary of the domain of integration the knowledge of the shape derivative of the state equation with respect to the boundary variations is not required. Therefore, the results obtained for the particular case of the energy functional cannot be directly generalized to the case of an arbitrary shape functional.

In this paper the derivative is defined for an arbitrary shape functional and evaluated for solutions of scalar elliptic equations and the system of elasticity in the plane.

2. Elliptic equation in $\mathbb{R}^2$. Assume that $\Omega \subset \mathbb{R}^2$ is a bounded domain with the boundary $\partial\Omega = \Gamma_1 \cup \Gamma_2$, $0 \in \Omega$. Let $K = [k_{ij}]_{2\times 2}, k_{ij} \in \mathbb{R}, i, j = 1, 2$, be a symmetric positive definite matrix.

We consider the following elliptic equation with nonhomogeneous Dirichlet–Neumann boundary conditions:

$$\begin{aligned} \operatorname{div}(K \cdot \nabla u) &= f \quad \text{in} \quad \Omega, \\ u &= g \quad \text{on} \quad \Gamma_1, \\ \frac{\partial u}{\partial n_K} &= h \quad \text{on} \quad \Gamma_2. \end{aligned} \tag{2.1}$$

Let λ_1, λ_2 be the eigenvalues of K, ξ^1, $\xi^2 \in \mathbb{R}^2$ the corresponding eigenvectors, i.e., $K \cdot \xi^i = \lambda_i \xi^i$, $i = 1, 2$, and $R_\lambda = [\xi^1, \xi^2]_{2\times 2}$, a rotation matrix consisting of the eigenvectors. Using the matrix R_λ the following ellipse $E_\rho \subset \mathbb{R}^2$ depending on the small parameter $\rho > 0$ is defined:

$$E_\rho = \left\{ x = (x_1, x_2) \,|\, x = R_\lambda \cdot y \ , \ y = (y_1, y_2) \ , \ \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} \leq \rho^2 \right\}.$$

For sufficiently small $\rho > 0$ it is always possible to remove $\overline{E_\rho}$ from Ω, obtaining

$$\Omega_\rho = \Omega \setminus \overline{E_\rho} \ , \quad \partial\Omega_\rho = \partial\Omega \cup \partial E_\rho.$$

In such a domain we define the following system:

$$(\mathcal{P}(\Omega_\rho)) \qquad \left\{ \begin{array}{ccl} \operatorname{div}(K \cdot \nabla u_\rho) & = & f \quad \text{in} \quad \Omega_\rho, \\ u_\rho & = & g \quad \text{on} \quad \Gamma_1, \\ \frac{\partial u_\rho}{\partial n_K} & = & h \quad \text{on} \quad \Gamma_2, \\ \frac{\partial u_\rho}{\partial n_K} & = & 0 \quad \text{on} \quad \Gamma_\rho = \partial E_\rho, \end{array} \right.$$

which coincides with (2.1) for $\rho = 0$.

The shape functionals we shall consider have the following form:

$$\mathcal{J}_1(\Omega_\rho) = J_u(\rho) = \int_{\Omega_\rho} F(u_\rho)\, d\Omega, \tag{2.2}$$

$$\mathcal{J}_2(\Omega_\rho) = J_g(\rho) = \int_{\Omega_\rho} [\nabla u_\rho \cdot K \cdot \nabla u_\rho]^p\, d\Omega, \tag{2.3}$$

where $p \geq 1$ is selected in such a way that (2.3) is well defined, and F is a C^2 function of its argument, and, e.g., $|F(u)| \leq C_1|u|^2, |F''(u)| \leq C_2$, to assure the differentiability of the functional (2.2). The value of p depends on the types of admissible domains and the regularity of boundary data. We distinguish two typical cases of nonsmooth domains for which the results are applicable.

(A1) Pure cracks are admissible, even having different types of boundary conditions on both edges (i.e., Neumann and Dirichlet). Then $p = 1$ and g, h must be compatible with $u \in H^1(\Omega)$, which means that the boundary data g, h are selected in such a way that the solution to (2.1) is a weak solution in the Sobolev space $H^1(\Omega)$.

(A2) Reentrant corners with $\alpha < 2\pi$ are admissible and the same types of boundary conditions on both edges (Neumann–Neumann or Dirichlet–Dirichlet) are prescribed. Then $p = 2$ and g, h must be compatible with $u \in W^1_4(\Omega)$.

We refer the reader to [2] for the regularity of solutions to the elliptic equations in nonsmooth domains. Observe that the interior regularity of u in Ω is determined by the regularity of the right-hand side f for elliptic equations. The rather restrictive assumption $f \in C^1(\Omega)$ is sufficient for our purposes, but it is not optimal. On the other hand the formulas (2.4), (2.5), defined below at $x_0 = 0$, formally can be used to define functions $J''_u(x), J''_g(x), x \in \Omega$, which have the following property:

$$J''_u(x), J''_g(x) \in L^1_{\text{loc}}(\Omega)$$

for $u, v, w \in H^1(\Omega)$, $p = 1$, and $f \in L^2(\Omega)$.

The following form of topological derivatives is obtained.

Theorem 2.1. *Assume that $f \in C^1(\Omega)$ and the boundary data (g, h) satisfy* (A1) *or* (A2); *then*

$$J''_u(0) = -2\pi\sqrt{\lambda_1\lambda_2}[\,F(u(0)) + f(0)w(0) + 2\nabla u \cdot K \cdot \nabla w|_{x=0}\,], \tag{2.4}$$

and

$$J''_g(0) = -2\pi\sqrt{\lambda_1\lambda_2}[k(p)\|\nabla u \cdot K \cdot \nabla u\|^{2p}_{x=0} + f(0)v(0) + 2(\nabla u \cdot K \cdot \nabla v)|_{x=0}], \tag{2.5}$$

where the coefficient $k(p)$ takes the values

$$k(1) = 2, \qquad k(2) = 6.$$

The functions w, v are the adjoint state variables defined by (2.28), (2.29), *respectively.*

Remark 1. From (2.4), (2.5) it follows that the topological derivatives for the shape functionals (2.2), (2.3) take the form at $x = 0$,

$$\mathfrak{T}_1(0) = -\sqrt{\lambda_1\lambda_2}[\,F(u(0)) + f(0)w(0) + 2\nabla u \cdot K \cdot \nabla w|_{x=0}\,],$$

and

$$\mathfrak{T}_2(0) = -\sqrt{\lambda_1\lambda_2}[k(p)\|\nabla u \cdot K \cdot \nabla u\|^{2p}_{x=0} + f(0)v(0) + 2\nabla u \cdot K \cdot \nabla v|_{x=0}] \ ,$$

respectively.

Proof. The proof is divided into three steps. The first step consists in transformation of $(\mathcal{P}(\Omega_\rho))$ defined in $\Omega_\rho = \Omega \setminus \overline{E_\rho}$, into the simpler elliptic equation defined in the domain $\Omega_\rho = \Omega \setminus \overline{B_\rho}$ by using an appropriate change of variables. Here we denote $B_\rho = B_\rho(0)$, $0 \in \Omega$. In the second and third steps the formulas are derived for the latter equation and then translated to the original problem by the inverse change of variables.

Step 1. Let us make the substitution $y = Bx$, where $B = \Lambda^{-1/2}R_\lambda$, and $\Lambda = \operatorname{diag}(\lambda_1, \lambda_1)$. Since $\nabla_x = B^T\nabla_y$, (2.1) is transformed to the Laplace equation, the ellipse E_ρ is transformed onto the ball B_ρ centered at 0, and the resulting domain is $\Omega_\rho = \Omega \setminus \overline{B_\rho}$. To keep notation simple, we shall use the same notation for the transformed problem as for (2.1). Therefore, the transformed state equation takes the following form:

$$\begin{aligned} \Delta u &= f \quad \text{in} \quad \Omega, \\ u &= g \quad \text{on} \quad \Gamma_1, \\ \frac{\partial u}{\partial n} &= h \quad \text{on} \quad \Gamma_2. \end{aligned} \tag{2.6}$$

The corresponding equation in the domain Ω_ρ with the hole B_ρ has the form

$$\begin{aligned} \Delta u_\rho &= f \quad \text{in} \quad \Omega_\rho, \\ u_\rho &= g \quad \text{on} \quad \Gamma_1, \\ \frac{\partial u_\rho}{\partial n} &= h \quad \text{on} \quad \Gamma_2, \\ \frac{\partial u_\rho}{\partial n} &= 0 \quad \text{on} \quad \Gamma_\rho = \partial B_\rho. \end{aligned} \tag{2.7}$$

The resulting shape functionals after the change of variables take the form

$$J_u(\rho) = \sqrt{\lambda_1\lambda_2}\int_{\Omega_\rho} F(u_\rho)\, d\Omega, \tag{2.8}$$

$$J_g(\rho) = \sqrt{\lambda_1\lambda_2}\int_{\Omega_\rho} [\nabla u_\rho \cdot \nabla u_\rho]^p\, d\Omega. \tag{2.9}$$

This is due to the fact that $K = R_\lambda \Lambda R_\lambda^T$. To make the notation still simpler, we shall compute derivatives of the following functionals:

$$I_u(\rho) = \int_{\Omega_\rho} F(u_\rho)\, d\Omega, \tag{2.10}$$

$$I_g(\rho) = \int_{\Omega_\rho} [\nabla u_\rho \cdot \nabla u_\rho]^p\, d\Omega. \tag{2.11}$$

Step 2. In the sequel we denote by $(\cdot)'$ the derivative $\partial(\cdot)/\partial\rho$, which can be considered as a particular case of the shape derivative. We refer the reader to [9] for the details on the shape differentiability of integral shape functionals and solutions to partial differential equations of elliptic type.

By an application of (C.2) it follows that

$$I_u'(\rho) = \int_{\Omega_\rho} F_u'(u_\rho)u_\rho' \, d\Omega - \int_{\Gamma_\rho} F(u_\rho)\, dS, \tag{2.12}$$

$$I_g'(\rho) = \int_{\Omega_\rho} 2p||\nabla u_\rho||^{2p-2}(\nabla u_\rho \cdot \nabla u_\rho')\, d\Omega - \int_{\Gamma_\rho} \left(\frac{\partial u_\rho}{\partial \tau}\right)^{2p} dS. \tag{2.13}$$

The weak solution $u_\rho \in H_g^1(\Omega_\rho)$ to (2.7) satisfies the integral identity

$$\int_{\Omega_\rho} \nabla u_\rho \cdot \nabla \phi \, d\Omega = \int_{\Gamma_2} h\phi \, dS - \int_{\Omega_\rho} f\phi \, d\Omega \quad \forall \phi \in H_{\Gamma_1}^1(\Omega_\rho), \tag{2.14}$$

where for $\rho \geq 0$ such that $B_\rho \subset \Omega$,

$$H_g^1(\Omega_\rho) = \{\psi \in H^1(\Omega_\rho) \mid \psi = g \quad \text{on} \quad \Gamma_1\},$$

$$H_{\Gamma_1}^1(\Omega_\rho) = \{\psi \in H^1(\Omega_\rho) \mid \psi = 0 \quad \text{on} \quad \Gamma_1\},$$

and we use the convention that the restriction to Ω_ρ of a function $\phi \in H_{\Gamma_1}^1(\Omega)$ is denoted by ϕ.

The strong shape derivative u_ρ' of the solution u_ρ to (2.14) is defined by the relation

$$u_\rho' = \dot{u}_\rho - \nabla u_\rho \cdot V \ ,$$

where $\dot{u}_\rho$ is the material derivative, and V is an appropriate vector field [9]. Furthermore, $u_\rho' \in H_{\Gamma_1}^1(\Omega_\rho)$ satisfies the integral identity [9],

$$\int_{\Omega_\rho} \nabla u_\rho' \cdot \nabla \phi \, d\Omega - \int_{\Gamma_\rho} \frac{\partial u_\rho}{\partial \tau}\frac{\partial \phi}{\partial \tau}\, dS = \int_{\Gamma_\rho} f\phi \, dS, \tag{2.15}$$

for all test functions $\phi \in H_{\Gamma_1}^1(\Omega_\rho) \cup H^2(\Omega_\rho)$. The regularity of u_ρ' is determined by the regularity of $\dot{u}_\rho$ and the gradient ∇u_ρ since V is sufficiently regular.

The adjoint state equation for the functional I_u is defined as follows: find $w_\rho \in H_{\Gamma_1}^1(\Omega_\rho)$ such that

$$-\int_{\Omega_\rho} \nabla w_\rho \cdot \nabla \phi \, d\Omega = \int_{\Omega_\rho} F_u'(u_\rho)\phi \, d\Omega \quad \forall \phi \in H_{\Gamma_1}^1(\Omega_\rho); \tag{2.16}$$

and for the functional I_g, find $v_\rho \in H_{\Gamma_1}^1(\Omega_\rho)$ such that

$$-\int_{\Omega_\rho} \nabla v_\rho \cdot \nabla \phi \, d\Omega = \int_{\Omega_\rho} 2p||\nabla u_\rho||^{2p-2}(\nabla u_\rho \cdot \nabla \phi)\, d\Omega \quad \forall \phi \in H_{\Gamma_1}^1(\Omega_\rho). \tag{2.17}$$

The strong shape derivative $w_\rho' \in H_{\Gamma_1}^1(\Omega_\rho)$ of the solution w_ρ to (2.16) satisfies the following integral identity [9]:

$$-\int_{\Omega_\rho} \nabla w_\rho' \cdot \nabla \phi \, d\Omega + \int_{\Gamma_\rho} \frac{\partial w_\rho}{\partial \tau}\frac{\partial \phi}{\partial \tau}\, dS = \int_{\Omega_\rho} F_u''(u_\rho)u_\rho'\phi \, d\Omega - \int_{\Gamma_\rho} F_u'(u_\rho)\phi \, dS \tag{2.18}$$

for all test functions $\phi \in H^1_{\Gamma_1}(\Omega_\rho) \cup H^2(\Omega_\rho)$.

The strong shape derivative $v'_\rho \in H^1_{\Gamma_1}(\Omega_\rho)$ of the solution v_ρ to (2.17) is defined in the similar way:

$$\begin{aligned}(2.19)\quad -\int_{\Omega_\rho} \nabla v'_\rho \cdot \nabla \phi \, d\Omega + \int_{\Gamma_\rho} \frac{\partial v_\rho}{\partial \tau}\frac{\partial \phi}{\partial \tau}\, dS = -\int_{\Gamma_\rho} 2p\|\nabla u_\rho\|^{2p-2}\frac{\partial u_\rho}{\partial \tau}\frac{\partial \phi}{\partial \tau}\, dS \\ + \int_{\Omega_\rho} 2p||\nabla u_\rho||^{2p-2}(\nabla u'_\rho \cdot \nabla \phi)\, d\Omega \\ + \int_{\Omega_\rho} 2p(2p-2)||\nabla u_\rho||^{2p-4}(\nabla u'_\rho \cdot \nabla u_\rho)(\nabla u_\rho \cdot \nabla \phi)\, d\Omega\end{aligned}$$

for all test functions $\phi \in H^1_{\Gamma_1}(\Omega_\rho) \cup H^2(\Omega_\rho)$. Using $\phi = u'_\rho \in H^1_{\Gamma_1}(\Omega_\rho)$ as a test function, the following form of the derivatives (2.12), (2.13) is obtained:

$$(2.20)\qquad I'_u(\rho) = -\int_{\Gamma_\rho} \left[F(u_\rho) + f w_\rho + \frac{\partial u_\rho}{\partial \tau}\frac{\partial w_\rho}{\partial \tau}\right] dS,$$

$$(2.21)\qquad I'_g(\rho) = -\int_{\Gamma_\rho} \left[\left(\frac{\partial u_\rho}{\partial \tau}\right)^{2p} + f v_\rho + \frac{\partial u_\rho}{\partial \tau}\frac{\partial v_\rho}{\partial \tau}\right] dS.$$

Since all integrands are bounded,

$$(2.22)\qquad \lim_{\rho \to 0+} I'_u(\rho) = \lim_{\rho \to 0+} I'_g(\rho) = 0.$$

By differentiating (2.21) once more, in view of (C.1) we get

$$\begin{aligned} I''_u(\rho) &= \int_{\Gamma_\rho} \left[\frac{\partial F(u_\rho)}{\partial n} + \frac{\partial (f w_\rho)}{\partial n} + \frac{\partial}{\partial n}\left(\frac{\partial u_\rho}{\partial \tau}\frac{\partial w_\rho}{\partial \tau}\right)\right] dS \\ &- \int_{\Gamma_\rho} \left[F'_u(u_\rho)u'_\rho + f w'_\rho + \left(\frac{\partial u_\rho}{\partial \tau}\frac{\partial w_\rho}{\partial \tau}\right)'\right] dS \\ &- \frac{1}{\rho}\int_{\Gamma_\rho} \left[F(u_\rho) + f w_\rho + \frac{\partial u_\rho}{\partial \tau}\frac{\partial w_\rho}{\partial \tau}\right] dS \\ (2.23)\qquad &= I_1(\rho) + I_2(\rho) + I_3(\rho). \end{aligned}$$

Observe, that $\frac{\partial}{\partial n} = -\frac{\partial}{\partial r}$ on Γ_ρ. Now, according to (A.4),

$$(2.24)\qquad \frac{\partial u_\rho}{\partial \tau} = \frac{1}{r}\frac{\partial u_\rho}{\partial \theta} = -a\left(\frac{\rho^2}{r^2} + 1\right)\sin\theta + b\left(\frac{\rho^2}{r^2} + 1\right)\cos\theta + O(\rho^{1-\epsilon}).$$

Hence

$$\frac{\partial}{\partial n}\frac{\partial u_\rho}{\partial \tau} = -2a\frac{\rho^2}{r^3}\sin\theta + 2b\frac{\rho^2}{r^3}\cos\theta + O(\rho^{-\epsilon}) \underset{r=\rho}{=} -\frac{2a}{\rho}\sin\theta + \frac{2b}{\rho}\cos\theta + O(\rho^{-\epsilon}).$$

Similarly,

$$\frac{\partial}{\partial \rho}\left(\frac{\partial u_\rho}{\partial \tau}\right) = -2a\frac{\rho}{r^2}\sin\theta + 2b\frac{\rho}{r^2}\cos\theta + O(\rho^{-\epsilon}) \underset{r=\rho}{=} -\frac{2a}{\rho}\sin\theta + \frac{2b}{\rho}\cos\theta + O(\rho^{-\epsilon}).$$

Taking this into account leads to

$$\frac{\partial}{\partial n}\left(\frac{\partial u_\rho}{\partial \tau}\frac{\partial w_\rho}{\partial \tau}\right) - \left(\frac{\partial u_\rho}{\partial \tau}\frac{\partial w_\rho}{\partial \tau}\right)' = O(\rho^{-\epsilon}),$$

and the first two integrals cancel out,

$$\lim_{\rho\to 0+}[I_1(\rho) + I_2(\rho)] = 0.$$

We use for w_ρ the following expansion, the existence of which is proven in the Appendix:

$$w_\rho = w(0) + c\left(\frac{\rho^2}{r} + r\right)\cos\theta + d\left(\frac{\rho^2}{r} + r\right)\sin\theta + O(\rho^{2-\epsilon}). \tag{2.25}$$

Taking into account (2.24),

$$\lim_{\rho\to 0+} I_3(\rho) = -2\pi F(u(0)) - 2\pi f(0)w(0) - 4\pi(ac + bd).$$

As a result

$$I_u''(0) = -2\pi[\, F(u(0)) + f(0)w(0) + 2(\nabla u \cdot \nabla w)|_{x=0}\,], \tag{2.26}$$

and similarly

$$I_g''(0) = -2\pi[\, k(p)\,||\nabla u(0)||^{2p} + f(0)v(0) + 2(\nabla u \cdot \nabla v)|_{x=0}\,], \tag{2.27}$$

where the coefficient $k(p)$ takes on the values $k(1) = 2$, $k(2) = 6$. For $\rho = 0$ the adjoint state variables w, v, satisfy the integral identities

$$w \in H^1_{\Gamma_1}(\Omega) : \int_\Omega \nabla w \cdot \nabla\phi \, d\Omega = -\int_\Omega F_u'(u)\phi \, d\Omega, \tag{2.28}$$

$$v \in H^1_{\Gamma_1}(\Omega) : \int_\Omega \nabla v \cdot \nabla\phi \, d\Omega = -\int_\Omega 2p||\nabla u||^{2p-2}(\nabla u \cdot \nabla\phi)\, d\Omega \tag{2.29}$$

for all test functions $\phi \in H^1_{\Gamma_1}(\Omega)$.

In a special case, for $p = 1$, $\Gamma_2 = \emptyset$, and $g = 0$, it follows that

$$\int_\Omega \nabla v \cdot \nabla\phi \, d\Omega = \int_\Omega f\phi \, d\Omega \quad \forall\phi \in H^1_0(\Omega);$$

hence $v = 2u$. The function $k(p)$ is obtained by the integration

$$\int_0^{2\pi} (-2a\sin\theta + 2b\cos\theta)^{2p}\, d\theta = k(p)(a^2 + b^2)^p.$$

Step 3. The proof is completed by the change of variables $x = B^{-1}y$. □

The matrix K in the definition of J_g may be replaced in fact by an arbitrary matrix, say H. However, in such a case it is not possible in general to get a simple closed form of the expression

$$A(u,p) = \lim_{\rho\to 0+}\frac{1}{\rho}\int_{\Gamma_\rho} [\nabla u_\rho \cdot \tilde{H} \cdot \nabla u_\rho]^p \, dS,$$

where $\tilde{H} = BHB^T$. We must introduce locally the orthogonal coordinate system (e_r, e_θ); see the expansions in elasticity in the Appendix. Denote $c = \cos\theta,\ s = \sin\theta$; it follows that in this frame of reference the matrix $\tilde{H}$ transforms like a second-order tensor $\tilde{H} \to \hat{H} = R(\theta)\tilde{H}R(\theta)^T$, where

$$R(\theta) = \left[\begin{array}{cc} c & -s \\ s & c \end{array}\right].$$

After substituting the expansion for u_ρ and keeping in mind that $\frac{\partial u_\rho}{\partial r} = 0$ on Γ_ρ, so that $\nabla u_\rho = [0, \frac{\partial u_\rho}{\partial \tau}]^T$ on Γ_ρ, we get

$$A(u,p) = 4\int_0^{2\pi} [\hat{h}_{22}(-as+bc)^2]^p \, d\theta$$
$$= 4\int_0^{2\pi} [(\tilde{h}_{11}c^2 - 2\tilde{h}_{12}cs + \tilde{h}_{22}s^2)(-as+bc)^2]^p \, d\theta.$$

Having computed the integral, we must again express it in terms of H. The assumption $\tilde{H} = I$ is adopted in the paper in order to simplify the obtained formulas.

Remark 2. Let us note that the condition $J'(0) = 0$ depends on the shape functional. If a shape functional depends actually on the boundary $\partial\Omega$ of the domain Ω, then in general $J'(0) \neq 0$. For $\mathcal{J}(\partial\Omega) = \int_{\partial\Omega} = |\partial\Omega|$ we have $\mathcal{J}(\partial\Omega_\rho) = |\partial\Omega| + 2\pi\rho$ and

$$\frac{\mathcal{J}(\partial\Omega_\rho) - \mathcal{J}(\partial\Omega)}{|\partial B_\rho(x)|} = 1\ .$$

Remark 3. The following function is used for the definition of the so-called Morrey spaces $L^{p,\lambda}(\Omega)$, $p \geq 1, \lambda \geq 0$,

$$\mathfrak{g}(x) = \sup_{0<\rho<1} \rho^{-\lambda} \int_{B_\rho(x)} |u|^p dx = \sup_{0<\rho<1} \rho^{-\lambda} \left[\int_\Omega |u|^p dx - \int_{\Omega\setminus\overline{B_\rho(x)}} |u|^p dx\right]$$

for $u \in L^p(\Omega)$; see, e.g., [1] for details. However, the function $\mathfrak{g}(x)$ is not useful in applications to the shape optimization.

Let us point out that the difference between the topological derivative and the function $\mathfrak{g}(x)$ is substantial, since for our applications the function $u = u_\rho$ is given by a solution of the partial differential equation defined in the domain $\Omega_\rho = \Omega \setminus \overline{B_\rho(x)}$ and we would rather consider, e.g., the function

$$\mathfrak{h}(x) = \sup_{0<\rho<1} \rho^{-\lambda} \left[\int_\Omega |u_\Omega|^p dx - \int_{\Omega_\rho} |u_{\Omega_\rho}|^p dx\right]\ .$$

3. Test cases for Laplace equation. The explicit formulas for the derivatives obtained in the previous section are presented for three examples.

Example 1. Let $\Omega = B_R(0)$, and $u(x,y) = x$, so that

$$\Delta u = 0 \quad \text{in} \ \ \Omega, \qquad u = x \quad \text{on} \ \ \partial\Omega.$$

The solution u_ρ to $(\mathcal{P}(\Omega_\rho))$ takes the form

$$u_\rho = \frac{R^2}{R^2+\rho^2}\left(\frac{\rho^2}{r} + r\right)\cos\theta,$$

and the adjoint state w is given by

$$w = \frac{1}{4}(r^3 - R^2 r)\cos\theta$$

for the functional

$$J_u(\rho) = \int_{\Omega_\rho} u_\rho^2 \, d\Omega.$$

Hence

$$J_u(\rho) = \pi \left(\frac{R^2}{R^2+\rho^2}\right)^2 \left[\rho^4(\ln R - \ln\rho) + \rho^2 R^2 - \frac{5}{4}\rho^4 + \frac{1}{4}R^4\right]$$

and simple calculations show that

$$J_u''(0) = \pi\, R^2.$$

Notice that $\nabla u(0) = [1,0]$, $\nabla w(0) = [-1/4R^2, 0]$, so according to (2.4),

$$J_u''(0) = (-4\pi)(-1/4\, R^2) = \pi R^2.$$

In general, the expression for J_u'' has the form

$$J_u'' = -2\pi \left[x^2 + \frac{1}{2}(3x^2 + y^2 - R^2)\right] = -\pi\,[5x^2 + y^2 - R^2].$$

Hence the level set $J_u'' \leq 0$ is an ellipse with the boundary

$$5x^2 + y^2 - R^2 = 0.$$

Consider the second functional

$$J_g(\rho) = \int_{\Omega_\rho} ||\nabla u_\rho||^2 \, d\Omega$$

in this case with the adjoint state variable $v = 0$, in view of $f = 0$. Thus (2.5) leads to

$$J_g''(0) = -4\pi\, ||\nabla u(0)||^2 = -4\pi.$$

On the other hand,

$$J_g(\rho) = 2\pi \left(\frac{R^2}{R^2+\rho^2}\right)^2 \left[\frac{1}{2}R^2 - \frac{1}{2}\rho^4 R^{-2}\right] \quad \text{thus} \quad J_g''(0) = -4\pi$$

and the inequality $J_g'' < 0$ holds in Ω.

This example is generic, since any regular scalar function resembles in the small neighborhood of the point an inclined plane. Since here $u(x,y) = x = r\cos\theta$, we may write u_ρ as

$$u_\rho = u + \frac{\rho^2}{r}\cos\theta - \frac{\rho^2}{R^2+\rho^2}\left(\frac{\rho^2}{r} + r\right)\cos\theta,$$

which agrees with the expansion (A.3), with $a = 1$ and $b = 0$. The pointwise convergence holds in $\Omega \setminus \{0\}$. In the 1-D case the equivalent problem would be

$$u'' = 0 \quad \text{in} \quad (-1, 1), \qquad u(-1) = -1, \;\; u(1) = 1,$$

so that $u(x) = x$. However, here u_ρ is defined by

$$u''_\rho = 0 \quad \text{in} \quad (-1, -\rho), \qquad u(-1) = -1, \;\; u'(-\rho) = 0,$$

$$u''_\rho = 0 \quad \text{in} \quad (\rho, 1), \qquad u'(\rho) = 0, \;\; u(1) = 1.$$

Therefore,

$$u_\rho(x) = \begin{cases} -1 & \text{for } x \in [-1, -\rho], \\ 1 & \text{for } x \in [\rho, 1], \end{cases}$$

and no convergence in $\Omega \setminus \{0\}$ takes place. This shows that the asymptotic expansion is valid only in higher dimensions.

Example 2. Let us consider the equation

$$\Delta u = -1, \qquad u = 0 \quad \text{on} \quad \partial\Omega,$$

where $\Omega = B_R(0)$. Hence

$$u = \frac{1}{4}(R^2 - r^2),$$

$$u_\rho = u + \frac{1}{2}\rho^2 \,\ln(r/R).$$

Observe that (A.3) holds, since $\nabla u(0) = [0, 0]$. The adjoint state w is given by

$$w = -\frac{1}{32}\,r^4 + \frac{1}{8}\,R^2 r^2 - \frac{3}{32}\,R^4.$$

Hence the gradient of w vanishes at 0 and from (2.4) it follows that

$$J''_u(0) = -2\pi \left[\left(\frac{1}{4}R^2\right)^2 + (-1)\left(-\frac{3}{32}\,R^4\right) \right] = -\frac{5}{16}\,\pi R^4.$$

Explicit computations give the same result.

Again, we may compute the general expression for J''_u. After appropriate transformations,

$$J''_u = -\frac{1}{16}\pi R^4 \left[7\left(\frac{r}{R}\right)^4 - 16\left(\frac{r}{R}\right)^2 + 5 \right];$$

hence the level set $J''_u \leq 0$ is the circle $r \leq 0.6R$.

The gradient of the functional J_g is obtained after some simple calculations. We have $v = 2u$, so that $\nabla v(0) = [0, 0]$ as well. In addition

$$||\nabla u||^2 = \left(\frac{\partial u_\rho}{\partial r}\right)^2 = -\frac{r^2 - \rho^2}{2r};$$

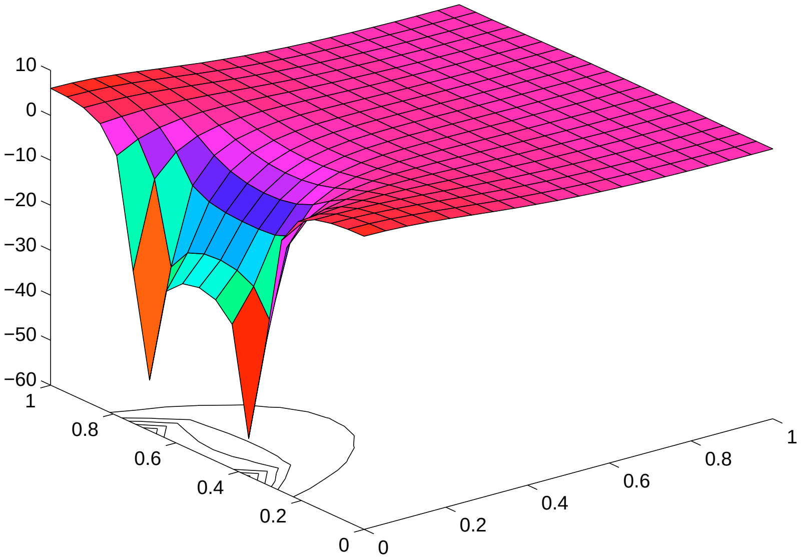


FIG. 3.1. *A graph of J_u'' and its* 0*-level line.*

hence

$$J_g''(0) = \pi R^2.$$

From our formula

$$J_g''(0) = -2\pi \left[(-1) \left(\frac{1}{2} R^2 \right) \right] = \pi R^2.$$

In general,

$$J_g'' = -4\pi \, [r^2 - R^2/4],$$

and the level set of $J_g'' \geq 0$ is the ring $r \geq R/2$.

Example 3. Let us consider the homogeneous Laplace equation $\Delta u = 0$ in $\Omega = [0,1] \times [0,1]$. The boundary conditions are prescribed as follows:

$$\begin{aligned} u = 0 \quad &\text{on} \quad \Gamma_0 = \partial\Omega - \{0\} \times [0,1], \\ u = 1 \quad &\text{on} \quad \Gamma_1 = \{0\} \times [0.3, 0.7], \\ \frac{\partial u}{\partial n} = 0 \quad &\text{on} \quad \Gamma_n = \partial\Omega \setminus (\Gamma_0 \cup \Gamma_1). \end{aligned}$$

The functional $J_u(\rho) = \int_{\Omega_\rho} u_\rho^2$ is defined for $\Omega_\rho = \Omega \setminus B_\rho(x)$; here $x \in \Omega$ stands for the center of the ball $B_\rho(x)$. The distribution of its second derivative as a function of $x \in \Omega$, computed numerically, is shown in Figure 3.1.

4. Plane elasticity problems. Let us consider the elasticity equations in the plane

$$\begin{aligned} A^T DAu = f \quad &\text{in} \quad \Omega, \\ u = g \quad &\text{on} \quad \Gamma_1, \\ B^T DAu = h \quad &\text{on} \quad \Gamma_2, \end{aligned} \tag{4.1}$$

and the same system in the domain with the circular hole $B_\rho(x_0) \subset \Omega$ centered at $x_0 \in \Omega$, $\Omega_\rho = \Omega \setminus \overline{B_\rho(x_0)}$,

$$\begin{aligned} A^T DAu^\rho &= f \quad \text{in} \quad \Omega_\rho, \\ u^\rho &= g \quad \text{on} \quad \Gamma_1, \\ B^T DAu^\rho &= h \quad \text{on} \quad \Gamma_2, \\ B^T DAu^\rho &= 0 \quad \text{on} \quad \Gamma_\rho. \end{aligned} \tag{4.2}$$

Assuming that $0 \in \Omega$, we can consider the case $x_0 = 0$. Here $u = (u_1, u_2)^T$ denotes the displacement field, g is a given displacement field on the fixed part Γ_1 of the boundary, and h is a traction given on the loaded part Γ_2 of the boundary. Finally, the volume forces are denoted by f. In addition, the differential operator is introduced,

$$A = \begin{bmatrix} \frac{\partial}{\partial x_1} & , & 0 \\ 0 & , & \frac{\partial}{\partial x_2} \\ \frac{\partial}{\partial x_1} & , & \frac{\partial}{\partial x_1} \end{bmatrix},$$

and the matrix of material (Lame) coefficients is denoted by

$$D = \begin{bmatrix} \lambda + 2\mu & , & \lambda & , & 0 \\ \lambda & , & \lambda + 2\mu & , & 0 \\ 0 & , & 0 & , & \mu \end{bmatrix}.$$

The following matrix is used for the Neumann boundary conditions:

$$B^T = \begin{bmatrix} n_1 & , & 0 & , & n_2 \\ 0 & , & n_2 & , & n_1 \end{bmatrix},$$

where $n = [n_1, n_2]^T$ is the unit outward normal vector on $\partial\Omega_\rho$. In this notation the stress tensor is replaced by the vector $\sigma = [\sigma_{11}, \sigma_{22}, \sigma_{12}]^T$, strain tensor is given by the vector $\varepsilon = [\varepsilon_{11}, \varepsilon_{22}, \gamma_{12}]^T$ (observe that $\gamma_{12} = 2\varepsilon_{12}$), and the surface tractions are defined by the formulas

$$\varepsilon = A \cdot u, \quad \sigma = D \cdot \varepsilon, \quad t = B \cdot \sigma. \tag{4.3}$$

The principal stresses associated with the displacement field u are denoted by $\sigma_I(u), \sigma_{II}(u)$; the trace of the stress tensor $\sigma(u)$ is denoted by $\text{tr}\sigma(u) = \sigma_I(u) + \sigma_{II}(u)$.

The first shape functional under consideration depends on the displacement field

$$J_u(\rho) = \int_{\Omega_\rho} F(u^\rho)\, d\Omega, \qquad F(u^\rho) = (u^\rho \cdot H \cdot u^\rho)^p = ((u^\rho)^T H u^\rho)^p\ . \tag{4.4}$$

In fact, F may be any C^2 function, similar to the case of the Laplace equation. It is also useful in the framework of elasticity to introduce the yield functional of the form

$$J_\sigma(\rho) = \int_{\Omega_\rho} [\sigma(u^\rho) \cdot S \cdot \sigma(u^\rho)]^p\, d\Omega = \int_{\Omega_\rho} [\sigma(u^\rho)^T S \sigma(u^\rho)]^p\, d\Omega, \tag{4.5}$$

where S is an isotropic matrix. Isotropicity means here that S may be expressed as

$$S = [s_{ij}] = \begin{bmatrix} l + 2m & l & 0 \\ l & l + 2m & 0 \\ 0 & 0 & 4m \end{bmatrix},$$

where l, m are any real constants. Their values for particular yield criteria are given in numerical examples. The following assumptions assure that J_u, J_σ are well defined for solutions of the elasticity system.

(B1) Pure cracks are admissible, even having different types of boundary conditions prescribed on both edges (i.e., tractions and displacements). Then $p = 1$ and g, h must be compatible with $u \in H^1(\Omega; \mathbb{R}^2)$.

(B2) Reentrant corners with $\alpha < 2\pi$ and the same types of boundary conditions are prescribed on both edges of each corner (traction–traction or displacement–displacement). Then $p = 2$ and g, h must be compatible with $u \in W_4^1(\Omega; \mathbb{R}^2)$.

The interior regularity of u in Ω is determined by the regularity of the right-hand side f of the elasticity system. For simplicity the following notation is used for functional spaces:

$$\begin{aligned} H_g^1(\Omega_\rho) &= \{\psi = (\psi_1, \psi_2) \in H^1(\Omega_\rho; \mathbb{R}^2) \mid \psi = g \quad \text{on} \quad \Gamma_1\}, \\ H_{\Gamma_1}^1(\Omega_\rho) &= \{\psi = (\psi_1, \psi_2) \in H^1(\Omega_\rho; \mathbb{R}^2) \mid \psi = 0 \quad \text{on} \quad \Gamma_1\}, \\ H_{\Gamma_1}^1(\Omega) &= \{\psi = (\psi_1, \psi_2) \in H^1(\Omega; \mathbb{R}^2) \mid \psi = 0 \quad \text{on} \quad \Gamma_1\}; \end{aligned}$$

here we use the convention that, e.g., $H_g^1(\Omega_\rho)$ stands for the Sobolev space of vector functions $H_g^1(\Omega_\rho; \mathbb{R}^2)$.

The weak solutions to the elasticity systems are defined as follows.

Find $u^\rho \in H_g^1(\Omega_\rho)$ such that, for every $\phi \in H_{\Gamma_1}^1(\Omega)$,

$$-\int_{\Omega_\rho} (Au^\rho)^T DA\phi \, d\Omega + \int_{\Gamma_2} h^T \phi \, dS = \int_{\Omega_\rho} f^T \phi \, d\Omega. \tag{4.6}$$

The adjoint state equations for the functional J_u are introduced.

Find $w^\rho \in H_{\Gamma_1}^1(\Omega_\rho)$ such that, for every $\phi \in H_{\Gamma_1}^1(\Omega)$,

$$-\int_{\Omega_\rho} (Aw^\rho)^T DA\phi \, d\Omega = \int_{\Omega_\rho} F_u'(u^\rho)^T \phi \, d\Omega. \tag{4.7}$$

Finally, $v^\rho \in H_{\Gamma_1}^1(\Omega_\rho)$ is the adjoint state for J_σ and satisfies for all test functions $\phi \in H_{\Gamma_1}^1(\Omega)$ the following integral identity:

$$-\int_{\Omega_\rho} (Av^\rho)^T DA\phi \, d\Omega = 2p \int_{\Omega_\rho} [\sigma(u^\rho)^T S \sigma(u^\rho)]^{(p-1)} \sigma(u^\rho)^T SDA\phi \, d\Omega. \tag{4.8}$$

Now we may formulate the following result.

THEOREM 4.1. *Assume that the distributed force is sufficiently regular*, $f \in C^1 (\Omega; \mathbb{R}^2)$, *and* (B1) *or* (B2) *is satisfied; then*

$$J_u''(0) = -2\pi \left[F(u) + f^T w + \frac{1}{E}(a_u a_w + 2 b_u b_w \cos 2\delta) \right]_{x=x_0}, \tag{4.9}$$

$$J_\sigma''(0) = -2\pi \left[s_{22}^p K_p(a_u, b_u) + f^T v + \frac{1}{E}(a_u a_v + 2 b_u b_v \cos 2\delta) \right]_{x=x_0}. \tag{4.10}$$

Some of the terms in (4.9), (4.10) require explanation. The function K_p takes the values

$$K_p(a, b) = \begin{cases} a^2 + 2b^2 & \text{for } p = 1, \\ a^4 + 6b^4 + 12a^2b^2 & \text{for } p = 2. \end{cases}$$

We denote

$$\begin{aligned} a_u &= \mathrm{tr}\sigma(u), & b_u &= \sigma_I(u) - \sigma_{II}(u),\\ a_w &= \mathrm{tr}\sigma(w), & b_w &= \sigma_I(w) - \sigma_{II}(w),\\ a_v &= \mathrm{tr}\sigma(v), & b_v &= \sigma_I(v) - \sigma_{II}(v). \end{aligned}$$

Finally, the angle δ denotes the angle between principal stress directions for displacement fields u and w in (4.9) and for displacement fields u and v in (4.10).

Proof. Let us calculate the derivatives of the functional $J_u(\rho)$ with respect to the parameter ρ, which determines the size of the hole B_ρ, by using the material derivative method. This leads to

$$J'_u(\rho) = \int_{\Omega_\rho} F'_u(u^\rho)^T u^{\rho\prime}\, d\Omega - \int_{\Gamma_\rho} F(u^\rho)\, dS, \tag{4.11}$$

and in the same way for the state equation

$$-\int_{\Omega_\rho} (Au^{\rho\prime})^T DA\phi\, d\Omega + \int_{\Gamma_\rho} (Au^\rho)^T DA\phi\, dS = -\int_{\Gamma_\rho} f^T\phi\, dS, \tag{4.12}$$

where $u^{\rho\prime}$ is the shape derivative.

After substitution of the test functions $\phi = w^\rho$ in the state equation, $\phi = u^{\rho\prime}$ in the adjoint state equation, we get

$$\begin{aligned} J'_u(\rho) &= -\int_{\Gamma_\rho} [(Au^\rho)^T DAw^\rho + F(u^\rho) + f^T w^\rho]\, dS \\ &= -\int_{\Gamma_\rho} \left[\frac{1}{E}\sigma_{\theta\theta}(w^\rho)\sigma_{\theta\theta}(u^\rho) + F(u^\rho) + f^T w^\rho\right] dS, \end{aligned} \tag{4.13}$$

since for the displacement fields u^ρ, w^ρ the boundary conditions $\sigma_{\theta\theta} = \tau_{r\theta} = 0$ on Γ_ρ are prescribed; here $\sigma_{\theta\theta}, \tau_{r\theta}$ denote the components of the stress tensor in the reference frame tied with normal and tangent unit vectors on Γ_ρ.

It is obvious that

$$J'_u(\rho) \underset{\rho\to 0+}{\longrightarrow} 0;$$

therefore, we compute the second derivative

$$J''_u(\rho) = I_1(\rho) - I_2(\rho) - I_3(\rho), \tag{4.14}$$

where

$$\begin{aligned} I_1(\rho) &= \int_{\Gamma_\rho} \frac{\partial}{\partial n}\left[\frac{1}{E}\sigma_{\theta\theta}(w^\rho)\sigma_{\theta\theta}(u^\rho) + F(u^\rho) + f^T w^\rho\right] dS,\\ I_2(\rho) &= \int_{\Gamma_\rho} \left[\frac{1}{E}(\sigma_{\theta\theta}(w^\rho)\sigma_{\theta\theta}(u^\rho))' + F'_u(u^\rho)u^{\rho\prime} + f^T w^{\rho\prime}\right] dS,\\ I_3(\rho) &= \frac{1}{\rho}\int_{\Gamma_\rho} \left[\frac{1}{E}\sigma_{\theta\theta}(w^\rho)\sigma_{\theta\theta}(u^\rho) + F(u^\rho) + f^T w^\rho\right] dS. \end{aligned}$$

Taking into account that $\frac{\partial}{\partial n} = -\frac{\partial}{\partial r}$ on Γ_ρ and using the expansion (B.6), we obtain

$$\frac{\partial}{\partial n}\sigma_{\theta\theta}(u^\rho) = \left[a_u\frac{\rho^2}{r^3} - 6b_u\frac{\rho^4}{r^5}\cos 2\theta\right] + O(\rho^{-\epsilon}) \underset{r=\rho}{=} a_u\frac{1}{\rho} - 6b_u\frac{1}{\rho}\cos\theta + O(\rho^{-\epsilon}).$$

Similarly,

$$\frac{\partial}{\partial \rho}\sigma_{\theta\theta}(u^\rho) = \left[a_u \frac{\rho}{r^2} - 6b_u \frac{\rho^3}{r^4}\cos 2\theta\right] + O(\rho^{-\epsilon}) \underset{r=\rho}{=} a_u \frac{1}{\rho} - 6b_u \frac{1}{\rho}\cos\theta + O(\rho^{-\epsilon}).$$

This means that the singular terms cancel out

$$\frac{\partial}{\partial n}\sigma_{\theta\theta}(u^\rho) - \frac{\partial}{\partial \rho}\sigma_{\theta\theta}(u^\rho) = O(\rho^{-\epsilon})$$

and

$$I_1(\rho) - I_2(\rho) \underset{\rho\to 0+}{\longrightarrow} 0.$$

Now we express the asymptotic expansion for $\sigma_{\theta\theta}(w^\rho)$ in the reference frame tied with principal stress directions for the displacement field u_ρ, and *not* with its own field w_ρ:

$$\sigma_{\theta\theta}(w^\rho) = \frac{1}{2}a_w\left(1 + \frac{\rho^2}{r^2}\right) - \frac{1}{2}b_w\left(1 + 3\frac{\rho^4}{r^4}\right)\cos 2(\theta - \delta) + O(\rho^{1-\epsilon}). \tag{4.15}$$

This leads to

$$\begin{aligned}
&\lim_{\rho\to 0+}\int_0^{2\pi} \sigma_{\theta\theta}(u^\rho)\sigma_{\theta\theta}(w^\rho)\, d\theta \\
&\quad = \int_0^{2\pi} [a_u - 2b_u\cos 2\theta][a_w - 2b_w\cos 2(\theta - \delta)]\, d\theta \\
&\quad = 2\pi[a_u a_w + 2b_u b_w \cos 2\delta],
\end{aligned} \tag{4.16}$$

and the final expression for the second derivative of J_u results.

In the case of J_σ the integral terms become

$$I_1(\rho) = \int_{\Gamma_\rho} \frac{\partial}{\partial n}\left[\frac{1}{E}\sigma_{\theta\theta}(v^\rho)\sigma_{\theta\theta}(u^\rho) + (s_{22}\sigma_{\theta\theta}(u^\rho)^2)^p + f^T v^\rho\right] dS,$$

$$I_2(\rho) = \int_{\Gamma_\rho} \left[\frac{1}{E}(\sigma_{\theta\theta}(v^\rho)\sigma_{\theta\theta}(u^\rho))' + +2ps_{22}^p\sigma_{\theta\theta}(u^\rho)^{2p-1}\sigma_{\theta\theta}(u^\rho)' + f^T v^{\rho\prime}\right] dS,$$

$$I_3(\rho) = \frac{1}{\rho}\int_{\Gamma_\rho} \left[\frac{1}{E}\sigma_{\theta\theta}(v^\rho)\sigma_{\theta\theta}(u^\rho) + (s_{22}\sigma_{\theta\theta}(u^\rho)^2)^p + f^T v^\rho\right] dS.$$

Again, in the same way as before,

$$I_1(\rho) - I_2(\rho) \underset{\rho\to 0+}{\longrightarrow} 0.$$

The function K_p is defined by the expression

$$K_p(a, b) = \frac{1}{2\pi}\int_0^{2\pi} (a - 2b\cos 2\theta)^{2p}\, d\theta.$$

The proof of Theorem 4.1 is completed. □

Remark 4. The matrix in the definition of J_σ, in fact, may be arbitrary, similar to the case of the scalar equation, and not only isotropic. However, it is difficult to imagine such a need for the isotropic material. Anyway, in the general case, we would have to transform S according to the known rules determined by the rotation of the reference frame. Then in the definition of $I_3(\rho)$ instead of s_{22} we would have an expression containing all the elements of S and trigonometric functions of θ. The integration is again possible but leads to more complicated formulas.

5. Examples for plane elasticity.

Example 4. Let us take the square domain, with the side length a. It is fixed on small segments of the length $a/10$ at the lower and upper part of the left side. The elastic body is pulled by the downward force distributed over the segment of the length $a/3$ located in the middle of the right side. The initial and distorted configurations are shown in Figure 5.1. The material Lame coefficients satisfy relation $\lambda = \mu$. We consider the functional J_u with $p = 8$ (approximating maximal displacement) and the following three types of J_σ, corresponding to the following yield criteria.

1. The elastic energy yield criterion (rarely used), which is equivalent modulo a proportionality factor (assuming $\lambda = \mu$ for Lame coefficients) to the following relation:

$$\sigma_{red}^2 = 3\sigma_{11}^2 + 3\sigma_{22}^2 - 2\sigma_{11}\sigma_{22} + 8\sigma_{12}^2. \tag{5.1}$$

This in turn corresponds to the isotropic matrix S with $l = -1$, $m = 2$, and J_σ given by (4.5).

2. The Huber yield criterion (frequently used), which is equivalent modulo a proportionality factor to the following relation:

$$\sigma_{red}^2 = 2\sigma_{11}^2 + 2\sigma_{22}^2 - 2\sigma_{11}\sigma_{22} + 6\sigma_{12}^2. \tag{5.2}$$

This in turn corresponds to the isotropic matrix S with $l = -1$, $m = 3/2$, and J_σ given by (4.5).

3. The maximal shear stress yield criterion (often used), which is equivalent modulo proportionality factor to the following relation:

$$\sigma_{red}^2 = \sigma_{11}^2 + \sigma_{22}^2 - 2\sigma_{11}\sigma_{22} + 4\sigma_{12}^2. \tag{5.3}$$

This in turn corresponds to the isotropic matrix S with $l = -1$, $m = 1$, and J_σ given by (4.5).

The second derivatives of these functionals are shown in Figures 5.1–5.4. The displacement functional here has the form $J_u = \int_\Omega (u_1^2 + u_2^2)^4 \, d\Omega$. The energy yield criterion is similar to the compliance functional considered in [8]. The level lines are distributed uniformly across the range of functions, and the lighter shades denote smaller values. The distributions of integrand functions and the densities of the second derivatives of functionals look similar; they are not, however, proportional to each other. The regions where the values of J'' are the smallest constitute the possible locations of holes. The same comments apply to the next example.

Example 5. Let us take the elongated rectangle, fixed on both left and right sides and loaded by the downward force over the small segment in the middle of the upper side. Its initial and distorted configuration are shown in Figure 5.5. Again we consider the same yield functions under assumption $\lambda = \mu$. The numerical results are shown in Figures 5.5–5.8.

Appendix A. Asymptotic expansions for Laplace equation in $\mathbb{R}^2$. Let us consider the equation

$$\begin{aligned} \Delta u &= f \quad \text{in} \quad \Omega, \\ u &= g \quad \text{on} \quad \Gamma_1, \\ \frac{\partial u}{\partial n} &= h \quad \text{on} \quad \Gamma_2, \end{aligned} \tag{A.1}$$

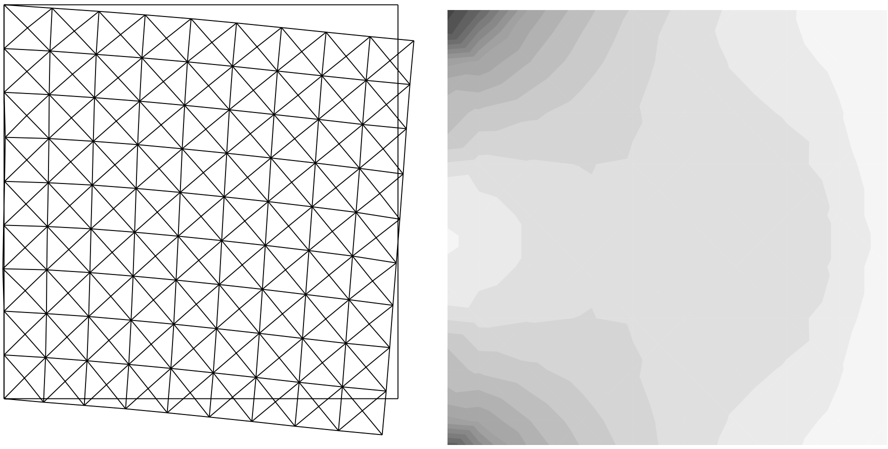


FIG. 5.1. *The square (original and distorted) and the distribution of the J''_u density.*

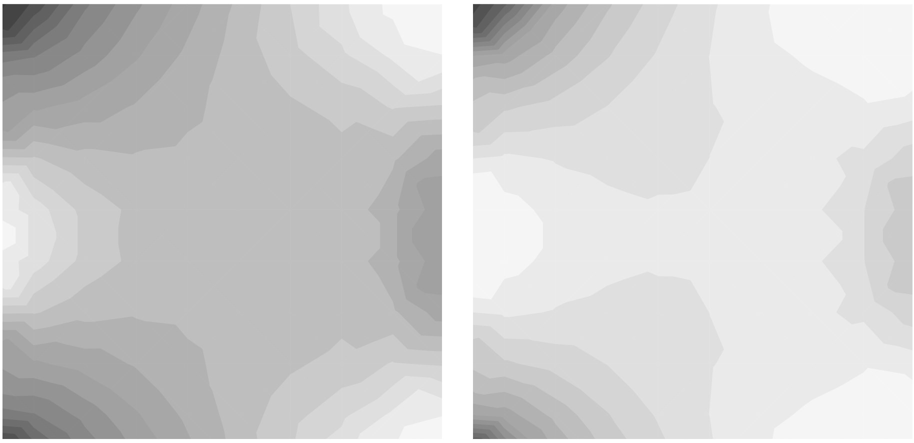


FIG. 5.2. *The distribution of the elastic energy function* (5.1) *and the corresponding J''_σ density.*

with $f \in C^1(\Omega)$ thus the solution $u \in C^3(\Omega)$. We drill a hole at $x_0 \in \Omega$, denoted $B_\rho(x_0),\ \rho < d(x_0, \Gamma)$, and define the set $\Omega_\rho = \Omega \setminus \overline{B_\rho(x_0)}$. Now consider

$$\begin{aligned} \Delta u_\rho &= f \quad \text{in} \quad \Omega_\rho, \\ u_\rho &= g \quad \text{on} \quad \Gamma_1, \\ \frac{\partial u_\rho}{\partial n} &= h \quad \text{on} \quad \Gamma_2, \\ \frac{\partial u_\rho}{\partial n} &= 0 \quad \text{on} \quad \Gamma_\rho = \partial B_\rho(x_0). \end{aligned} \tag{A.2}$$

Assume for simplicity that $x_0 = 0$. Then we have the following asymptotic expansion relations. Denote

$$\nabla u|_{x=0} = [a, b]^T.$$

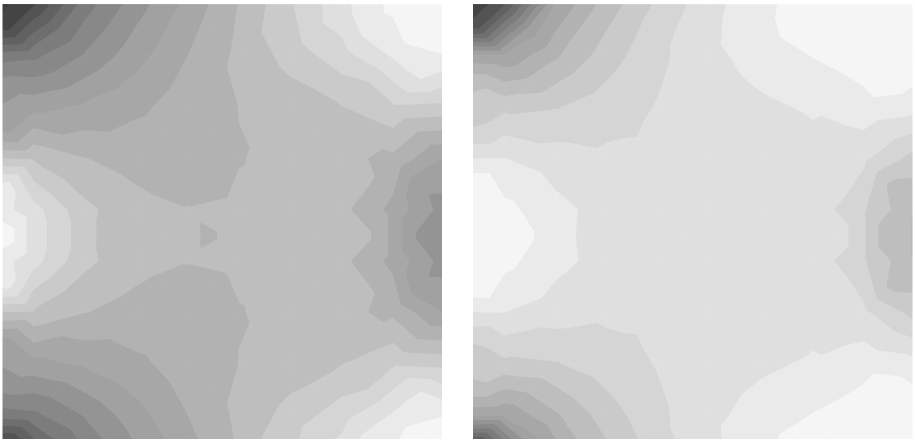


FIG. 5.3. *The distribution of the Huber yield function* (5.2) *and the corresponding* J''_σ *density.*

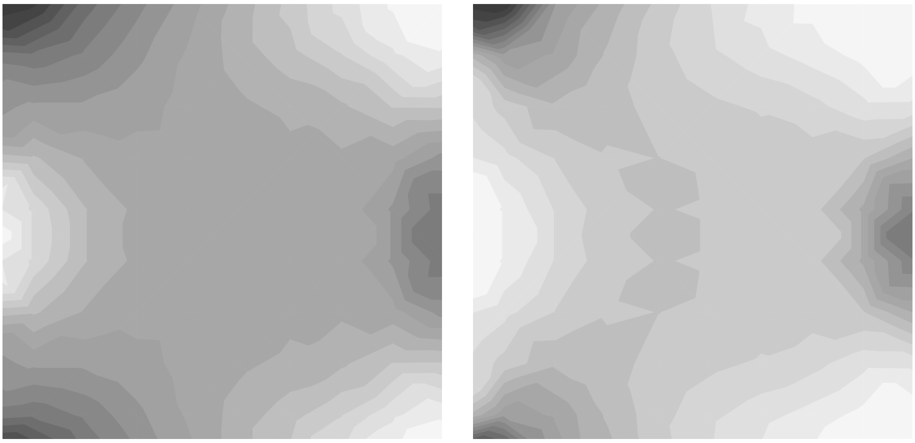


FIG. 5.4. *The distribution of the maximal shear stress yield function* (5.3) *and the corresponding* J''_σ *density.*

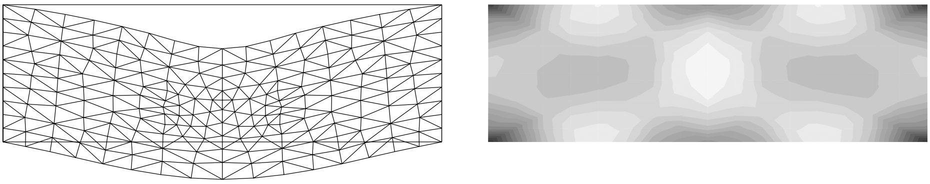


FIG. 5.5. *The object (original and distorted) and the distribution of the* J''_u *density.*

The solution u as a function of r, θ, can be expressed for $r \geq \rho$ as follows (see [5], Satz 4, and [3], [6]):

$$u_\rho = u + a\frac{\rho^2}{r}\cos\theta + b\frac{\rho^2}{r}\sin\theta + \mathcal{R}, \tag{A.3}$$

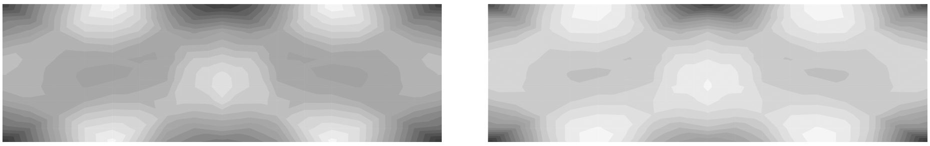


FIG. 5.6. *The distribution of the elastic energy function* (5.1) *and the corresponding* J''_σ *density.*

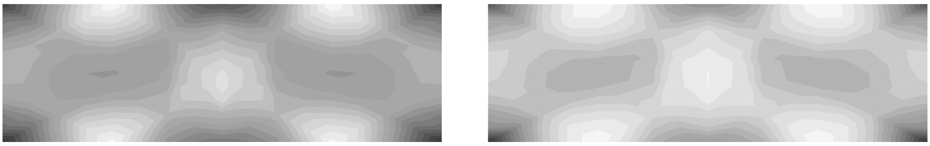


FIG. 5.7. *The distribution of the Huber yield function* (5.2) *and the corresponding* J''_σ *density.*

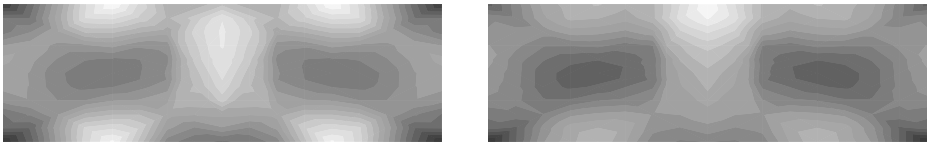


FIG. 5.8. *The distribution of the maximal shear stress yield function* (5.3) *and the corresponding* J''_σ *density.*

where

$$\mathcal{R} = \rho^2 \left[O\left(\frac{\rho}{r}\right) + l(\rho, r) \right],$$

and $l(\rho, r)$ may contain finite powers of $\ln\rho, \ln r$. Hence $\mathcal{R} = O(\rho^{2-\epsilon})$ for any $\epsilon > 0$.

The above formula gives the asymptotic expansion in the function space to which u belongs, the solution to (A.1). Besides, for smooth $f \in C^1(\Omega)$, u is three times continuously differentiable in an open neighborhood of B_ρ.

Therefore, in the ring $\rho \leq r \leq 2\rho$, taking into account the regularity of u in the neighborhood of $x_0 = 0$, we have the expansion

$$u_\rho = u(0) + a\left(\frac{\rho^2}{r} + r\right)\cos\theta + b\left(\frac{\rho^2}{r} + r\right)\sin\theta + O(\rho^{2-\epsilon}) \ , \tag{A.4}$$

where $u(0)$ denotes the value at x_0 of the solution to (A.1).

The above formulas are given in the polar coordinate system with the center at $x_0 = 0$, which coincides with the center of the ball B_ρ. In particular, from (A.4) it follows that

$$\left.\frac{\partial u_\rho}{\partial \tau}\right|_{r=\rho} = \left.\frac{1}{\rho}\frac{\partial u_\rho}{\partial \theta}\right|_{r=\rho} = 2(-a\sin\theta + b\cos\theta) + O(\rho^{1-\epsilon}). \tag{A.5}$$

Appendix B. Asymptotic expansions for the elasticity system in $\mathbb{R}^2$. Let us consider the systems (4.1) and (4.2) and assume that the coordinate system is aligned with the principal stress directions, so that $\sigma_{12} = 0$. Denote also

$$a_u = [\sigma_I(u) + \sigma_{II}(u)]|_{x=0}, \tag{B.1}$$

$$b_u = [\sigma_I(u) - \sigma_{II}(u)]|_{x=0}. \tag{B.2}$$

Let us introduce the polar coordinate system (r, θ). At each point in the plane we define also the orthogonal coordinate axes, still denoted by (r, θ), and defined by the unit vectors e_r, e_θ, directed along r and perpendicular to it, counterclockwise. Given the displacement field u, we may compute the components of the strain field (in the orthogonal system and using the polar coordinates):

$$\begin{aligned} \varepsilon_{rr} &= \frac{\partial u_r}{\partial r}, \\ \varepsilon_{\theta\theta} &= \frac{u_r}{r} + \frac{1}{r}\frac{\partial u_\theta}{\partial \theta}, \\ \gamma_{r\theta} &= \frac{1}{r}\frac{\partial u_r}{\partial \theta} + \frac{\partial u_\theta}{\partial r} - \frac{u_\theta}{r}. \end{aligned} \tag{B.3}$$

The corresponding isotropic Hook law has the form

$$\begin{aligned} \varepsilon_{rr} &= \frac{1}{E}(\sigma_{rr} - \nu\sigma_{\theta\theta}), \\ \varepsilon_{\theta\theta} &= \frac{1}{E}(\sigma_{\theta\theta} - \nu\sigma_{rr}), \\ \gamma_{r\theta} &= \frac{1}{G}\tau_{r\theta}, \end{aligned} \tag{B.4}$$

where $G = E/2(1+\nu)$. Then, similar to the Laplace case, the following expansion holds (see, e.g., [4] and [6]) in the ring $\rho \le r \le 2\rho$:

$$\begin{aligned} u_r^\rho = u_r(0) &+ \frac{a_u}{8Gr}[(\kappa-1)r^2 + 2\rho^2] \\ &+ \frac{b_u}{4Gr}\left[(\kappa+1)\rho^2 + r^2 - \frac{\rho^4}{r^2}\right]\cos 2\theta + O(\rho^{2-\epsilon}), \\ u_\theta^\rho = u_\theta(0) &- \frac{b_u}{4Gr}\left[(\kappa-1)\rho^2 + r^2 + \frac{\rho^4}{r^2}\right]\sin 2\theta + O(\rho^{2-\epsilon}), \end{aligned} \tag{B.5}$$

where $\kappa = (3-\nu)/(1+\nu)$ for plane stress and

$$\begin{aligned} u_r(0) &= \lim_{r\to 0} u_r(r,\theta), \\ u_\theta(0) &= \lim_{r\to 0} u_\theta(r,\theta). \end{aligned}$$

The corresponding expressions for the stresses have the form

$$\begin{aligned} \sigma_{rr}(u^\rho) &= \frac{1}{2}\left[a_u\left(1 - \frac{\rho^2}{r^2}\right) + b_u\left(1 - 4\frac{\rho^2}{r^2} + 3\frac{\rho^4}{r^4}\right)\cos 2\theta\right] + O(\rho^{1-\epsilon}), \\ \sigma_{\theta\theta}(u^\rho) &= \frac{1}{2}\left[a_u\left(1 + \frac{\rho^2}{r^2}\right) - b_u\left(1 + 3\frac{\rho^4}{r^4}\right)\cos 2\theta\right] + O(\rho^{1-\epsilon}), \\ \tau_{r\theta}(u^\rho) &= -\frac{1}{2}b_u\left(1 + 2\frac{\rho^2}{r^2} - 3\frac{\rho^4}{r^4}\right)\sin 2\theta + O(\rho^{1-\epsilon}). \end{aligned} \tag{B.6}$$

Observe that due to the free edge condition on the boundary of the hole, we have

$$\sigma_{rr}(u^\rho) = \tau_{r\theta}(u^\rho) = 0 \quad \text{on} \quad \partial B_\rho.$$

Appendix C. Derivatives of general functionals. Denote $(\cdot)'_\rho = \partial(\cdot)/\partial\rho$. Then we know [9] that for general G,

$$\left[\int_{\Gamma_\rho} G(u_\rho)\, dS\right]'_\rho = \int_{\Gamma_\rho} \left[G'_u(u_\rho)u'_\rho - \frac{\partial G(u_\rho)}{\partial n}\right] dS + \frac{1}{\rho}\int_{\Gamma_\rho} G(u_\rho)\, dS, \tag{C.1}$$

$$\left[\int_{\Omega_\rho} G(u_\rho)\, d\Omega\right]'_\rho = \int_{\Omega_\rho} G'_u(u_\rho)u'_\rho\, d\Omega - \int_{\Gamma_\rho} G(u_\rho)\, dS. \tag{C.2}$$

The formulas (C.1), (C.2) follow from the general formulas for the shape derivatives of integral functionals; we refer the reader to [9] for the details.

Appendix D. Asymptotic expansions for the adjoint function. The adjoint variable w_ρ satisfies the boundary value problem

$$\begin{aligned} \Delta w_\rho &= F'(u_\rho) \quad \text{in} \quad \Omega_\rho,\\ w_\rho &= 0 \quad \text{on} \quad \Gamma_1,\\ \frac{\partial u_\rho}{\partial n} &= 0 \quad \text{on} \quad \Gamma_2 \cup \Gamma_\rho. \end{aligned} \tag{D.1}$$

We know also that

$$u_\rho = u + \psi_\rho(r)\cdot(a\cos\theta + b\sin\theta) + \mathcal{R},$$

where $\psi_\rho(r) = \rho^2/r$ for $r \geq \rho$. Let us neglect for a moment $\mathcal{R}$, which is proportional to a higher power of ρ. Then, due to the regularity of F,

$$F'(u_\rho) = F'_u(u) + F''(u)\psi_\rho(r)\cdot(a\cos\theta + b\sin\theta) + O(\psi_\rho^2).$$

Again, we neglect the last term, quadratic with respect to ρ. The function $\psi_\rho(r)\cdot(a\cos\theta + b\sin\theta)$ may be smoothly extended to $\bar\psi_\rho$, defined on the whole Ω by putting

$$\bar\psi_\rho = \begin{cases} \frac{1}{2}(3\rho - r^2/\rho)\cdot(a\cos\theta + b\sin\theta) & \text{if } r < \rho,\\ \psi_\rho(r)\cdot(a\cos\theta + b\sin\theta) & \text{if } r \geq \rho. \end{cases}$$

In the next step we consider the problems

$$\begin{aligned} &\Delta p = F'_u(u) \quad \text{in} \quad \Omega, \qquad p = 0 \quad \text{on} \quad \Gamma_1, \quad \frac{\partial p}{\partial n} = 0 \quad \text{on} \quad \Gamma_2,\\ &\Delta q = \bar\psi_\rho\cdot F''_u(u) \quad \text{in} \quad \Omega, \quad q = 0 \quad \text{on} \quad \Gamma_1, \quad \frac{\partial q}{\partial n} = 0 \quad \text{on} \quad \Gamma_2. \end{aligned}$$

The function p is equal to the adjoint variable in the domain without a hole and, therefore, has the expansion

$$p_\rho = p + \psi_\rho(r)\cdot(c\cos\theta + d\sin\theta) + \mathcal{R},$$

where $[c, d] = \nabla p(0)$. There remains to analyze the expansion of q. We shall use here the results of [2, Chapter 2], and [7, Chapter 2].

Let $2\delta < \operatorname{dist}(0, \Gamma) - \rho$ and Ω_δ correspond to the domain Ω with the balls of radii δ centered around the boundary corner points removed. Then, due to the regularity of $\bar{\psi}_\rho$,

$$\begin{aligned} \|q\|_{H^2(\Omega_{2\delta})} &\leq C_\delta \|\bar{\psi}_\rho\|_{L^2(\Omega_{2\delta})}, \\ \|q\|_{H^3(\Omega_{2\delta})} &\leq C_\delta \|\bar{\psi}_\rho\|_{H^1(\Omega_{2\delta})}, \end{aligned}$$

and therefore both q and ∇q are pointwise bounded in $\Omega_{2\delta}$. Moreover, it may be checked by direct computations that

$$\begin{aligned} &\|\bar{\psi}_\rho\|_{L^2(\Omega_{2\delta})} \leq \rho^2(\Lambda_1 + \Lambda_2 |\ln \rho|) = O(\rho^{2-\epsilon}), \\ &\|\bar{\psi}_\rho\|_{H^1(\Omega_{2\delta})} \leq \Lambda\rho. \end{aligned}$$

Hence

$$q(0) = O(\rho^{2-\epsilon}), \qquad |\nabla q(0)| = O(\rho). \tag{D.2}$$

In the last step we consider the expansion of q with respect to the hole $B_{\rho_1}(0)$. It has the form

$$q_{\rho_1} = q(0) + \left(1 + \frac{\rho_1^2}{r^2}\right)(\nabla q(0) \cdot x) + O(\rho_1^{2-\epsilon}),$$

where $x = [r \cos\theta, r \sin\theta]$. By putting $\rho_1 = \rho$ and taking into account (D.2) we get

$$q_\rho = O(\rho^{2-\epsilon}),$$

and therefore the leading part of the expansion of w_ρ coincides with the expansion of p_ρ, i.e., the adjoint variable with a fixed right-hand side. The neglected parts contribute only to the higher order terms.

REFERENCES

[1] M. Giaquinta, *Introduction to Regularity Theory for Nonlinear Elliptic Systems*, Birkhäuser-Verlag, Basel, Switzerland, 1993.

[2] P. Grisvard, *Singularities in Boundary Value Problems*, Springer-Verlag, Berlin, New York, 1992.

[3] D. Göhde, *Singuläre Störung von Randvertproblemen durch ein kleines Loch im Gebiet*, Z. Anal. Anwendungen, 4 (1985), pp. 467–477.

[4] H. G. Hahn, *Elastizitätstheorie*, Teubner, Stuttgart, 1985.

[5] A. Herwig, *Elliptische randvertprobleme zweiter Ordnung in Gebieten mit einer Fehlstelle*, Z. Anal. Anwendungen, 8 (1989), pp. 153–161.

[6] A. M. Il'in, *Matching of Asymptotic Expansions of Solutions of Boundary Value Problems*, Transl. Math. Monogr., 102, AMS, Providence, 1992.

[7] S. A. Nazarov and B. A. Plamenevsky, *Elliptic Problems in Domains with Piecewise Smooth Boundaries*, de Gruyter Exposition in Mathematics 13, Walter de Gruyter, Berlin, New York, 1994.

[8] A. Shumacher, *Topologieoptimierung von Bauteilstrukturen unter Verwendung von Lochpositionierungkriterien*, Ph.D. thesis, Universität–Gesamthochschule–Siegen, Siegen, 1995.

[9] J. Sokołowski and J-P. Zolesio, *Introduction to Shape Optimization. Shape Sensitivity Analysis*, Springer-Verlag, Berlin, New York, 1992.

[10] J. Sokołowski and A. Żochowski, *On topological derivative in shape optimisation*, INRIA-Lorraine, Rapport de Recherche No. 3170, 1997.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 4, pp. 1273–1290

PONTRYAGIN MAXIMUM PRINCIPLE FOR OPTIMAL CONTROL OF VARIATIONAL INEQUALITIES*

MAÏTINE BERGOUNIOUX[†] AND HOUSNAA ZIDANI[‡]

Abstract. In this paper we investigate optimal control problems governed by variational inequalities. We present a method for deriving optimality conditions in the form of Pontryagin's principle. The main tools used are the Ekeland's variational principle combined with penalization and spike variation techniques.

Key words. variational inequalities, optimal control, Pontryagin principle

AMS subject classifications. 49J20, 49M29

PII. S0363012997328087

1. Introduction.

The purpose of this paper is to present a method for deriving a Pontryagin-type maximum principle as a first-order necessary condition of optimal controls for problems governed by variational inequalities. We allow various kinds of constraints to be imposed on the state. To be more precise, we consider the following variational inequality:

$$\frac{\partial y}{\partial t} + Ay + f(y) + \partial\varphi(y) \ni u \quad \text{in } Q = \Omega \times]0, T[, \tag{1.1a}$$

$$y = 0 \qquad \text{on } \Sigma = \Gamma \times]0, T[, \tag{1.1b}$$

$$y(0) = y_o \quad \text{in } \Omega, \tag{1.1c}$$

where $\Omega \subset \mathbb{R}^n$, $T > 0$, u is a distributed control, A is a second-order elliptic operator, and $\frac{\partial y}{\partial t}$ denotes the derivative of y with respect to t; $\partial\varphi(y)$ is the subdifferential of the function φ at y. We shall give all the definitions we need in section 3 and (1.1) will be made clear as well. The control variable u and the state variable y must satisfy constraints of the form

(1.2a)
$$u \in U_{ad} = \{\ u \in L^p(Q) \mid u(x,t) \in \mathcal{K}_U(x,t) \text{ almost everywhere (a.e.) in } Q\ \} \subset L^p(Q),$$

where $\mathcal{K}_U$ is a measurable set-valued mapping from Q with closed values in $\mathcal{P}(\mathbb{R})$ ($\mathcal{P}(\mathbb{R})$ being the set of all subsets of $\mathbb{R}$), and where

$$\Phi(y) \in \mathcal{C} \tag{1.2b}$$

with $1 < p < \infty$, Φ is a $\mathcal{C}^1$ mapping from $C(\overline{Q})$ into $C(\overline{Q})$, and $\mathcal{C} \subset C(\overline{Q})$ is a closed convex subset with finite codimension.

*Received by the editors October 1, 1997; accepted for publication (in revised form) July 6, 1998; published electronically May 28, 1999.
http://www.siam.org/journals/sicon/37-4/32808.html

[†]Département de Mathématiques, UMR 6628, Université d'Orléans, B.P. 6759, 45067 Orléans Cedex 2, France (maitine@labomath.univ-orleans.fr).

[‡]Laboratoire MIP, UMR 5640, UFR MIG, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse, France (zidani@mip.ups.tlse.fr).

The control problem is

$$(\mathcal{P}) \qquad \inf\{J(y,u) \mid y \in C(\overline{Q}), u \in U_{ad}, (y,u) \text{ satisfies } (1.1),(1.2)\},$$

where the cost functional is defined by

$$(1.3) \qquad J(y,u) = \int_Q F(x,t,y(x,t),u(x,t))\,dx\,dt + \int_\Omega L(x,y(x,T))dx.$$

Many authors (for example, Barbu [2], Mignot–Puel [17], Yong [23], Bonnans–Tiba [6], Bonnans–Casas [5], and Bergounioux [3]) have already considered control problems for variational inequalities from the theoretical or numerical point of view. Here we are interested in optimality conditions in the form of Pontryagin's principle. The existence of an optimal solution is assumed a priori. The novelty of this paper is twofold: We obtain the optimality conditions in Pontryagin's form and we think that our hypotheses seem to be minimal. In essence we ask for the state equation to be well posed and assume differentiability of data with respect to the state. We allow various kinds of constraints to be added on the control u and on the state. However, we restrict the study to the case in which φ is the indicator function of the closed convex set $K_o = \{z \in C(\overline{Q}) \mid z \geq 0\}$ so that the variational inequality (1.1) becomes the so-called obstacle problem.

To get Pontryagin's principle, we use a method based on penalization of state constraints and on Ekeland's principle combined with diffuse perturbations [16, 20]. These techniques already have been used by many authors in the case of optimal control of parabolic or elliptic equations [5, 16, 21]. Some of these techniques also have been used for control problems governed by variational inequalities [5, 23, 4]. In those papers, the variational inequality is approximated via the Moreau–Yosida approximation of the maximal monotone operator $\partial\varphi$.

Here we use another idea based on the formulation of (1.1) with a slackness variable and the regularity of its solution. In fact, the solution of (1.1) is also a weak solution of

$$(1.4) \qquad \frac{\partial y}{\partial t} + Ay + f(y) = u + \xi \quad \text{in } Q, \qquad y = 0 \quad \text{on } \Sigma, \qquad y(0) = y_o \quad \text{in } \Omega,$$

where ξ is the Lagrange multiplier associated with the variational inequality and is introduced as an additional control variable. Therefore we obtain a problem $(\widetilde{\mathcal{P}})$ equivalent to $(\mathcal{P})$, with constraints on both the control variable and the state variable as well as coupled state-control constraints. We first give a Pontryagin's principle for $(\widetilde{\mathcal{P}})$. For this, we adapt the proof given in [21, 24, 7] to problem $(\widetilde{\mathcal{P}})$. Next we derive optimality conditions for $(\mathcal{P})$ from those for $(\widetilde{\mathcal{P}})$.

2. Assumptions.

Let Ω be an open, smooth (with a $\mathcal{C}^2$ boundary Γ for example), and bounded domain of $\mathbb{R}^n$ $(2 \leq n)$. In this paper we suppose that

$$p > n.$$

Remark 2.1. We must emphasize that this choice of p is not optimal. Indeed, we should distinguish the integers p (for the L^p-space of the distributed control u) and q (for the L^q-space of the initial value y_o). The optimal choice should be $u \in L^p(Q)$ with $p > \frac{n}{2} + 1$ and $y_o \in W_o^{1,q}(\Omega)$ with $q > n$; at each occurrence we note how the assumptions that follow could be weakened from this point of view. To make the presentation clearer we simply assume that $p = q > n$.

In addition we make the following assumptions.

(A1) A is a linear elliptic differential operator defined by

$$\begin{gathered} Ay = -\sum_{i,j=1}^{n} \partial_{x_i}(a_{ij}(x)\partial_{x_j} y) + a_0(x)y \quad \text{with} \\ a_{ij} \in \mathcal{C}^2(\overline{\Omega}) \ for\ i,j = 1 \cdots n, \\ a_0 \in L^\infty(\Omega), \sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq m_o \sum_{i=1}^{n} \xi_i^2 \ \ \forall x \in \overline{\Omega}, \forall \xi \in \mathbb{R}^{\mathrm{n}}, m_o > 0. \end{gathered} \tag{2.1}$$

(A2) $f\ : \mathbb{R} \to \mathbb{R}$ is a monotone increasing, globally Lipschitz $\mathcal{C}^1$-function.

Remark 2.2. The monotonicity assumption on f can be relaxed and replaced by

$$\exists c_o \in \mathbb{R} \qquad f'_y \geq c_o.$$

An appropriate translation shows that we retrieve the case where f is monotonically increasing, so we assume this for the sake of simplicity.

On the other hand one could consider a mapping f from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$ depending on both y and u. The method would work in the same way. (In what follows, we denote the real function $f : \mathbb{R} \to \mathbb{R}$ and the Nemytski operator associated to $f : y(\cdot) \mapsto f(y(\cdot))$ in $L^p(Q)$ by the same symbol f.)

(A3) $\varphi\ : W^{1,p}_o(\Omega) \to \mathbb{R} \cup \{+\infty\}$ is a proper (i.e., nonidentically equal to $+\infty$), convex, lower semicontinuous function such that $0 \in \text{dom } \varphi$.

(A4) $y_o \in \text{dom } \varphi$.

(A5) For every $(y,u) \in \mathbb{R}^2$, $F(\cdot,y,u)$ is measurable on Q. For almost every $(x,t) \in Q$, for every $u \in \mathbb{R}$, $F(x,t,\cdot,u)$ is $\mathcal{C}^1$ on $\mathbb{R}$. For almost every $(x,t) \in Q$, $F(x,t,\cdot)$ and $F'_y(x,t,\cdot)$ are continuous on $\mathbb{R}^2$. The following estimate holds:

$$|F(x,t,y,u)| + |F'_y(x,t,y,u)| \leq (M_1(x,t) + m_1|u|^p)\eta(|y|),$$

where $M_1 \in L^1(Q)$, $m_1 \geq 0$, and η is a nondecreasing function from $\mathbb{R}^+$ to $\mathbb{R}^+$.

(A6) For every $y \in \mathbb{R}$, $L(\cdot,y)$ is measurable on Ω. For almost every $x \in \Omega$, $L(x,\cdot)$ is $\mathcal{C}^1$ on $\mathbb{R}$. The following estimate holds:

$$|L(x,y)| + |L'_y(x,y)| \leq M_2(x)\eta(|y|),$$

where $M_2 \in L^1(\Omega)$, η is as in (A5).

(A7) Φ is a $\mathcal{C}^1$ mapping from $C(\overline{Q})$ into $C(\overline{Q})$, and $\mathcal{C}$ is a closed convex subset of $C(\overline{Q})$ with finite codimension.

We recall that for $p \in \mathbb{N}$

$$W^{1,p}(\Omega) = \{y \in L^p(\Omega) \mid \nabla y \in L^p(\Omega)^n \ \} \ \text{ and}$$
$$W^{2,1,p}(Q) = \left\{ y \in L^p(Q) \mid Dy,\ D^2y,\ \frac{\partial y}{\partial t} \in L^p(Q) \right\}.$$

3. Existence and regularity of solutions to the variational inequality. Let V and H be Hilbert spaces such that $V \subset H \subset V'$ with continuous and dense injections. We denote by $(\cdot,\cdot)_V$ the V-scalar product, $\langle \cdot,\cdot \rangle$ the duality product between V and V', and $\|\cdot\|_V$ the V-norm. We consider a linear, continuous V-elliptic operator $\mathcal{A}$ from V to V' and ϕ a convex, proper, and lower semicontinuous function from V to $\mathbb{R} \cup \{+\infty\}$. Then we may define the variational inequality

$$\begin{cases} \dfrac{\partial y}{\partial t}(t) + \mathcal{A}y(t) + \partial\phi(y)(t) \ni u(t) \text{ a.e. t in } [0,T], \\ y(0) = y_o \end{cases} \tag{3.1}$$

in the following (variational) sense:

$$\left\langle \frac{\partial y}{\partial t}(t) + \mathcal{A}y(t), y(t) - z \right\rangle + \phi(y(t)) - \phi(z) \leq \langle\, f(t), y(t) - z\rangle \quad \text{a.e. } t \in (0,T)\ \forall z \in V. \tag{3.2}$$

Here $\partial\phi(y(t))$ denotes the subdifferential of ϕ at $z = y(t) \in V$ [8]:

$$\partial\phi(z) = \{\ z^* \in V' \mid \phi(z) - \phi(\zeta) \leq \langle z - \zeta, z^*\rangle\ \forall \zeta \in V\ \}. \tag{3.3}$$

Now we set $V = H^1_o(\Omega)$ and $H = L^2(\Omega)$; we let g be a primitive function of f (such that $g(0) = 0$ for example) and define

$$\phi = \varphi + g, \tag{3.4}$$

where φ is given by (A3). Then $\partial\phi = g' + \partial\varphi = f + \partial\varphi$ (g is the regular part of ϕ). Therefore (1.1) makes sense in the (3.1) form with $\mathcal{A} = A$ and we may give a first existence and regularity result as in the following theorem.

THEOREM 3.1. *Set $p \geq 2$; let $u \in L^p(Q)$ and $y_o \in W^{1,p}_o(\Omega)$. Assume that*

$$\exists \gamma \in L^p(\Omega) \cap \partial\varphi(y_o); \tag{3.5}$$

then (1.1) *has a unique solution $y \in W^{2,1,p}(Q)$.*

Proof. We first use a result of Tiba [22, Theorem 4.5, p. 26] that ensures that if β is a maximal monotone graph $\subset \mathbb{R} \times \mathbb{R}$, $u \in L^p(Q)$ and $y_o \in W^{1,p}_o(\Omega)$, then the parabolic variational inequality

$$\begin{cases} \dfrac{\partial y}{\partial t} + Ay + \beta(y) & \ni u \quad \text{a.e. in } Q,\\ y(0,x) & = y_o(x) \quad \text{a.e. on } \Omega,\\ y(t,x) & = 0 \quad \text{a.e. on } \Sigma \end{cases}$$

has a unique solution in $W^{2,1,p}(Q)$ if the compatibility relation

$$\begin{gathered} 0 \in \text{ dom } \beta,\ y_o(x) \in \text{ dom } \beta \text{ a.e. in } \Omega,\\ \exists \gamma \in L^p(\Omega) \text{ such that } \gamma(x) \in \beta(y_o(x)) \text{ a.e. in } \Omega \end{gathered} \tag{3.6}$$

is fulfilled. One can apply this result to $\beta = f + \partial\varphi$, which is a maximal monotone graph since f is monotone increasing and φ is convex, lower semicontinuous, and proper. It remains to check (3.6), that is,

$$\exists \gamma \in L^p(\Omega) \text{ such that } \gamma(x) \in f(y_o(x)) + \partial\varphi(y_o(x)) \ \text{ a.e. in } \Omega.$$

This is equivalent to

$$\exists \gamma \text{ such that } \gamma + f(y_o) \in L^p(\Omega), \text{ and } \gamma(x) \in \partial\varphi(y_o(x)) \text{ a.e. in } \Omega.$$

Since f is globally Lipschitz then $f(y_o) \in L^p(\Omega)$ and we get the result. □

We set

$$\xi = u - \frac{\partial y}{\partial t} - Ay - f(y) \in L^p(Q)$$

(since f is globally Lipschitz and $y \in W^{2,1,p}(Q)$). In addition, $\xi(t) \in \partial\varphi(y(t))$ almost everywhere in $]0,T[$; using the characterization of the subdifferential of a function in Banach spaces this gives

$$\varphi(y(t)) + \varphi^*(\xi(t)) - \langle y(t), \xi(t) \rangle = 0 \text{ a.e in }]0,T[. \tag{3.7}$$

In this last relation $\langle , \rangle$ denotes the duality product between $V = W_o^{1,p}(\Omega)$ and V', and φ^* is the conjugate function of φ. For more details refer to Barbu–Precupanu [1] or Ekeland–Temam [13]. It follows that the variational inequality (1.1) is equivalent to

$$\begin{aligned} \frac{\partial y}{\partial t} + Ay + f(y) &= u + \xi && \text{in } Q, \\ y &= 0 && \text{on } \Sigma, \\ y(x,0) &= y_o(x) && \text{in } \Omega \end{aligned} \tag{3.8}$$

and (3.7). Because $y_o \in W_o^{1,p}(\Omega)$ and $(u,\xi) \in L^p(Q) \times L^p(Q)$, the solution y of equation (3.8) belongs to $C(\overline{Q}) \cap W^{2,1,p}(Q)$. More precisely, we have the following theorem.

THEOREM 3.2. (i) *If $p > n/2+1$ and $(u,\xi,y_o) \in L^p(Q) \times L^p(Q) \times \mathcal{C}(\overline{\Omega})$, then* (3.8) *has a unique weak solution $y_{u\xi}$ in $W(0,T) \cap C(\overline{Q})$ which satisfies*

$$\|y_{u\xi}\|_{\infty,Q} \le C_1(\|u\|_{p,Q} + \|\xi\|_{p,Q} + \|y_o\|_{\infty,\Omega} + 1),$$

where $C_1 = C_1(T,\Omega,m_0,n,p)$. Moreover, for every $\varepsilon > 0$, $y_{u\xi}$ is Hölder continuous on $[\varepsilon,T] \times \overline{\Omega}$ and belongs to $W^{2,1,p}(\Omega \times]\varepsilon,T[)$.
(ii) *If $p > n$ and $(u,\xi,y_o) \in L^p(Q) \times L^p(Q) \times W_o^{1,p}(\Omega)$,* (3.8) *has a unique weak solution $y_{u\xi}$ in $W^{2,1,p}(Q) \cap C(\overline{Q})$.*

Proof. The existence of a unique weak solution $y_{u\xi}$ in $W(0,T) \cap C(\overline{Q})$ for (3.8) can be proved as in the case of the Robin boundary condition (see Raymond–Zidani [20, 21]). The Hölder continuity result holds thanks to [9]. Point (ii) can be found in Bergounioux–Tröltzsch [4]. □

4. Optimal control of the obstacle problem.

4.1. The obstacle problem. Now we focus on the very case of control of the obstacle problem, where

$$K_o = \{ z \in W_o^{1,p}(\Omega) \mid z \ge 0 \text{ a.e. in } \Omega \ \} \tag{4.1}$$

and φ is the indicator function of K_o:

$$\varphi(z) = \begin{cases} 0 & \text{if } z \in K_o, \\ +\infty & \text{else.} \end{cases}$$

It is clear that $0 \in \text{dom } \varphi = K_o$. Moreover, the compatibility condition (3.5) is fulfilled with $\gamma = 0$ so that Theorem 3.1 is valid. On the other hand, the (classical) calculus of φ^* shows that relation (3.7) is equivalent to

$$\begin{gathered} y(t) \ge 0 \text{ in } \Omega \quad \forall t \in]0,T[\ , \quad \xi(t) \ge 0 \text{ in } \Omega, \\ \text{and } \int_\Omega y(t,x)\ \xi(t,x)\ dx = 0 \ \text{ a.e. } t \in]0,T[, \end{gathered} \tag{4.2}$$

that is, at last

$$y \geq 0 \text{ in } Q \ , \xi \geq 0 \text{ a.e. in } Q, \text{ and } \int_Q y(t,x)\ \xi(t,x)\ dx\ dt = 0.$$

We may summarize in the following theorem.

THEOREM 4.1. *Assume $p > n$, $(u, y_o) \in L^p(Q) \times W_o^{1,p}(\Omega)$; then the variational inequality*

$$\frac{\partial y}{\partial t} + Ay + f(y) + \partial\varphi(y) \ni u \quad \text{in } Q, \quad y = 0 \quad \text{on } \Sigma, \quad y(0) = y_o \quad \text{in } \Omega, \tag{4.3}$$

where φ is the indicator function of K_o, has a unique solution $y \in \mathcal{C}(\overline{Q}) \cap W^{2,1,p}(Q)$. Moreover, it is equivalent to

$$\begin{cases} \dfrac{\partial y}{\partial t} + Ay + f(y) = u + \xi \quad \text{in } Q, \quad y = 0 \quad \text{on } \Sigma, \quad y(x,0) = y_o(x) \quad \text{in } \Omega, \\ \xi \geq 0, \quad y \geq 0, \quad \displaystyle\int_Q y(t,x)\ \xi(t,x)\ dx\ dt = 0. \end{cases} \tag{4.4}$$

In the following we denote

$$V_{ad} = \{\xi \in L^p(Q) \mid \xi \geq 0 \text{ a.e. in } Q\ \}. \tag{4.5}$$

4.2. Pontryagin principle. Now we consider the following problem $(\widetilde{\mathcal{P}})$: Minimize $J(y, u)$ subject to

$$\frac{\partial y}{\partial t} + Ay + f(y) = u + \xi \text{ in } Q, \quad y = 0 \text{ on } \Sigma,\ y(.,0) = y_o \text{ in } \Omega, \tag{4.6a}$$

$$\tilde{\Phi}(y) \in \widetilde{\mathcal{C}} \qquad (\text{“pure” state constraint}), \tag{4.6b}$$

$$(u, \xi) \in U_{ad} \times V_{ad} \qquad (\text{“pure” control constraints}), \tag{4.6c}$$

$$\int_Q y(t,x)\ \xi(t,x)\ dx\ dt = 0 \qquad (\text{mixed state-control integral constraints}), \tag{4.6d}$$

where

$$\tilde{\Phi}(y) = (\Phi(y), y) \text{ and } \widetilde{\mathcal{C}} = \mathcal{C} \times \{y \in \mathcal{C}(\overline{Q}) \mid y \geq 0\ \}. \tag{4.7}$$

The results of section 3 yield that problems $(\mathcal{P})$ and $(\widetilde{\mathcal{P}})$ are equivalent. In particular if $(\bar{y}, \bar{u})$ is a solution of $(\mathcal{P})$, then there exists $\bar{\xi} \in L^p(Q)$ such that $(\bar{y}, \bar{u}, \bar{\xi})$ is an optimal solution of $(\widetilde{\mathcal{P}})$ with $\bar{\xi} = \partial\bar{y}/\partial t + A\bar{y} + f(\bar{y}) - \bar{u}$. Let us mention that we are interested not in existence results (although we will give an example in the last section of this paper) but in optimality conditions for $(\bar{y}, \bar{u})$. Consequently, we study optimality conditions for $(\bar{y}, \bar{u}, \bar{\xi})$ to get those for $(\bar{y}, \bar{u})$.

Let us define the Hamiltonian functions by

$$H_1(x,t,y,u,q,\nu) = \nu F(x,t,y,u) + q\ u \tag{4.8}$$

for every $(x,t,y,u,q,\nu) \in Q \times \mathbb{R}^4$, and

$$H_2(y,\xi,q,\lambda) = q\ \xi + \lambda\ y\ \xi \tag{4.9}$$

for every $(y,\xi,q,\lambda) \in \mathbb{R}^4$.

THEOREM 4.2 (Pontryagin principle for $(\widetilde{\mathcal{P}})$). *If* (A1)−(A7) *are fulfilled and if* $(\bar{y},\bar{u},\bar{\xi})$ *is a solution of* $(\widetilde{\mathcal{P}})$*, then there exist* $\bar{q} \in L^1(0,T;W_o^{1,1}(\Omega))$*,* $\bar{\nu} \in \mathbb{R}$*,* $\bar{\lambda} \in \mathbb{R}$*, and* $(\bar{\mu},\bar{\theta}) \in \mathcal{M}(\overline{Q}) \times \mathcal{M}(\overline{Q})$ *(*$\mathcal{M}(\overline{Q})$ *is the space of Radon measures on* $\overline{Q}$*), such that*

$$(\bar{\nu},\bar{\lambda},\bar{\mu},\bar{\theta}) \neq 0, \quad \bar{\nu} \geq 0, \tag{4.10a}$$

$$\forall z \in \{ z \in \mathcal{C}(\overline{Q}) \mid z \geq 0 \}\ \ \langle \bar{\mu}, z - \bar{y} \rangle_{\overline{Q}} \leq 0,\ \ \textit{and}\ \forall z \in \mathcal{C}\ \ \langle \bar{\theta}, z - \Phi(\bar{y}) \rangle_{\overline{Q}} \leq 0, \tag{4.10b}$$

$$\begin{cases} -\dfrac{\partial \bar{q}}{\partial t} + A^*\bar{q} + f'_y(\bar{y})\bar{q} = \bar{\nu}F'_y(x,t,\bar{y},\bar{u}) + \bar{\mu}|_Q + [\Phi'(\bar{y})^*\bar{\theta}]|_Q + \bar{\lambda}\bar{\xi} & \textit{in } Q, \\ \bar{q} = 0 \quad \textit{on } \Sigma, \qquad \bar{q}(T) = \bar{\nu}L'_y(x,\bar{y}(T)) + \bar{\mu}|_{\overline{\Omega}_T} + [\Phi'(\bar{y})^*\bar{\theta}]|_{\overline{\Omega}_T} & \textit{in } \Omega, \end{cases} \tag{4.10c}$$

$$\bar{q} \in L^{\delta'}(0,T;W_o^{1,d'}(\Omega)) \qquad \textit{for every } (\delta,d) \textit{ satisfying } \frac{n}{2d} + \frac{1}{\delta} < \frac{1}{2}, \tag{4.10d}$$

$$\begin{aligned} H_1(x,t,\bar{y}(x,t),\bar{u}(x,t),\bar{q}(x,t),\bar{\nu}) = \min_{u \in \mathcal{K}_U(x,t)} \\ H_1(x,t,\bar{y}(x,t),u(x,t),\bar{q}(x,t),\bar{\nu}) \quad \textit{a.e. in } Q, \end{aligned} \tag{4.10e}$$

$$H_2(\bar{y}(x,t),\bar{\xi}(x,t),\bar{q}(x,t),\bar{\lambda}) = \min_{\xi \in \mathbb{R}^+} H_2(\bar{y}(x,t),\xi,\bar{q}(x,t),\bar{\lambda}) \qquad \textit{a.e. in } Q, \tag{4.10f}$$

where $\bar{\mu}|_Q$ *(resp.,* $[\Phi'(\bar{y})^*\bar{\theta}]|_Q$*) is the restriction of* $\bar{\mu}$ *(resp.,* $[\Phi'(\bar{y})^*\bar{\theta}]$*) to* Q*,* $\bar{\mu}|_{\overline{\Omega}_T}$ *(resp.,* $[\Phi'(\bar{y})^*\bar{\theta}]|_{\overline{\Omega}_T}$*) is the restriction of* $\bar{\mu}$ *(resp.,* $[\Phi'(\bar{y})^*\bar{\theta}]$*) to* $\overline{\Omega} \times \{T\}$*,* $\langle \cdot,\cdot \rangle_{\overline{Q}}$ *denotes the duality product between* $\mathcal{M}(\overline{Q})$ *and* $C(\overline{Q})$*,* A^* *is the adjoint operator of* A*, and* $\frac{1}{d} + \frac{1}{d'} = 1$*.*

Remark 4.1. We briefly describe these relations: $(\bar{\mu},\bar{\theta})$ are the multipliers associated with the state constraints; $\bar{\mu}$ corresponds to $y \geq 0$; and an immediate consequence of relation (4.10b) is the complementarity result $\bar{\mu} \leq 0$, $\langle \bar{\mu}, \bar{y} \rangle_{\overline{Q}} = 0$. $\bar{\theta}$ is associated to the (general) constraint $\Phi(y) \in \mathcal{C}$. $\bar{\lambda}$ is the multiplier associated to the integral constraint $\int_Q y(t,x)\ \xi(t,x)\ dx\ dt = 0$, and $\bar{q}$ is the classical adjoint state which takes into account the cost functional via $\bar{\nu}$.

Condition (4.10a) is a nontriviality condition. We must emphasize that we get (a priori) *nonqualified* optimality conditions. If $\bar{\nu} \neq 0$, the problem is qualified.

Remark 4.2. One may note that if $\bar{\xi} = 0$, then it could happen that $\bar{\nu} = \bar{\mu} = \bar{\theta} = 0$ and $\bar{\lambda} \neq 0$, so that $\bar{q} = 0$; therefore, the optimality system could appear to be useless. However, this is the case where the solution $(\bar{y},\bar{u})$ is the solution of a control problem governed by a classical semilinear parabolic equation, since we have

$\partial\bar y/\partial t + A\bar y + f(\bar y) = \bar u$ and the associated optimality systems are well known for this kind of problem. We refer for instance to [20].

Theorem 4.3 (Pontryagin principle for $(\mathcal{P})$). *If* (A1)–(A7) *are fulfilled and if* $(\bar y, \bar u)$ *is a solution of* $(\mathcal{P})$*, then there exists* $\bar q \in L^1(0,T;W_o^{1,1}(\Omega))$*,* $\bar\nu \in \mathbb{R}$*,* $\bar\lambda \in \mathbb{R}$*,* $(\bar\mu,\bar\theta) \in \mathcal{M}(\overline{Q}) \times \mathcal{M}(\overline{Q})$ *such that* (4.10a), (4.10b), (4.10d), *and* (4.10e) *hold. Moreover, we have*

(4.11a)

$$\begin{cases} -\dfrac{\partial \bar q}{\partial t} + A^*\bar q + f_y'(\bar y)\bar q = \bar\nu F_y'(x,t,\bar y,\bar u) + \bar\mu_{|Q} \\ \qquad +[\Phi'(\bar y)^*\bar\theta]_{|Q} + \bar\lambda\left(\dfrac{\partial \bar y}{\partial t} + A\bar y + f(\bar y) - \bar u\right) \quad \textit{in } Q, \\ \qquad \bar q = 0 \quad \textit{on } \Sigma, \qquad \bar q(T) = \bar\nu L_y'(x,\bar y(T)) + \bar\mu|_{\overline{\Omega}_T} + [\Phi'(\bar y)^*\bar\theta]|_{\overline{\Omega}_T} \quad \textit{in } \Omega, \end{cases}$$

$$\text{(4.11b)} \qquad \bar q(x,t)\left(\frac{\partial \bar y}{\partial t} + A\bar y + f(\bar y) - \bar u\right)(x,t) = 0 \quad \textit{a.e. } (x,t) \in Q.$$

Remark 4.3. Relation (4.11b) is a pointwise complementarity condition. Therefore, $\bar q$ may be viewed as a Lagrange multiplier associated with the pointwise constraint

$$\left(\frac{\partial y}{\partial t} + Ay + f(y) - u\right)(x,t) \geq 0.$$

Let us recall a regularity result for a weak solution of parabolic equation with measures as data, as follows.

Proposition 4.1. *Let* μ *be in* $\mathcal{M}_b(\overline{Q} \setminus (\overline{\Omega} \times \{0\} \cup \overline{\Sigma}))$ *and let* a *be in* $L^p(Q)$ *satisfying*

$$a \geq C_0, \quad \|a\|_{L^p(Q)} \leq M,$$

where $M > 0$*. Consider the equation*

$$\text{(4.12)} \quad -\frac{\partial q}{\partial t} + A^*q + aq = \mu_Q \quad \textit{in } Q, \qquad q = 0 \quad \textit{on } \Sigma, \qquad q(T) = \mu_{\overline{\Omega}_T} \quad \textit{on } \overline{\Omega},$$

where $\mu = \mu_Q + \mu_{\overline{\Omega}_T}$ *is a bounded Radon measure on* $\overline{Q} \setminus (\overline{\Omega} \times \{0\} \cup \overline{\Sigma})$*,* μ_Q *is the restriction of* μ *to* Q*, and* $\mu_{\overline{\Omega}_T}$ *is the restriction of* μ *to* $\overline{\Omega} \times \{T\}$*. Equation* (4.12) *admits a unique weak solution* $q \in L^1(0,T;W_o^{1,1}(\Omega))$*. For every* (δ, d) *satisfying* $d > 2$*,* $\delta > 2$*,* $\frac{n}{2d} + \frac{1}{\delta} < \frac{1}{2}$*,* $q \in L^{\delta'}(0,T;W_o^{1,d'}(\Omega))$*, and we have*

$$\|q\|_{L^{\delta'}(0,T;W_o^{1,d'}(\Omega))} \leq C_2 \|\mu\|_{\mathcal{M}_b(\overline{Q}\setminus(\overline{\Omega}\times\{0\}\cup\overline{\Sigma}))},$$

where $C_2 = C_2(T,\Omega,n,C_0,M,p,\delta,d)$ *is independent of* a*. Moreover, there exists a function* $q(0) \in L^1(\Omega)$ *such that*

$$\int_Q q\left\{\frac{\partial y}{\partial t} + Ay + ay\right\} dxdt = \langle y, \mu\rangle_b - \langle y(0), q(0)\rangle_{C(\overline{\Omega})\times\mathcal{M}(\overline{\Omega})}$$

for every $y \in Y = \{y \in W(0,T) \cap C(\overline{Q}) \mid \frac{\partial y}{\partial t} + Ay \in L^p(Q), y = 0 \textit{ in } \Sigma\}$*, where* $\langle\cdot,\cdot\rangle_b$ *denotes the duality product between* $C_b(\overline{Q} \setminus (\overline{\Omega} \times \{0\} \cup \overline{\Sigma}))$ *and* $\mathcal{M}_b(\overline{Q} \setminus (\overline{\Omega} \times \{0\} \cup \overline{\Sigma}))$*.*

($C_b(\overline{Q}\setminus(\overline{\Omega}\times\{0\}\cup\overline{\Sigma}))$ denotes the space of bounded continuous functions on $\overline{Q}\setminus(\overline{\Omega}\times\{0\}\cup\overline{\Sigma})$, while $\mathcal{M}_b(\overline{Q}\setminus(\overline{\Omega}\times\{0\}\cup\overline{\Sigma}))$ denotes the space of bounded Radon measures on $\overline{Q}\setminus(\overline{\Omega}\times\{0\}\cup\overline{\Sigma})$, that is, the topological dual of $C_o(\overline{Q}\setminus(\overline{\Omega}\times\{0\}\cup\overline{\Sigma}))$.)

Proof. The proof is the same as the one given in [19] for the Neumann boundary conditions (see also [7]). An easy adaptation of this proof yields the previous result. However, for the convenience of the reader we recall that q is the weak solution of (4.12) if and only if q belongs to $L^1(0,T;W_o^{1,1}(\Omega))$, $aq\in L^1(Q)$, and for every $\varphi\in C^1(\overline{Q})$ satisfying $\varphi(x,0)=0$ on $\overline{\Omega}$ and $\varphi(\cdot)=0$ on Σ we have

$$\int_Q\left\{q\frac{\partial\varphi}{\partial t}+\Sigma_{i,j}a_{ij}D_j\varphi D_iq+a\varphi q\right\}dx\,dt=\langle\varphi,\mu\rangle_b.$$

As in [7], we can prove that the weak solution q belongs to $L^{\delta'}(0,T;W_o^{1,d'}(\Omega))$ for every (δ,d) satisfying the condition

$$d>2,\quad \delta>2,\quad \frac{n}{2d}+\frac{1}{\delta}<\frac{1}{2}. \tag{4.13}$$

We remark that the set of pairs (δ,d) satisfying the above condition is nonempty. We remark also that if (δ,d) satisfies (4.13), if a belongs to $L^p(Q)$, and if q belongs to $L^{\delta'}(0,T;W_o^{1,d'}(\Omega))$, then $aq\in L^1(Q)$. Now, since $q\in L^{\delta'}(0,T;W_o^{1,d'}(\Omega))$ (where (δ,d) satisfies (4.24)), and since

$$\mathrm{div}_{xt}\left((\Sigma_j a_{ij}D_jq)_{1\le i\le n},q\right)=\frac{\partial q}{\partial t}-Aq \text{ belongs to } \mathcal{M}_b(Q),$$

then we can define the normal trace of the vector field $((\sum_j a_{ij}D_jq)_{1\le i\le n},q)$ in the space $W^{\frac{-1}{m},m}(\partial Q)$ (for some $1<m<\frac{n+1}{n}$). If we denote by $\gamma_o((\sum_j a_{ij}D_jq)_{1\le i\le n},q)$ this normal trace, we can prove (see Theorem 4.2 in [19]) that this normal trace belongs to $\mathcal{M}(\partial Q)$ and the restriction of $\gamma_o((\sum_j a_{ij}D_jq)_{1\le i\le n},q)$ to $\overline{\Omega}\times\{T\}$ is equal to $\mu_{\overline{\Omega}_T}$, and if $q(0)$ is the measure on $\overline{\Omega}$ which satisfies the Green formula of our Theorem 3.2, then $-q(0)$ is the restriction of $\gamma_o((\sum_j a_{ij}D_jq)_{1\le i\le n},q)$ to $\overline{\Omega}\times\{0\}$. In fact, it can be proved that $q(0)$ belongs to $L^1(\Omega)$ (see Theorem 4.3 in [19]). □

4.3. Proof of Theorems 4.2–4.3. First we assume that Theorem 4.2 is valid. As mentioned before, if $(\bar y,\bar u)$ is an optimal solution for $(\mathcal{P})$, then $(\bar y,\bar u,\bar\xi)$ is a solution for $(\widetilde{\mathcal{P}})$, where $\bar\xi=\frac{\partial\bar y}{\partial t}+A\bar y+f(\bar y)-\bar u\ \in L^p(Q)$. Thanks to Theorem 4.2, there exist $(\bar\nu,\bar\lambda,\bar\mu,\bar q)$ such that (4.10) holds. Replacing $\bar\xi$ by its value in (4.10c) obviously leads to (4.11a). Furthermore, relation (4.10f) implies

$$(\bar q(x,t)+\bar\lambda\bar y(x,t))\ (\bar\xi(x,t)-\xi)\le 0\quad\text{a.e. } (x,t)\in Q\ \forall\xi\in\mathbb{R}^+,$$

which gives

$$(\bar q(x,t)+\bar\lambda\bar y(x,t))\ \bar\xi(x,t)=0\quad\text{a.e. } (x,t)\in Q.$$

Since $\bar y(x,t)\,\bar\xi(x,t)=0$ a.e. in Q we obtain (4.11b). This concludes the proof of Theorem 4.3. □

It remains to show that Theorem 4.2 is valid. Note that Pontryagin's principle for a control problem with unbounded controls, with pointwise state constraints, and with state-control constraints in integral form already have been studied in [7]. For the convenience of the reader, we give the main ideas of the proof.

Step 1: *Metric space of controls.* In this paper we shall consider control problems for which the state constraints (4.6b) and the state-control integral constraints (4.6d) are penalized. These problems are constructed in such a way to make $(\bar y, \bar u, \bar\xi)$ an approximate solution. The idea is to apply the Ekeland variational principle next. For this we have to define a metric space of controls, endowed with the so-called Ekeland distance d, to make the mapping $(u,\xi) \longmapsto y_{u\xi}$ continuous from this metric space into $\mathcal{C}(\overline{Q})$. Thanks to Theorem 3.2, this continuity condition will be realized if convergence in the metric space of controls implies convergence in $L^p(Q) \times L^p(Q)$. Here, since we deal with (generally) unbounded controls, the convergence in $(U_{ad} \times V_{ad}, d)$ does not imply the convergence in $L^p(Q) \times L^p(Q)$ (see [14, p. 227]). To overcome this difficulty, as in [24, 20], we define a new metric as follows. For $0 < k < \infty$, we set

$$U_{ad}(\bar u, k) = \{u \in U_{ad} \mid |u(x,t) - \bar u(x,t)| \le k \quad \text{a.e. } (x,t) \in Q\},$$

$$V_{ad}(\bar \xi, k) = \{\xi \in V_{ad} \mid |\xi(x,t) - \bar \xi(x,t)| \le k \quad \text{a.e. } (x,t) \in Q\}.$$

We endow the control space with Ekeland's metric:

$$d((u_1,\xi_1),(u_2,\xi_2)) = \mathcal{L}^{n+1}(\{(x,t \mid u_1(x,t) \neq u_2(x,t)\}) + \mathcal{L}^{n+1}(\{(x,t) \mid \xi_1(x,t) \neq \xi_2(x,t)\}),$$

where $\mathcal{L}^{n+1}$ denotes the Lebesgue measure in $\mathbb{R}^{n+1}$. Then, as in [24, 20], we can prove the following lemma.

LEMMA 4.1. *$(U_{ad}(\bar u, k) \times V_{ad}(\bar\xi, k), d)$ is a complete metric space for the distance d, and the mapping which associates $(y_{u\xi}, J(y_{u\xi}, u))$ with (u,ξ) is continuous from $(U_{ad}(\bar u, k) \times V_{ad}(\bar\xi, k), d)$ into $\mathcal{C}(\overline{Q}) \times \mathbb{R}$.*

In [7], the authors have used another method to build the metric space of controls. This construction was adapted to the type of constraints they have considered.

Step 2: *Penalized problems.* Since $\mathcal{C}(\overline{Q})$ is separable, there exists a norm $|\cdot|_{\mathcal{C}(\overline{Q})}$, which is equivalent to the norm $\|\cdot\|_{\mathcal{C}(\overline{Q})}$ such that $(\mathcal{C}(\overline{Q}), |\cdot|_{\mathcal{C}(\overline{Q})})$ is strictly convex and $\mathcal{M}(\overline{Q})$, endowed with the dual norm of $|\cdot|_{\mathcal{C}(\overline{Q})}$ (denoted by $|\cdot|_{\mathcal{M}(\overline{Q})}$), also is strictly convex (see [11, Corollary 2, p. 148, or Corollary 2, p. 167]). Let $\mathbb{K}$ be a convex subset of $\mathcal{C}(\overline{Q})$. We define the distance function to $\mathbb{K}$ (for the new norm $|\cdot|_{\mathcal{C}(\overline{Q})}$) by

$$\delta_{\mathbb{K}}(\zeta) = \inf_{z \in \mathbb{K}} |\zeta - z|_{\mathcal{C}(\overline{Q})}.$$

Since $\mathbb{K}$ is convex, then $\delta_{\mathbb{K}}$ is convex and Lipschitz of rank 1, and we have

$$\limsup_{\substack{\rho \searrow 0, \\ \zeta' \to \zeta}} \frac{\delta_{\mathbb{K}}(\zeta' + \rho z) - \delta_{\mathbb{K}}(\zeta')}{\rho} = \max\{\langle \xi, z\rangle_{\overline{Q}} \mid \xi \in \partial\delta_{\mathbb{K}}(\zeta)\} \tag{4.14}$$

for every $\zeta, z \in \mathcal{C}(\overline{Q})$, where $\partial\delta_{\mathbb{K}}(\zeta)$ is the subdifferential of $\delta_{\mathbb{K}}$ at (ζ). Moreover, as $\mathbb{K}$ is a closed convex subset of $\mathcal{C}(\overline{Q})$ it is proved in [16, Lemma 3.4] that for every $\zeta \notin \mathbb{K}$, and every $\xi \in \partial\delta_{\mathbb{K}}(\zeta)$, $|\xi|_{\mathcal{M}(\overline{Q})} = 1$. Since $\partial\delta_{\mathbb{K}}(\zeta)$ is convex in $\mathcal{M}(\overline{Q})$ and $(\mathcal{M}(\overline{Q}), |\cdot|_{\mathcal{M}(\overline{Q})})$ is strictly convex, then if $\zeta \notin \mathbb{K}$, $\partial\delta_{\mathbb{K}}(\zeta)$ is a singleton and $\delta_{\mathbb{K}}$ is Gâteaux-differentiable at ζ. Let us notice that when $\mathbb{K} := \{z \in \mathcal{C}(\overline{Q}) \mid z \ge 0\}$, the distance function to $\mathbb{K}$ is given by $\delta_{\mathbb{K}}(\zeta) = |\zeta^-|_{\mathcal{C}(\overline{Q})}$, where $\zeta^- = \min(0,\zeta)$.
Endowing $\mathcal{C}(\overline{Q}) \times \mathcal{C}(\overline{Q})$ with the product norm we have similarly $\delta_{\widetilde{\mathcal{C}}}(\tilde\Phi(y))^2 = |y^-|^2_{\mathcal{C}(\overline{Q})} + \delta_{\mathcal{C}}(\Phi(y))^2$ ($\widetilde{\mathcal{C}}$ is defined by (4.7)). Let us consider the penalized functional

$$J_\varepsilon(y,u,\xi) = \left\{ \left[\left(J(y,u) - J(\bar y, \bar u) + \varepsilon^2\right)^+ \right]^2 + \delta_{\widetilde{\mathcal{C}}}(\tilde\Phi(y))^2 + \left(\int_Q y(x,t)\xi(x,t)\,dx\,dt \right)^2 \right\}^{\frac{1}{2}}.$$

With such a choice, for every $\varepsilon > 0$ and $k > 0$, $(\bar{y}, \bar{u}, \bar{\xi})$ is a ε^2-solution of the penalized problem

$$(\mathcal{P}_{k,\varepsilon}) \quad \inf\{J_\varepsilon(y,u,\xi) \mid y \in \mathcal{C}(\overline{Q}), (u,\xi) \in U_{ad}(\bar{u},k)\times V_{ad}(\bar{\xi},k),\ (y,u,\xi) \text{ satisfies } (4.6a)\},$$

i.e.,

$$\inf(\mathcal{P}_{k,\varepsilon}) \le J_\varepsilon(\bar{y},\bar{u},\bar{\xi}) \le \inf(\mathcal{P}_{k,\varepsilon}) + \varepsilon^2$$

(since $\inf(\mathcal{P}_{k,\varepsilon}) \ge 0$ and $J_\varepsilon(\bar{y},\bar{u},\bar{\xi}) = \varepsilon^2$).

For every $k > 0$, we choose $\varepsilon(k) = \varepsilon_k \le \frac{1}{k^{2p}}$ and we denote by $(\mathcal{P}_k)$ the penalized problem $(\mathcal{P}_{k,\varepsilon_k})$. Thanks to Ekeland's principle [13, p. 30], for every $k \ge 1$ there exists $(u_k, \xi_k) \in U_{ad}(\bar{u},k) \times V_{ad}(\bar{\xi},k)$ such that

$$d((u_k,\xi_k),(\bar{u},\bar{\xi})) \le \varepsilon_k \le \frac{1}{k^{2p}}, \tag{4.15a}$$

$$J_{\varepsilon_k}(y_k,u_k,\xi_k) \le J_{\varepsilon_k}(y_{u\xi},u,\xi) + \varepsilon_k\ d((u_k,\xi_k),(u,\xi)) \tag{4.15b}$$

for every $(u,\xi) \in U_{ad}(\bar{u},k) \times V_{ad}(\bar{\xi},k)$ (y_k and $y_{u\xi}$ being the states corresponding respectively to (u_k,ξ_k) and (u,ξ)). In view of the definition of ε_k, we have $\lim_k \|u_k - \bar{u}\|_{p,Q} = \lim_k \|\xi_k - \bar{\xi}\|_{p,Q} = 0$. Indeed, $\mathcal{L}^{n+1}(\{(x,t) \mid u_k(x,t) \ne \bar{u}(x,t)\}) + \mathcal{L}^{n+1}(\{(x,t) \mid \xi_k(x,t) \ne \bar{\xi}(x,t)\}) \le \frac{1}{k^{2p}}$, and $|u_k(x,t) - \bar{u}(x,t)| \le k$, $|\xi_k(x,t) - \bar{\xi}(x,t)| \le k$ a.e. on Q. Thus $\|u_k - \bar{u}\|_{p,Q} \le \frac{1}{k}$, $\|\xi_k - \bar{\xi}\|_{p,Q} \le \frac{1}{k}$.

To exploit the approximate optimality conditions (4.15), we introduce a particular perturbation of (u_k,ξ_k).

Step 3: *Diffuse perturbations.* For fixed (u_o,ξ_o) in $U_{ad} \times V_{ad}$, we denote by (u_{ok},ξ_{ok}) $(k > 0)$ the pair of functions in $U_{ad}(\bar{u},k) \times V_{ad}(\bar{\xi},k)$ defined by

$$u_{ok}(x,t) = \begin{cases} u_o(x,t) & \text{if } |u_o(x,t) - \bar{u}(x,t)| \le k, \\ \bar{u}(x,t) & \text{if not,} \end{cases} \tag{4.16a}$$

$$\xi_{ok}(x,t) = \begin{cases} \xi_o(x,t) & \text{if } |\xi_o(x,t) - \bar{\xi}(x,t)| \le k, \\ \bar{\xi}(x,t) & \text{if not.} \end{cases} \tag{4.16b}$$

Observe that for every $k \ge 1$, (u_{ok},ξ_{ok}) belongs to $U_{ad}(\bar{u},k) \times V_{ad}(\bar{\xi},k)$, and that $(u_{ok},\xi_{ok})_k$ converges to (u_o,ξ_o) in $L^p(Q) \times L^p(Q)$. Applying Theorem 4.1 of [7] (see also [24, 21] for more details), we deduce the existence of measurable sets E^k_ρ with $\mathcal{L}^{n+1}(E^k_\rho) = \rho\mathcal{L}^{n+1}(Q)$, such that if we denote by (u^ρ_k,ξ^ρ_k) the pair of controls defined by

(4.17)
$$u^\rho_k(x,t) = \begin{cases} u_k(x,t) & \text{on } Q \setminus E^k_\rho, \\ u_{ok}(x,t) & \text{on } E^k_\rho, \end{cases} \qquad \xi^\rho_k(x,t) = \begin{cases} \xi_k(x,t) & \text{on } Q \setminus E^k_\rho, \\ \xi_{ok}(x,t) & \text{on } E^k_\rho \end{cases}$$

and if y^ρ_k is the state corresponding to (u^ρ_k,ξ^ρ_k), then we have

$$y^\rho_k = y_k + \rho z_k + r^\rho_k, \qquad \lim_{\rho\to 0} \frac{1}{\rho}|r^\rho_k|_{\mathcal{C}(\overline{Q})} = 0, \tag{4.18a}$$

$$J(y^\rho_k,u^\rho_k) = J(y_k,u_k) + \rho\Delta_k J + o(\rho), \tag{4.18b}$$

$$\int_Q y_k^\rho \xi_k^\rho \, dx\ dt = \int_Q y_k \xi_k \, dx\ dt + \rho \int_Q \left[z_k \xi_k + y_k(\xi_{ok} - \xi_k) \right] \, dx\ dt + o(\rho), \tag{4.18c}$$

where z_k is the weak solution of

$$\frac{\partial z_k}{\partial t} + A z_k + f'_y(y_k) z_k = u_k - u_{ok} + \xi_k - \xi_{ok} \quad \text{in } Q, \qquad z_k = 0 \quad \text{on } \Sigma, \qquad z_k(0) = 0 \quad \text{in } \Omega,$$

and

$$\Delta_k J = \int_Q \Big[F'_y(x,t,y_k,u) z_k + F(x,t,y_k,u_{ok}) - F(x,t,y_k,u_k) \Big] \, dx\ dt + \int_\Omega L'_y(x, y_k(T)) z_k(T) \, dx.$$

Setting $(u, \xi) = (u_k^\rho, \xi_k^\rho)$ in (4.15b), it follows that

$$\limsup_{\rho \to 0} \frac{J_{\varepsilon_k}(y_k, u_k, \xi_k) - J_{\varepsilon_k}(y_k^\rho, u_k^\rho, \xi_k^\rho)}{\rho} \le \varepsilon_k \mathcal{L}^{n+1}(Q). \tag{4.19}$$

Taking (4.18) and the definition of J_{ε_k} into account, we get

(4.20)
$$-\nu_k \Delta_k J - \langle \mu_k, z_k \rangle_{\overline{Q}} - \langle \theta_k, \Phi'(y_k) z_k \rangle_{\overline{Q}} - \lambda_k \Big[\langle \xi_k, z_k \rangle_{\overline{Q}} + \langle y_k, \xi_{ok} - \xi_k \rangle_{\overline{Q}} \Big] \le \varepsilon_k \mathcal{L}^{n+1}(Q),$$

where

$$\nu_k = \frac{(J(y_k,u_k) - J(\bar y, \bar u) + \varepsilon_k^2)^+}{J_{\varepsilon_k}(y_k,u_k,\xi_k)}, \lambda_k = \frac{\left(\displaystyle\int_Q y_k(x,t) \xi_k(x,t) \, dx\ dt \right)}{J_{\varepsilon_k}(y_k,u_k,\xi_k)},$$

$$\mu_k = \begin{cases} \dfrac{|y_k^-|_{\mathcal{C}(\overline{Q})} \nabla |y_k^-|_{\mathcal{C}(\overline{Q})}}{J_{\varepsilon_k}(y_k,u_k,\xi_k)} & \text{if } |y_k^-|_{\mathcal{C}(\overline{Q})} \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$\theta_k = \begin{cases} \dfrac{\delta_{\mathcal{C}}(\Phi(y_k) \nabla \delta_{\mathcal{C}}(\Phi(y_k))}{J_{\varepsilon_k}(y_k,u_k,\xi_k)} & \text{if } \delta_{\mathcal{C}}(\Phi(y_k)) \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

For every $k > 0$, we consider the weak solution q_k of

(4.21)
$$\begin{cases} -\dfrac{\partial q_k}{\partial t} + A^* q_k + f'_y(y_k) q_k = \nu_k F'_y(x,t,y_k,u_k) + \mu_k|_Q + [\Phi'(y_k)^* \theta_k]|_Q + \lambda_k \xi_k \quad \text{in } Q, \\ \quad q_k = 0 \quad \text{on } \Sigma, \qquad q_k(T) = \nu_k L'_y(x, y_k(T)) + [\Phi'(y_k)^* \theta_k]|_{\overline{\Omega}_T} + \mu_k|_{\overline{\Omega}_T} \quad \text{in } \Omega, \end{cases}$$

where $\mu_k|_Q$ (resp., $[\Phi'(y_k)^*\theta_k]|_Q$) is the restriction of μ_k (resp., $[\Phi'(y_k)^*\theta_k]$) to Q, and $\mu_k|_{\overline{\Omega}_T}$ (resp., $[\Phi'(y_k)^*\theta_k]|_{\overline{\Omega}_T}$) is the restriction of μ_k (resp., $[\Phi'(y_k)^*\theta_k]$) to $\overline{\Omega} \times \{T\}$. By using the Green formula of Proposition 4.1 with z_k, we obtain

$$\int_Q \nu_k F'_y(x,t,y_k,u_k) z_k \, dx \, dt + \lambda_k \int_Q z_k(x,t) \xi_k(x,t) \, dx \, dt + \int_\Omega \nu_k L'_y(x, y_k(T)) z_k(T) \, dx$$
$$+ \langle \mu_k, z_k \rangle_{\overline{Q}} + \langle \theta_k, \Phi'(y_k) z_k \rangle_{\overline{Q}} = \int_Q q_k \left(u_{ok} - u_k + \xi_{ok} - \xi_k \right) \, dx \, dt.$$

With this equality, (4.20), and the definition of $\Delta_k J$, we get

$$\begin{aligned}&\int_Q [\nu_k F(x,t,y_k,u_k) + q_k u_k + q_k \xi_k + \lambda_k y_k \xi_k]\, ds\ dt\\ &\qquad \le \int_Q [\nu_k F(s,t,y_k,u_{ok}) + q_k u_{ok} + q_k \xi_{ok} + \lambda_k y_k \xi_{ok}]\, ds\, dt + \frac{1}{k^{2p}} \mathcal{L}^{n+1}(Q)\end{aligned} \tag{4.22}$$

for every $k > 0$ and every $(u_o, \xi_o) \in U_{ad} \times V_{ad}$ (where (u_{ok}, ξ_{ok}) is defined with respect to (u_o, ξ_o)).

Step 4. *Convergence of sequence $(\nu_k, \lambda_k, \mu_k, \theta_k, q_k)_k$; Pontryagin principle.* Observing that $\nu_k^2 + \lambda_k^2 + |\mu_k|^2_{\mathcal{M}(\overline{Q})} + |\theta_k|^2_{\mathcal{M}(\overline{Q})} = 1$, there exist $(\bar\nu, \bar\lambda, \bar\mu, \bar\theta) \in \mathbb{R}^+ \times \mathbb{R}^+ \times \mathcal{M}(\overline{Q}) \times \mathcal{M}(\overline{Q})$ and a subsequence, still denoted by $(\nu_k, \lambda_k, \mu_k, \theta_k)_k$, such that

$$\nu_k \to \bar\nu, \quad \lambda_k \to \bar\lambda, \quad \mu_k \rightharpoonup \bar\mu \text{ and } \ \theta_k \rightharpoonup \bar\theta \ \text{ weak}^* \text{ in } \mathcal{M}(\overline{Q}).$$

With the same arguments as in [21, Section 6.2, Step 4], we prove that $(q_k)_k$, or at least a subsequence of $(q_k)_k$, weakly converges to $\bar q$ in $L^{\delta'}(0,T; W_0^{1,d'}(\Omega))$ for every (δ, d) such that $\frac{n}{2d} + \frac{1}{\delta} < \frac{1}{2}$. Recall that $(u_k, \xi_k)_k$ converges to $(\bar u, \bar\xi)$ in $L^p(Q) \times L^p(Q)$. Hence y_k also converges to $\bar y$. Passing to the limit when k tends to infinity in (4.22) gives

$$\int_Q \big[H_1(x,t,\bar y, \bar u, \bar q, \bar\nu) + H_2(\bar y, \bar\xi, \bar q, \bar\lambda)\big]\, dx\ dt \le \int_Q \big[H_1(x,t,\bar y, u, \bar q, \bar\nu) + H_2(\bar y, \xi, \bar q, \bar\lambda)\big]\, dx\ dt$$

for every $(u, \xi) \in U_{ad} \times V_{ad}$. This inequality is equivalent to

$$\begin{aligned}&\int_Q H_1(x,t,\bar y(x,t), \bar u(x,t), \bar q(x,t), \bar\nu)\, dx\, dt\\ &\qquad = \min_{u \in U_{ad}} \int_Q H_1(x,t,\bar y(x,t), u(x,t), \bar q(x,t), \bar\nu)\, dx\, dt\end{aligned} \tag{4.23a}$$

$$\int_Q H_2(\bar y(x,t), \bar\xi(x,t), \bar q(x,t), \bar\lambda)\, dx\ dt = \min_{\xi \in V_{ad}} \int_Q H_2(\bar y(x,t), \xi(x,t), \bar q(x,t), \bar\lambda)\, dx\ dt. \tag{4.23b}$$

Now, by using Lebesgue's points argument (see [21, 24]), we obtain (4.10e) and (4.10f). On the other hand, it is clear that $\bar\nu \ge 0$. Moreover, from the definitions of μ_k and θ_k, we deduce

$$\langle \mu_k, z - y_k\rangle_{\overline{Q}} \le 0 \ \ \forall z \in \{z \in \mathcal{C}(\overline{Q}) \mid z \ge 0\} \quad \text{and } \langle \theta_k, z - \Phi(y_k)\rangle_{\overline{Q}} \le 0 \ \ \forall z \in \mathcal{C}. \tag{4.24}$$

When k tends to infinity, we obtain (4.10b) and a part of (4.10a). It remains to prove that $(\bar\nu, \bar\lambda, \bar\mu, \bar\theta)$ is nonzero; for this, we recall that $\nu_k^2 + \lambda_k^2 + |\mu_k|^2_{\mathcal{M}(\overline{Q})} + |\theta_k|^2_{\mathcal{M}(\overline{Q})} = 1$.

If $(\bar\nu, \bar\lambda) \neq 0$, then the proof is complete. If not, we can prove that $|\bar\mu|_{\mathcal{M}(\overline{Q})} + |\bar\theta|_{\mathcal{M}(\overline{Q})} > 0$.

First we recall that $\mathcal{C}$ has a finite codimension in $\mathcal{C}(\overline{Q})$ and that $\{z \in \mathcal{C}(\overline{Q}) \mid z \geq 0\}$ is a subset of $\mathcal{C}(\overline{Q})$ with a nonempty interior. Then $\widetilde{\mathcal{C}}$ is a subset of $\mathcal{C}(\overline{Q}) \times \mathcal{C}(\overline{Q})$ with a finite codimension. Moreover, from (4.24) we deduce that for every $(z_1, z_2) \in \widetilde{\mathcal{C}}$

$$\begin{aligned}\langle \mu_k, z_2 - \bar{y}\rangle_{\overline{Q}} + \langle \theta_k, z_1 - \Phi(\bar{y})\rangle_{\overline{Q}} &\leq \langle \mu_k, y_k - \bar{y}\rangle_{\overline{Q}} \\ +\langle \theta_k, \Phi(y_k) - \Phi(\bar{y})\rangle_{\overline{Q}} &\leq |y_k - \bar{y}|_{\mathcal{C}(\overline{Q})} + |\Phi(y_k) - \Phi(\bar{y})|_{\mathcal{C}(\overline{Q})}.\end{aligned}$$

The last right-hand side quantity tends to 0 as $k \to +\infty$. With this estimate and using $\lim_k |\mu_k|_{\mathcal{M}(\overline{Q})} + \lim_k |\theta_k|_{\mathcal{M}(\overline{Q})} = 1$, thanks to Lemma 3.6 of [16], we conclude that $(\bar{\mu}, \bar{\theta}) \neq 0$ when $(\bar{\nu}, \bar{\lambda}) = 0$. □

5. Examples. Let us consider the following optimal control problem where the cost functional is defined by

$$J(y, u) = \int_0^T [g(t, y(t)) + h(u(t))]\, dt + \psi(y(T)), \tag{5.1}$$

where

(A5*) the function $h : L^2(\Omega) \to \mathbb{R} \cup \{+\infty\}$ is convex and lower semicontinuous and there exist $c_1 > 0,\ c_2 \in \mathbb{R}$ such that

$$\forall u \in L^2(\Omega) \qquad h(u) \geq c_1 |u|^2_{L^2(\Omega)} - c_2, \tag{5.2}$$

(A6*) the function $g : [0, T] \times L^2(\Omega) \to \mathbb{R} \cup \{+\infty\}$ is measurable in t, $g(.,0) \in L^1(0, T)$, and for every $r > 0$ there exists $\gamma_r > 0$ independent of t such that

$$\begin{gathered}\forall t \in [0, T] \qquad |y|_{L^2(\Omega)} + |z|_{L^2(\Omega)} \leq r, \\ |g(t, y) - g(t, z)| + |\psi(y) - \psi(z)| \leq \gamma_r |y - z|_{L^2(\Omega)}.\end{gathered} \tag{5.3}$$

Conditions on g and ψ could be weakened. For more details one can refer to Barbu [2, p. 317].

Now we consider

$$(\mathcal{P}) \qquad \begin{cases}\text{Minimize } J(y(y_o, u), u), \\ u \in U_{ad}, \\ y(y_o, u) \text{ is the solution of (4.3),}\end{cases}$$

where U_{ad} is a nonempty, convex subset of $L^p(Q)$, closed for the $L^2(Q)$-topology, and p is an integer such that $n < p$. Although we are especially interested in optimality conditions for solutions of problem $(\mathcal{P})$, we give an existence result in the following theorem.

THEOREM 5.1. *For any $y_o \in K_o$ (defined by* (4.1)*), problem* $(\mathcal{P})$ *has at least one solution u. Moreover, the corresponding state belongs to $\mathcal{C}(\overline{Q}) \cap W^{2,1,p}(Q)$.*

Proof. One can find this result in Barbu [2, Proposition 1.1., p. 319] when $U_{ad} = L^2(Q)$. This is easily adapted to the case where U_{ad} is a closed convex subset of $L^2(Q)$. A priori estimations do not change so that we get the "suitable" convergence in the "suitable" spaces. The only modification concerns the cluster points of the control sequences. Because U_{ad} is convex and closed for the $L^2(Q)$-topology these points belong to U_{ad}. Because $U_{ad} \subset L^p(Q)$, we can use regularity results of Theorem 4.1. □

Remark 5.1. The assumption that U_{ad} has to be a convex subset of $L^p(Q)$ (for some $p > n$) closed for the $L^2(Q)$-topology may be difficult to ensure: for example

$U_{ad} = L^p(Q)$ is not suitable. However, we give more precise example sets U_{ad} in what follows. Let us refine the example. We set

$$J(y,u) = \frac{1}{2}\int_\Omega (y(x,T) - z_d(x))^2\,dx + \frac{N}{2}\int_Q u(x,t)^2\,dx\,dt \tag{5.4}$$

(with $N > 0$) so that with the previous notations we get

$$F(x,t,y,u) = \frac{N}{2}\,u^2,\ h(u(t)) = \frac{N}{2}\|u(t)\|^2_{L^2(\Omega)},$$
$$L(x,y) = \frac{1}{2}(y - z_d(x))^2,\ g(t,y(t)) \equiv 0,\ \psi(y(T)) = \frac{1}{2}\|y(T) - z_d\|^2_{L^2(\Omega)}.$$

It is easy to see that both (A5*) and (A6*) are fulfilled for such a choice of $h,\ g,\ \psi$. Therefore the optimal control problem

$$(\mathcal{P}_2)\qquad \begin{cases} \min\ J(y,u), \\ \dfrac{\partial y}{\partial t} + Ay + f(y) \geq u \ \text{ in } Q, \quad y = 0 \text{ on } \Sigma, \quad y(0) = y_o \text{ in } \Omega, \\ u \in U_{ad}, \\ y(x,t) \geq 0\ \forall (x,t) \in \overline{Q}, \end{cases}$$

where $y_o \in W^{1,p}_o(\Omega), y_o \geq 0,\ z_d \in L^2(\Omega)$, and U_{ad} is a nonempty, convex subset of $L^p(Q)$ closed for the $L^2(Q)$-topology, has an optimal solution.

We always assume, of course, that (A1) and (A2) are valid (one may choose $A = -\Delta$ for instance, where Δ is the Laplacian operator); we have already seen that (A3) and (A4) are fulfilled with the special choice of φ and y_o. It is also easy to see that (A5) and (A6) are ensured with F and L defined as above. Thus we may give optimality conditions for $(\mathcal{P}_2)$, as follows.

THEOREM 5.2. *Assume* (A1) *and* (A2) *are valid. Then problem* $(\mathcal{P}_2)$ *has an optimal solution* $(\bar y, \bar u) \in [W^{2,1,p}(Q) \cap \mathcal{C}(\overline{Q})] \times L^p(Q)$. *Moreover, there exist* $(\bar\nu, \bar\lambda, \bar\mu, \bar q) \in \mathbb{R} \times \mathbb{R} \times \mathcal{M}(\overline{Q}) \times L^1(0,T; W^{1,1}_o(\Omega))$ *such that the following optimality system holds:*

$$(\bar\nu, \bar\lambda, \bar\mu) \neq 0, \quad \bar\nu \geq 0, \tag{5.5a}$$

$$\forall z \in \{ z \in \mathcal{C}(\overline{Q}) \mid z \geq 0 \},\ \ \langle \bar\mu, z - \bar y\rangle_{\overline{Q}} \leq 0, \tag{5.5b}$$

$$\begin{cases} \dfrac{\partial \bar y}{\partial t} + A\bar y + f(\bar y) = \bar u + \bar\xi \quad \text{in } Q, \\ \bar y = 0 \quad \text{on } \Sigma, \qquad \bar y(0) = y_o \quad \text{in } \Omega, \end{cases} \tag{5.5c}$$

$$\bar y \geq 0,\ \bar\xi \in V_{ad},\ \bar u \in U_{ad}, \int_\Omega \bar y(t)\ \bar\xi(t)\ dx = 0 \quad \text{a.e. on } [0,T], \tag{5.5d}$$

$$\begin{cases} -\dfrac{\partial \bar q}{\partial t} + A^*\bar q + f'(\bar y)\bar q = \bar\mu_{|Q} + \bar\lambda\ \bar\xi \quad \text{in } Q, \\ \bar q = 0 \quad \text{on } \Sigma, \qquad \bar q(T) = \bar\nu[\bar y(T) - z_d] + \bar\mu_{|\overline{\Omega}_T} \quad \text{in } \Omega, \end{cases} \tag{5.5e}$$

$$([(\bar\nu N\bar u + \bar q)(u - \bar u)](x,t)) \le 0 \quad \forall\, u \in U_{ad}, \quad \textit{and a.e. } (x,t) \in Q, \tag{5.5f}$$

$$\bar q(x,t)\ \bar\xi(x,t) = 0 \quad \textit{a.e. } (x,t) \in Q, \tag{5.5g}$$

where $\bar\xi = \dfrac{\partial \bar y}{\partial t} + A\bar y + f(\bar y) - \bar u$.

Proof. This is a direct consequence of Theorem 4.2 where $\Phi = Id$ and $\mathcal{C}$ is the whole space. Considering the Hamiltonian functions and relations (4.10e) and (4.10f) give (5.5e) and (5.5f) immediately. □

We end this section with two examples for U_{ad}.

5.1. Case where U_{ad} is bounded in $L^\infty(Q)$.

Let us set

$$U_{ad} = \{\ u \in L^\infty(Q) \mid a(x,t) \le u(x,t) \le b(x,t) \ \text{ in } Q\ \},$$

where $a, b \in L^\infty(Q)$. U_{ad} is of course a convex subset of $L^p(Q)$ for any $p > n$. Moreover, we get the following lemma.

LEMMA 5.1. *U_{ad} is closed for the $L^2(Q)$-topology.*

Proof. Let $u_n \in U_{ad}$ converging to u in $L^2(Q)$. Then $u_n(x,t)$ converges to $u(x,t)$ a.e. in Q so that we get $a(x,t) \le u(x,t) \le b(x,t)$ a.e. in Q. Thus $u \in L^\infty(Q)$. It is clear that $u \in U_{ad}$. □

Therefore, in view of Remark 5.1, we get the result stated in the next theorem for $y_o = 0$ and

$$J(y,u) = \frac{1}{2}\int_\Omega (y(x,T) - z_d(x))^2\, dx + \frac{N}{2}\int_Q u^2(x,t)\, dx\ dt.$$

THEOREM 5.3. *Assume* (A1) *and* (A2) *are valid. Then problem* ($\mathcal{P}_2$) *has an optimal solution* $(\bar y, \bar u) \in [W^{2,1,p}(Q) \cap \mathcal{C}(\overline{Q})] \times L^p(Q)$ *for any* $p > n$. *Moreover there exists* $(\bar\nu, \bar\lambda, \bar\mu, \bar q) \in \mathbb{R} \times \mathbb{R} \times \mathcal{M}(\overline{Q}) \times L^1(0,T; W_o^{1,1}(\Omega))$ *such that* (5.5a)–(5.5d) *and* (5.5g) *hold with*

$$\begin{cases} -\dfrac{\partial \bar q}{\partial t} + A^*\bar q + f'(\bar y)\bar q = \bar\mu_{|Q} + \bar\lambda\ \bar\xi \quad \textit{in } Q, \\ \bar q = 0 \quad \textit{on } \Sigma, \qquad \bar q(T) = \bar\nu[\bar y(T) - z_d] + \bar\mu_{|\overline{\Omega}_T} \quad \textit{in } \Omega, \end{cases} \tag{5.6}$$

$$[(\bar\nu N\bar u + \bar q)(u - \bar u)](x,t) \le 0 \quad \forall\, u \in U_{ad} \quad \textit{and a.e. } (x,t) \in Q. \tag{5.7}$$

5.2. Case where $U_{ad} = \{u \in L^p(Q) \mid u(x,t) \ge 0$ a.e. in $Q\}$.

When $U_{ad} = \{u \in L^p(Q) \mid u(x,t) \ge 0$ a.e. in $Q\}$ and $y_o \ge 0$ in Ω, thanks to the maximum principle for parabolic equations, the constraint $y \ge 0$ is automatically fulfilled in (4.6b) so that the corresponding multiplier $\bar\mu$ is equal to 0 (or at least does not appear). Therefore the corresponding Pontryagin optimality system consists of (5.5a) and (5.5c)–(5.5g), where (5.5e) is replaced by

$$\begin{cases} -\dfrac{\partial \bar q}{\partial t} + A^*\bar q + f'(\bar y)\bar q = \bar\lambda\ \bar\xi \quad \text{in } Q, \\ \bar q = 0 \quad \text{on } \Sigma, \qquad \bar q(T) = \bar\nu[\bar y(T) - z_d] \quad \text{in } \Omega. \end{cases} \tag{5.8}$$

This implies in particular that $\bar q \in W^{2,1,p}(Q) \cap \mathcal{C}(\overline{Q})$.

For this simple example we can see that the optimality conditions (5.2) are not trivial because we cannot have $\bar\nu = \bar\lambda = 0$.

6. Conclusion. The optimality conditions we have obtained are given in a nonqualified form. So far it is difficult to compare precisely these results with those already existing, since they usually are in a qualified form [6, 5, 17] or they concern elliptic variational inequalities. Nevertheless we must emphasize that in this paper we obtain interesting informations about optimal solutions (at least in simple cases). Indeed, we have seen in Example 5 that (5.5e) provides precise information on the structure of the multipliers $\bar{\mu} + \bar{\xi}\ \bar{\lambda}$ for the distributed multiplier, for instance, and the adjoint state $\bar{q}$: the regular part of this adjoint state belongs to $\mathcal{C}(\overline{Q})$ while the nonsmooth part belongs to $L^1(0,T;W_o^{1,1}(\Omega))$. This information seems new, compared with that in Barbu [2, Section 5.1.4, p. 331], for example.

The method developed in [5, 23] for elliptic variational inequalities is still true for the parabolic case, but we think that this method does not allow the condition (4.11b) to be obtained. However, in [23, 5] the authors give a qualification assumption under which they can derive Pontryagin's principle in qualified form.

Since we now can preview the generic form of the Lagrange multipliers, we can check optimal control problems where the variational inequality is more general than the obstacle type or occurs on the boundary, with boundary control.

REFERENCES

[1] V. Barbu and T. Precupanu, *Convexity and Optimization in Banach Spaces*, Sijthoff and Noordhoff, Leyden, The Netherlands, 1978.

[2] V. Barbu, *Analysis and Control of Non Linear Infinite Dimensional Systems*, Math. Sci. Engrg. 190, Academic Press, New York, 1993.

[3] M. Bergounioux, *Optimal control of problems governed by abstract elliptic variational inequalities with state constraints,* SIAM J. Control Optim., 36 (1998), pp. 273–289.

[4] M. Bergounioux and F. Tröltzsch, *Optimality conditions and generalized bang-bang principle for a state constrained semilinear parabolic problem*, Numer. Funct. Anal. Optim., 15 (1996), pp. 517–537.

[5] F. Bonnans and E. Casas, *An extension of Pontryagin's principle for state constrained optimal control of semilinear elliptic equations and variational inequalities*, SIAM J. Control Optim., 33 (1995), pp. 274–298.

[6] J. F. Bonnans and D. Tiba, *Pontryagin's principle in the control of semilinear elliptic variational inequalities*, Appl. Math. Optim., 23 (1991), pp. 299–312.

[7] E. Casas, J.-P. Raymond, and H. Zidani, *Pontryagin's Principle for Local Solutions of Control Problems with Mixed Control-State Constraints*, MIP preprint, 1998.

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

[9] E. DiBenedetto, *Degenerate Parabolic Equations*, Springer-Verlag, New York, 1993.

[10] E. Di Benedetto, *On the local behavior of solutions of degenerate parabolic equations with measurable coefficients*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (13), (1986), pp. 487–535.

[11] J. Diestel, *Geometry of Banach Spaces—Selected Topics*, Lecture Notes in Math. 485, Springer-Verlag, Berlin, Heidelberg, New York, 1975.

[12] I. Ekeland, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.

[13] I. Ekeland and R. Temam, *Analyse convexe et problèmes variationnels*, Dunod-Gauthier-Villars, Paris, 1974.

[14] H. O. Fattorini and S. Sritharan, *Necessary and sufficient conditions for optimal controls in viscous flow problems*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 211–251.

[15] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Trans. Math. Monog. 23, AMS, Providence, RI, 1968.

[16] X.-J. and L.-J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser, Boston, Cambridge, MA, 1995.

[17] F. Mignot and J.-P. Puel, *Optimal Control in Some Variational Inequalities*, Math. Cont. Theo., 14 (1985), pp. 409–422.

[18] M. K. V. Murty and G. Stampacchia, *A variational inequality with mixed boundary conditions*, Israel J. Math., 13 (1972), pp. 188–224.

[19] J. P. Raymond, *Nonlinear boundary control of semilinear parabolic problems with pointwise state constraints*, Discontinuous Continuous Dynam. Systems, 9 (1997), pp. 341–370.

[20] J. P. Raymond and H. Zidani, *Hamiltonian Pontryagin's principles for control problems governed by semilinear parabolic equations*, Appl. Math. Optim., 39 (1999), pp. 143–177.

[21] J. P. Raymond and H. Zidani, *Pontryagin's principles for state-constrained control problems governed by parabolic equations with unbounded controls*, SIAM J. Control Optim., 36 (1998), pp. 1853–1871.

[22] D. Tiba, *Optimal Control of Nonsmooth Distributed Parameter Systems*, Lecture Notes in Math. 1459, Springer-Verlag, Berlin, 1990.

[23] J. Yong, *Pontryagin maximum principle for semilinear second order elliptic partial differential equations and variational inequalities with state constraints*, Differential Integral Equations, 5 (1992), pp. 1307–1334.

[24] H. Zidani, *Optimal Control Problem for Semilinear Parabolic Equations: Optimality Conditions and Numerical Approximations*, Thèse de doctorat, Université Paul Sabatier, Toulouse, France, 1996.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 4, pp. 1291–1309

RELAXATION OF CONTROL SYSTEMS UNDER STATE CONSTRAINTS*

HÉLÈNE FRANKOWSKA† AND FRANCO RAMPAZZO‡

Abstract. In this paper we provide a relaxation result for control systems under both equality and inequality constraints involving the state and the control. In particular, we show that the Mangasarian–Fromowitz constraint qualification allows us to rewrite constrained systems as differential inclusions with locally Lipschitz right-hand side. Then the Filippov–Ważewski relaxation theorem may be applied to show that ordinary solutions are dense in the set of relaxed solutions. If, besides agreeing with the above constraints, the state has to remain in a control-independent set K, then the Mangasarian–Fromowitz condition cannot hold. This case is investigated as well by means of a condition on the feasible velocities on the boundary of K.

Key words. control under state constraints, differential inclusion, relaxation

AMS subject classifications. 34A09, 34A60, 34H05, 49J15, 49J45, 49N35

PII. S0363012997331019

1. Introduction.

Let us consider a control system of the form

$$x'(t) = f(t, x(t), u(t)), \quad u(t) \in U, \tag{1}$$

where the time parameter t belongs to an interval $[t_0, T]$, the state x takes values in $\mathbf{R}^n$, U (the *control set*) is a subset of a complete separable metric space $\mathcal{Z}$, and the function f maps $[t_0, T] \times \mathbf{R}^n \times \mathcal{Z}$ into $\mathbf{R}^n$. The trajectory-control pairs are subject to the constraints

$$h(t, x(t), u(t)) \leq 0, \quad g(t, x(t), u(t)) = 0 \quad \text{almost everywhere (a.e.) in } [t_0, T], \tag{2}$$

where

$$h = (h_1, \dots, h_m) : [t_0, T] \times \mathbf{R}^n \times \mathcal{Z} \mapsto \mathbf{R}^m \quad \text{and} \quad g : [t_0, T] \times \mathbf{R}^n \times \mathcal{Z} \mapsto \mathbf{R}^p.$$

In this paper we study the closure of the solution set of the control system (1) under state constraints (2). A classical way to investigate this problem relies on the study of the feedback set-valued map

$$G(t, x) \; = \; \{u \in U \mid h(t, x, u) \leq 0, \;\; g(t, x, u) = 0\}$$

(see, for instance, [6, 7]). In fact, observe that a trajectory-control pair (x, u) solves (1), (2) if and only if

$$u(t) \in G(t, x(t)) \quad \text{a.e in } [t_0, T].$$

More generally, one can consider a set-valued map $G : [t_0, T] \times \mathbf{R}^n \mapsto U$ and the control system

$$\begin{cases} x'(t) = f(t, x(t), u(t)), \quad u(t) \in G(t, x(t)), \quad t \in [t_0, T], \\ x(t_0) = x_0. \end{cases} \tag{3}$$

*Received by the editors December 3, 1997; accepted for publication (in revised form) June 2, 1998; published electronically May 28, 1999. This research was supported by the Human Capital and Mobility Programme of the European Union, contract CHRX-CT94-0431.

http://www.siam.org/journals/sicon/37-4/33101.html

†Centre de Recherche Viabilité, Jeux, Contrôle, CNRS and Université de Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cx 16, France (frankows@viab.dauphine.fr).

‡Dipartimento di Matematica Pura e Applicata, Università di Padova, Via Belzoni 7, 35131 Padova, Italy (rampazzo@pdmat1.unipd.it).

In several dynamic optimization problems it is interesting to answer the following two questions:

$\mathcal{Q}1$. *Is the set of solutions to the system* (3) (in particular, to (1), (2)) *precompact in the space of continuous functions?*

$\mathcal{Q}2$. *What is its closure?*

While the answer to the first question comes from functional analysis (via a standard application of the Ascoli theorem), the answer to the second one can be provided in different ways. One can use for instance *generalized curves*, which were introduced within the context of calculus of variations by L. C. Young and extended to control systems by J. Warga [22] and E. J. McShane [16] (see also [23] and references therein). Generalized curves are defined as the trajectories corresponding to *relaxed controls.* To be more precise, let f satisfy some standard regularity assumptions (see section 3), and let $\mathcal{M}(U)$ denote the set of probability measures on U. $\mathcal{M}(U)$ can be regarded as a subset of $C(U)^*$, the dual of the space $C(U)$ of continuous real functions defined on U. Define the space of admissible relaxed controls as

$$\widehat{\mathcal{U}}^{adm}(t_0,T) = \{\mu \in L^\infty([t_0,T];\mathcal{M}(U)) | \text{ for a.e. } t \in [t_0,T],\ \operatorname{supp}(\mu_t) \subset G(t,x(t))\},$$

where $\operatorname{supp}(\mu_t)$ denotes the support of $\mu(t,\cdot)$, and $x(\cdot)$ is the unique continuous solution to the integral equation

$$x(t) = x(t_0) + \int_{t_0}^{t}\int_U f(s,x(s),u)\mu(s,du)ds.$$

It turns out that such a solution is absolutely continuous. It is called a *generalized curve* of (3). Under standard assumptions, if a sequence $x_n(\cdot)$ of solutions to (3) converges uniformly to some $x(\cdot)$, then $x(\cdot)$ is a generalized curve.

The second approach goes back to papers by L. Cesari [8], A. Lasota and C. Olech [14], and C. Olech [18] (see also the bibliography of [7] and [6]). It relies on Mazur's theorem and uses the convex hull of the right-hand side of (3). Absolutely continuous functions satisfying

$$x'(t) \in \overline{co}\, f(t,x(t),G(t,x(t))) \quad \text{a.e. in } [t_0,T]$$

are called *relaxed solutions of* (3). The above system is the relaxed (convexified) differential inclusion corresponding to (3) (of course, the meanings given here to expressions like *relaxed control*, *generalized curve*, and *relaxed solutions* are purely conventional). It is well known that under standard hypotheses the uniform limit of solutions to (3) is a relaxed trajectory. Actually, under general assumptions the set of generalized curves coincides with the set of relaxed trajectories (see also Theorem 3.4 below). Therefore the set of relaxed curves (or, equivalently, of generalized curves) contains the closure of the solutions to (3). In order to complete the answer to question $\mathcal{Q}2$, one has to solve the following problem:

$\mathcal{Q}3$. *Can we approximate every relaxed solution by "ordinary" trajectories of the control system* (3)*?*

A general answer to this question is contained in the celebrated Filippov–Ważewski theorem, in the framework of differential inclusions. For control systems, a partial answer is given, e.g., in [6, 7], where only the case when the set-valued map $G(\cdot,\cdot)$ is independent of the state is investigated.

E. N. Barron and J. Jensen have recently published a paper [5] concerning a related question. They studied the Mayer problem

$$\text{minimize } \psi(x(T))$$

over solutions to the constrained system (1), (2) satisfying $x(t_0) = x_0$ in the case when $g = 0$ (i.e., only inequality constraints are present). They were looking for an enlarged class of trajectories that should be considered in order for the above control problem to have a solution, and the infimum of the original problem was the same as the minimum of the relaxed problem. Let us remark that an answer to this problem does not necessarily imply a positive answer to question $\mathcal{Q}3$ (while, by virtue of the continuity of the map $x(\cdot) \mapsto \psi(x(T))$ the converse implication holds true), the latter being outside the declared goal of [5]. Actually our aim is to introduce a *strong extension* of the problem (see [21]), while the objective in [5] is to obtain a *weak extension* (see [9]). Incidentally, since our hypotheses are more general than those assumed in [5], our results show that the extension provided by E. N. Barron and J. Jensen is in fact strong. Under a rank condition recalled below and some technical assumptions, E. N. Barron and J. Jensen proposed the following answer to their problem.

Let $\widehat{\mathcal{U}}(t_0,T) = L^\infty([t_0,T];\mathcal{M}(U))$. The space of admissible relaxed controls is defined by

$$\widehat{\mathcal{U}}_h(t_0,T) = \left\{\mu \in \widehat{\mathcal{U}}(t_0,T) |\text{ for a.e. } t \in [t_0,T],\ \mu - \text{ess sup}_{u\in U}\ h(t,x(t),u) \le 0\right\},$$

where $x(\cdot)$ is the generalized curve corresponding to the relaxed control μ. The relaxed problem is

$$\text{minimize } \psi(x(T))$$

over all generalized curves satisfying $x(t_0) = x_0$. They showed that, under usual technical assumptions, the set of generalized curves of (1), (2) (with $g = 0$) starting at x_0 is compact, and so the relaxed problem has a solution. Next they proved that the value functions of the original and the relaxed problems coincide. For this aim they checked that these value functions solve the same Hamilton–Jacobi boundary value problem, which, in turn, has a unique solution.

Let us briefly recall the rank condition used by E. N. Barron and J. Jensen. Let $u \in G(t,x)$ and let us denote the set of all "active" indices by

$$I(t,x,u) = \left\{i_1 < \cdots < i_k \mid \forall\, 1 \le j \le k,\ h_{i_j}(t,x,u) = 0 \text{ and } h_q(t,x,u) \ne 0\, \forall\ q \ne i_j\right\}.$$

Define the map $h_{I(t,x,u)} = (h_{i_1},\ldots,h_{i_k})$. It is assumed in [5] that the rank of the matrix

$$\frac{\partial h_{I(t,x,u)}}{\partial u}(t,x,u)$$

is equal to the number of elements in $I(t,x,u)$ (to be more precise, the authors restrict their investigation to the case where the maps $h_1,\ldots,h_{m-1}$ do not depend on (t,x)). This rank condition is often met in the optimal control theory of the Bolza problem (see, for instance, [6, p. 36], where the author also considered the problem (1), (2) in the case when $g = 0$) but apparently there are no results in the literature answering in the whole generality the above-mentioned question $\mathcal{Q}3$, although a tool providing an answer has existed since 1962 in the form of the Filippov–Ważewski relaxation theorem for differential inclusions. Namely, to get a positive answer it is enough to assume that the set-valued map $x \rightsquigarrow f(t,x,G(t,x))$ is locally Lipschitz. On the other hand an example provided by Pliś [19] shows that the Lipschitz condition cannot be weakened to a continuity condition.

The first aim of this paper is to investigate questions $\mathcal{Q}1$–$\mathcal{Q}3$ in the presence of both equality and inequality constraints. We show that the Mangasarian–Fromowitz constraint qualification (of which the above rank condition is a particular case) yields the local Lipschitz continuity of the map $x \rightsquigarrow G(t,x)$. The crucial argument to achieve this result is a careful application of an inverse mapping theorem (extending a result from [11] for maps defined on a metric space. Once the Lipschitz continuity of $x \rightsquigarrow G(t,x)$ is established, if $f(t,\cdot,\cdot)$ is locally Lipschitz, the above mentioned Filippov–Ważewski theorem may be applied and question $\mathcal{Q}3$ may be positively answered.

The last part of the paper is devoted to the relaxation problem in a situation when the Mangasarian–Fromowitz constraint qualification is not verified, namely, when we add a state constraint of the form $k(x) \leq 0$ to the above constraints (2). Nevertheless we can provide a result concerning this case by assuming a generalization (provided in [17]) of a condition originally proposed—for relaxation purposes—by P. Loreti [15].

The outline of the paper is as follows. In section 2 we provide some preliminaries on set-valued inverse mapping theorems and differential inclusions. Section 3 is devoted to relaxation of (1), (2) under the Mangasarian–Fromowitz condition. Section 4 covers a case where this condition is violated. A concluding short section provides a simple application concerning the value function of the Mayer problem.

2. Preliminaries.

2.1. Inverse mapping theorems. This section is basically an adaptation of some results from [11] to the form we need in this paper. We denote by $B_\delta(x)$ and $B(x,\delta)$ the closed ball of center x and radius δ; by $\overset{o}{B}$ and B the open and closed unit balls, respectively; by Graph(G) the graph of the set-valued map $G : X \rightsquigarrow Y$,

$$\mathrm{Graph}(G) = \{(x,y) \in X \times Y \mid y \in G(x)\}\,;$$

and by Dom(G) its domain

$$\mathrm{Dom}(G) = \{x \in X \mid G(x) \neq \emptyset\}\,.$$

Observe the following simple fact.

Proposition 2.1. *Let φ be a map from a metric space X to a normed space Y and $\rho > 0$, $\varepsilon > 0$, $\delta > 0$, $\overline{x} \in X$ be given. Then, the following two statements are equivalent:*

i. *φ satisfies the uniform open mapping principle at $\overline{x}$: $\forall\ x \in B_\delta(\overline{x})$ and $0 \leq h \leq \varepsilon$, $\varphi(x) + \rho h \overset{o}{B} \subset\ \varphi(B_h(x))$.*

ii. *The inverse map φ^{-1} is pseudo-Lipschitz at $\varphi(\overline{x})$: $\forall\ x \in B_\delta(\overline{x})$ and $y \in Y$ satisfying $\|y - \varphi(x)\| < \rho\varepsilon$,*

$$\mathrm{dist}\left(x,\ \varphi^{-1}(y)\right) \leq \frac{1}{\rho}\ \|\varphi(x) - y\|\,.$$

Pseudo-Lipschitz maps were introduced in [1]. The proof of Proposition 2.1 is straightforward. Most recently they have been called Aubin continuous maps.

Definition 2.2 (see [11]). *The* upper variation *of $\varphi : X \mapsto Y$ at x is the closed subset of Y defined by*

$$\varphi^{\sharp 1}(x) = \mathrm{Limsup}_{h\to 0+}\ \frac{\varphi(B_h(x)) - \varphi(x)}{h}.$$

In other words $v \in \varphi^{\sharp 1}(x)$ if and only if there exist sequences $h_i \to 0+$, $v_i \to v$ such that $\varphi(x) + h_i v_i \in \varphi(B_{h_i}(x))$.

Let X be a Banach space, $K \subset X$, and $x \in K$. The *contingent cone* to K at x is defined by

$$T_K(x) = \mathrm{Limsup}_{h \to 0+} \frac{K - x}{h}$$

(see, for instance, [4, Chapter 4]).

Let X, Y be Banach spaces and $\varphi : X \mapsto Y$ be a Fréchet differentiable map (or locally Lipschitz and Gâteaux differentiable map). Consider a subset $K \subset X$, $\overline{x} \in K$, and let $\varphi_{|K}$ denote the restriction of φ to K. Then it is not difficult to check that

$$\varphi'(\overline{x})(T_K(\overline{x}) \cap B) \subset \left(\varphi_{|K}\right)^{\sharp 1}(\overline{x}), \tag{4}$$

so that

$$\varphi'(\overline{x})(\overline{co}\ (T_K(\overline{x}) \cap B)) \subset \overline{co}\ \left(\varphi_{|K}\right)^{\sharp 1}(\overline{x})$$

and, in particular,

$$\varphi'(\overline{x})(B) \subset \varphi^{\sharp 1}(\overline{x}). \tag{5}$$

THEOREM 2.3. *Consider a complete metric space (X, d_X), a Banach space Y with the norm Gâteaux differentiable away from zero and a continuous map $\varphi : X \mapsto Y$. Let $\overline{x} \in X$ and assume that for some $\varepsilon > 0$, $\rho > 0$*

$$\rho B \subset \bigcap_{x \in B_\varepsilon(\overline{x})} \overline{co}\, \varphi^{\sharp 1}(x).$$

Then for every $x \in B_{\varepsilon/2}(\overline{x})$ and $y \in Y$ satisfying $\|y - \varphi(x)\| < \rho\varepsilon/2$,

$$\mathrm{dist}\ \left(x,\ \varphi^{-1}(y)\right) \leq \frac{1}{\rho}\ \|\varphi(x) - y\|\,.$$

In order to prove Theorem 2.3 we shall make use of Ekeland's variational principle (see, e.g., [3, 4]), which we recall for the reader's convenience.

THEOREM 2.4. *Let (X, d) be a complete metric space and $f : X \mapsto \mathbf{R} \cup \{+\infty\}$ be an extended lower semicontinuous bounded from below function. Consider $x_0 \in X$ such that $f(x_0) \neq +\infty$ and $\varepsilon > 0$. Then there exists $\overline{x} \in X$, a solution to*

$$\begin{cases} \text{i.} & f(\overline{x}) + \varepsilon d(x_0, \overline{x}) \leq f(x_0), \\ \text{ii.} & \forall\, x \neq \overline{x},\ \ f(\overline{x}) < f(x) + \varepsilon d(x, \overline{x}). \end{cases}$$

Proof of Theorem 2.3. By Proposition 2.1 we have to show that for every $x_1 \in B_{\varepsilon/2}(\overline{x})$, $0 \leq h \leq \varepsilon/2$

$$\varphi(x_1) + \rho h \overset{o}{B} \subset \varphi(B_h(x_1)).$$

We fix such x_1, h and pick $y \in Y$ satisfying $\|y - \varphi(x_1)\| < \rho h$. We define $0 < \theta < 1$ by

$$\theta^2 = \|y - \varphi(x_1)\| / \rho h.$$

Applying Theorem 2.4 to the complete metric space $B_h(x_1)$ and the continuous function $x \mapsto \|y - \varphi(x)\|$, we prove the existence of $x_2 \in B_{\theta h}(x_1) \subset B_\varepsilon(\overline{x})$ such that

$$\forall\, x \in B_h(x_1),\ \ \|\varphi(x_2) - y\| \leq \|\varphi(x) - y\| + \theta\rho d_X(x, x_2). \tag{6}$$

It remains to show that y is equal to $\varphi(x_2)$. Indeed, assume for a moment that $y \neq \varphi(x_2)$. By differentiability of the norm, there exists $p \in Y^\star$ of $\|p\|_\star = 1$ such that $\forall h > 0,\ v \in Y$

$$\|\varphi(x_2) + hv - y\| = \|\varphi(x_2) - y\| + \langle p, hv\rangle + o_v(h), \tag{7}$$

where $\lim_{h\to 0+} o_v(h)/h = 0$. We fix $v \in \varphi^{\sharp 1}(x_2)$ and let $h_j \to 0+,\ v_j \to v$ be such that $\varphi(x_2) + h_j v_j \in \varphi(B_{h_j}(x_2))$. Then from (6) and (7) we obtain

$$0 \leq \langle p, h_j v\rangle + h_j \|v_j - v\| + \theta\rho h_j + o_v(h_j).$$

Dividing by h_j and taking the limit yields $\langle p, v\rangle \geq -\theta\rho$. Since $v \in \varphi^{\sharp 1}(x_2)$ is arbitrary we proved that

$$\forall\, v \in \overline{co}\left(\varphi^{\sharp 1}(x_2)\right), \quad \langle p, v\rangle \geq -\theta\rho. \tag{8}$$

By the assumption of the theorem, $\rho B \subset \overline{co}\left(\varphi^{\sharp 1}(x_2)\right)$. Thus (8) yields $-\rho \geq -\theta\rho$. Since $0 < \theta < 1$ and $\rho > 0$, we derived a contradiction. Consequently $\varphi(x_2) = y$. □

2.2. Relaxation of differential inclusions. Consider a set-valued map F from $[t_0, T] \times \mathbf{R}^n$ into subsets of $\mathbf{R}^n$ and the differential inclusion

$$x' \in F(t, x). \tag{9}$$

An absolutely continuous function $x : [t_0, T] \mapsto \mathbf{R}^n$ is called a *solution* to (9) if

$$x'(t) \in F(t, x(t)) \ \text{ a.e. in } \ [t_0, T]. \tag{10}$$

Denote by $\mathcal{S}_{[t_0,T]}(x_0)$ the set of solutions to (9) starting at $x_0 \in \mathbf{R}^n$ at time t_0 and defined on the time interval $[t_0, T]$. Let us recall a result concerning the density of the set of solutions to the differential inclusion

$$\left\{ \begin{array}{lcl} x'(t) & \in & F(t, x(t)) \ \text{ a.e. in } \ [t_0, T], \\ x(t_0) & = & x_0 \end{array} \right. \tag{11}$$

into the set of solutions to the convexified (relaxed) differential inclusion

$$\left\{ \begin{array}{lcl} x'(t) & \in & \overline{co}\, F(t, x(t)) \ \text{ a.e. in } \ [t_0, T], \\ x(t_0) & = & x_0. \end{array} \right. \tag{12}$$

Assume

$$\left\{ \begin{array}{ll} \text{i.} & \forall\, (t, x) \in [t_0, T] \times \mathbf{R}^n, \ \ F(t, x) \ \text{ is closed,} \\ \text{ii.} & \forall\, x \in \mathbf{R}^n, \ F(\cdot, x) \text{ is measurable,} \\ \text{iii.} & \exists\, \gamma \in L^1(t_0, T), \ \text{for a.e. } t \in [t_0, T], \forall\, x \in \mathbf{R}^n, \ \sup_{v \in F(t,x)} \|v\| \leq \gamma(t)(1 + \|x\|). \end{array} \right. \tag{13}$$

Observe that if F satisfies (13), then so does the set-valued map $(t, x) \rightsquigarrow \overline{co}\,(F(t, x))$. If in addition for almost every t, $F(t, \cdot)$ is upper semicontinuous, then it is well known, thanks to the Dunford–Pettis criterion, the Mazur theorem, and the convergence theorem (see, for instance, [4, 6, 7]) that under assumptions (13) the set of solutions to (12) is compact in the space of continuous functions $C(t_0, T)$. This and the Filippov–Ważewski theorem (see, for instance, [2] for the time-independent case, [4, p. 402], or [12]) yield the following well-known result.

THEOREM 2.5. *Suppose* (13) *and that*

$$\left\{ \begin{array}{l} \forall\, \rho > 0, \exists\, k \in L^1(t_0, T) \ \textit{ such that for almost all } \ t \in [t_0, T], \\ \textit{the map } \ F(t, \cdot) \ \textit{ is } \ k(t) - \textit{Lipschitz on } \ B_\rho(0). \end{array} \right. \tag{14}$$

Then the closure of $\mathcal{S}_{[t_0,T]}(x_0)$ *in* $C(t_0,T)$ *is compact and equal to the set of solutions to the relaxed inclusion* (12).

Remark. The convergence theorem can be found, e.g., in [4] for time-independent maps. However it can be extended with exactly the same proof to all measurable with respect to time set-valued maps. □

2.3. Control systems with state-dependent control sets. Let $\mathcal{Z}$ be a complete separable metric space, let $G : [t_0,T] \times \mathbf{R}^n \rightsquigarrow \mathcal{Z}$ be a given set-valued map, and consider the control system

$$x' = f(t,x,u), \quad u \in G(t,x), \quad t \in [t_0,T]. \tag{15}$$

Such systems were considered, e.g., in [6, 7]. An absolutely continuous function $x : [t_0,T] \mapsto \mathbf{R}^n$ is called a solution to (15) if for some measurable selection $u(t) \in G(t,x(t))$ we have

$$x'(t) = f(t,x(t),u(t)) \text{ a.e. in } [t_0,T].$$

We introduce the set-valued map $F : [t_0,T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ defined by

$$F(t,x) = f(t,x,G(t,x)) = \{f(t,x,v) \mid v \in G(t,x)\},$$

and replace (15) by the differential inclusion (10). We impose the following assumptions:

$$\begin{cases} \forall\,(x,u) \in \mathbf{R}^n \times \mathcal{Z}, \ f(\cdot,x,u) \text{ is measurable,} \\ \forall\, t \in [t_0,T], \ f(t,\cdot,\cdot) \text{ is continuous,} \\ G \text{ is upper semicontinuous and has closed nonempty images.} \end{cases} \tag{16}$$

Observe that if in addition $G(t,x)$ are compact, then F has compact images.

From the measurable selection theorem (see, for instance, [4, Chapter 8]), Theorem 2.6 follows.

THEOREM 2.6. *If* (16) *holds true, then the sets of solutions to control system* (15) *and differential inclusion* (10) *do coincide.*

Thanks to the above result Theorem 2.5 can be applied to the control system (15) whenever the map $x \rightsquigarrow f(t,x,G(t,x))$ satisfies the required Lipschitz property. This holds, for example, whenever $G(\cdot,\cdot)$ is locally Lipschitz and $f(t,\cdot,\cdot)$ is $k(t)$-Lipschitz for some $k \in L^1$. On the other hand it is well known that if the Lipschitz property fails, then the conclusion of Theorem 2.5 may fail as well. A corresponding example was provided by Pliś in 1962.

Example (derived from [19, Pliś]). We recall first the Pliś construction.

Consider the map $F : \mathbf{R}^2 \rightsquigarrow \mathbf{R}^2$

$$F(x_1,x_2) = \left\{(v_1,v_2) \mid v_1 \in \{-1,+1\}\,,\ v_2 = \sqrt{|x_2|} + |x_1|\right\}.$$

This map is continuous but not locally Lipschitz around zero. For every $x \in \mathcal{S}_{[0,T]}(0)$ solving (11) for $t_0 = 0$ and $x_0 = 0$ we have $x_2(t) \geq t^2/4$, while $x \equiv 0$ solves the relaxed inclusion (12). Thus the conclusion of Theorem 2.5 is not valid in this case (trajectories of the original inclusion are not dense in the trajectories of the convexified one).

Consider a C^∞-map $g : \mathbf{R}^2 \times \mathbf{R}^2 \mapsto \mathbf{R}$ such that $g(x,u) = 0$ if and only if $u \in F(x)$. Such a map exists since the set Graph(F) is closed. Let $f(t,x,u) = u$ and $G(t,x) = \{u \mid g(x,u) = 0\}$. Then $F(x) = f(t,x,G(t,x))$. This implies that solutions to

$$x'(t) = f(t,x(t),u(t)), \quad g(x(t),u(t)) = 0, \quad x(0) = 0$$

are not dense in the solutions to

$$x'(t) \in \overline{co}\, f(t, x, G(t, x)), \;\; x(0) = 0.$$

3. Solutions of control systems under constraints. In this section we study relaxation of control systems under constraints. Speaking of the *closure* of the solution's set of a given Cauchy problem we always mean the closure in the space of continuous functions $C(t_0, T)$ endowed with the supremum norm. Consider a complete separable metric space $\mathcal{Z}$, real numbers $t_0 < T$, a map (describing the dynamics)

$$f : [t_0, T] \times \mathbf{R}^n \times \mathcal{Z} \mapsto \mathbf{R}^n,$$

and the control set $U \subset \mathcal{Z}$. We associate with this data the control system

$$x'(t) = f(t, x(t), u(t)), \quad u(t) \in U. \tag{17}$$

Consider the maps defining state-control constraints

$$h = (h_1, \dots, h_m) : [t_0, T] \times \mathbf{R}^n \times \mathcal{Z} \mapsto \mathbf{R}^m \;\text{ and }\; g : [t_0, T] \times \mathbf{R}^n \times \mathcal{Z} \mapsto \mathbf{R}^p.$$

We investigate solutions to (17) under additional constraints

$$h(t, x(t), u(t)) \le 0, \quad g(t, x(t), u(t)) = 0 \quad \text{a.e. in } [t_0, T]. \tag{18}$$

A way to study the above problem is to introduce the set-valued map

$$G(t, x) = \{u \in U \mid h(t, x, u) \le 0, \;\; g(t, x, u) = 0\}, \tag{19}$$

where $h(t, x, u) \le 0$ means $h_i(t, x, u) \le 0$ for every $i = 1, \dots, m$.

Observe that if a trajectory-control pair (x, u) solves (17), (18), then

$$u(t) \in G(t, x(t)) \quad \text{a.e. in } [t_0, T]. \tag{20}$$

Naturally, the converse statement holds true as well, if (x, u) satisfies (17) and (20), then (18) is satisfied.

Another type of problem that will be considered in this section is the one with equality constraints independent of controls. Let $\varphi : [t_0, T] \times \mathbf{R}^n \mapsto \mathbf{R}^p$. The state constraints are given by

$$\varphi(t, x(t)) = 0, \;\; h(t, x(t), u(t)) \le 0. \tag{21}$$

Observe that if (x, u) solves (17), (21), and φ is differentiable, then for almost all $t \in [t_0, T]$ we have

$$\frac{\partial \varphi}{\partial t}(t, x(t)) + \frac{\partial \varphi}{\partial x}(t, x(t)) f(t, x(t), u(t)) = 0.$$

Setting

$$G_1(t, x) = \left\{ u \in U \mid \frac{\partial \varphi}{\partial t}(t, x) + \frac{\partial \varphi}{\partial x}(t, x) f(t, x, u) = 0, \; h(t, x, u) \le 0 \right\}$$

we deduce that $u(t) \in G_1(t, x(t))$ a.e.

PROPOSITION 3.1. *Assume that φ is differentiable and locally Lipschitz and let (x, u) be a trajectory-control pair of* (17) *such that $u(t) \in G_1(t, x(t))$ a.e. If $\varphi(t_0, x(t_0)) = 0$, then (x, u) satisfies the state-control constraints* (21). *Conversely, every trajectory-control pair (x, u) to* (17), (21) *verifies $u(t) \in G_1(t, x(t))$ a.e.*

In order to prove this proposition it is enough to differentiate the map $t \mapsto \varphi(t, x(t))$.

Observe that the above result reduces the constrained system (17), (21) to the constrained system (17), (18) with g given by

$$g(t,x,u) = \frac{\partial \varphi}{\partial t}(t,x) + \frac{\partial \varphi}{\partial x}(t,x) f(t,x,u). \tag{22}$$

We shall next provide sufficient conditions for Lipschitz continuity of the map $x \rightsquigarrow G(t,x)$. Recall that in view of the Example of section 2 such a property is needed to build a satisfactory relaxation theory.

3.1. Control-dependent constraints. In this section $\mathcal{Z}$ is a Banach space and $U = \mathcal{Z}$. Set

$$\mathbf{R}^k_+ = \{(\alpha_1, \ldots, \alpha_k) |\ \alpha_i \geq 0\ \forall\, i = 1, \ldots, k\},$$

and consider the set-valued map $G : [t_0, T] \times \mathbf{R}^n \rightsquigarrow \mathcal{Z}$ defined by (19). With every (t,x,u) we associate the set of all "active" indices

$$I(t,x,u) = \left\{ i_1 < \cdots < i_k \mid \forall\, 1 \leq j \leq k,\ h_{i_j}(t,x,u) = 0 \ \text{ and } \ h_q(t,x,u) \neq 0\, \forall\, q \neq i_j \right\},$$

and define

$$h_{I(t,x,u)}(s,y,v) = \left(h_{i_1}, \ldots, h_{i_k}\right)(s,y,v) \quad \forall\, (s,y,v) \in [t_0,T] \times \mathbf{R}^n \times \mathcal{Z}.$$

Assume that the constraints satisfy the Mangasarian–Fromowitz-type conditions:

$$\begin{cases} \frac{\partial h}{\partial u},\ \frac{\partial g}{\partial u} \ \text{ are continuous,} \\ \forall\, u \in G(t,x),\ \frac{\partial g}{\partial u}(t,x,u) \ \text{ is surjective,} \\ \forall\, u \in G(t,x),\ \exists\, v \in U,\ \ \frac{\partial h_{I(t,x,u)}}{\partial u}(t,x,u)(v) < 0,\ \ \frac{\partial g}{\partial u}(t,x,u)(v) = 0. \end{cases} \tag{23}$$

Remark. In [5] a relaxation problem is considered where only inequality constraints act on the system. The main assumption made in [5] is a rank condition, which is stronger than (23). In fact, assume that there is no g in (18) and, $\forall u \in G(t,x)$, denote by $k(t,x,u)$ the number of elements in $I(t,x,u)$. The rank condition states that the rank of the matrix

$$\frac{\partial h_{I(t,x,u)}}{\partial u}(t,x,u)$$

is equal to $k(t,x,u)$. This means that the linear operator

$$\frac{\partial h_{I(t,x,u)}}{\partial u}(t,x,u) : U \to \mathbf{R}^{k(t,x,u)}$$

is surjective. Hence there exists $v \in U$ as in (23). The converse statement in general does not hold. □

We also assume that

$$\begin{cases} & \forall\, (\bar t, \bar x) \in [t_0,T] \times \mathbf{R}^n,\ \exists\, r > 0,\ \varepsilon > 0,\ L > 0\ \forall\, (t,x) \in B_\varepsilon(\bar t, \bar x), \\ \text{i.} & \{u \in U \mid h(t,x,u) \leq r,\ \|g(t,x,u)\| \leq r\} \quad \text{is compact,} \\ \text{ii.} & \forall\, u \in U,\ (h,g)(\cdot,\cdot,u) \text{ is L-Lipschitz on } B_\varepsilon(\bar t, \bar x), \\ \text{iii.} & \forall \text{ bounded } K \subset U,\ \exists\, M > 0 \text{ such that } \forall\, (t,x) \in B_\varepsilon(\bar t, \bar x), \\ & h(t,x,\cdot) \ \text{ is M-Lipschitz on } \ K. \end{cases} \tag{24}$$

We claim that G is upper semicontinuous (see, for instance, [2] for the definition). By (24.i) it has compact images. Clearly its graph is closed. We check next that its domain is closed as well. Indeed if $(t,x) \in \overline{\mathrm{Dom}(G)}$, then there exist $(t_i,x_i) \to (t,x)$ and $u_i \in G(t_i,x_i)$. Because of assumption (24.ii) $\forall i$ large enough

$$\|g(t,x,u_i)\| = \|g(t,x,u_i) - g(t_i,x_i,u_i)\| \le r,$$

$$h(t,x,u_i) \le h(t,x,u_i) - h(t_i,x_i,u_i) \le r.$$

We deduce from (24.i) that the sequence u_i has an accumulation point u. Then $h(t,x,u) \le 0,\ g(t,x,u) = 0$ and therefore (t,x) is in the domain of G.

We next fix $(\bar t,\bar x) \in \mathrm{Dom}(G)$. By (24.ii), for some $\delta > 0$ and $\forall (t,x) \in B_\delta(\bar t,\bar x) \cap \mathrm{Dom}(G)$ we have

$$G(t,x) \subset \{u \in U \mid h(\bar t,\bar x,u) \le r,\ \ \|g(\bar t,\bar x,u)\| \le r\}\,.$$

Since the set-valued map G has a closed graph and since $\forall (t,x) \in B_\delta(\bar t,\bar x) \cap \mathrm{Dom}(G)$ it takes its values in a compact set, we deduce that G is upper semicontinuous at $(\bar t,\bar x)$. The domain of G being closed and the point $(\bar t,\bar x) \in \mathrm{Dom}(G)$ being arbitrary, we conclude that G is upper semicontinuous.

THEOREM 3.2. *Assume that* (23), (24) *hold true and that* $\mathrm{Dom}(G) \neq \emptyset$. *Then* $\mathrm{Dom}(G) = [t_0,T] \times \mathbf{R}^n$ *and G is locally Lipschitz on* $[t_0,T] \times \mathbf{R}^n$.

Proof. We first observe that assumption (23) yields

$$\forall\, u \in G(t,x),\ \ \mathrm{Im}\left(\frac{\partial h_{I(t,x,u)}}{\partial u}, \frac{\partial g}{\partial u}\right)(t,x,u) + \mathbf{R}^k_+ \times 0 = \mathbf{R}^{k+p}, \tag{25}$$

where k denotes the number of elements in $I(t,x,u)$. It is enough to show that G has nonempty values and is locally Lipschitz continuous on a neighborhood of every point of its domain. In particular, since we already know that this domain is closed, we deduce that it is equal to $[t_0,T] \times \mathbf{R}^n$.

We fix $(\bar t,\bar x) \in \mathrm{Dom}(G)$, and consider the compact set

$$Z(\bar t,\bar x) = \{u \in G(\bar t,\bar x) \mid I(\bar t,\bar x,u) \neq \emptyset\} \subset G(\bar t,\bar x).$$

Let $\varepsilon, r > 0, L > 0$ be as in assumption (24) and $\bar u \in Z(\bar t,\bar x)$. Consider $\rho(\bar u) > 0$ such that

$$4\rho(\bar u)B \subset \left(\frac{\partial h_{I(\bar t,\bar x,\bar u)}}{\partial u}, \frac{\partial g}{\partial u}\right)(\bar t,\bar x,\bar u)(B) + \big(\mathbf{R}^k_+ \cap B\big) \times \{0\}\,.$$

It exists because of (25) and the separation theorem. Since $\frac{\partial h}{\partial u}$ and $\frac{\partial g}{\partial u}$ are continuous, there exists $\varepsilon > \delta(\bar u) > 0$ such that

$$\begin{cases} \forall\,(t,x) \in B_{\delta(\bar u)}(\bar t,\bar x)\ \ \forall\, u \in B_{\delta(\bar u)}(\bar u), \\ 2\rho(\bar u)B \subset \left(\frac{\partial h_{I(\bar t,\bar x,\bar u)}}{\partial u}, \frac{\partial g}{\partial u}\right)(t,x,u)(B) + \big(\mathbf{R}^k_+ \cap B\big) \times \{0\}\,. \end{cases}$$

By continuity of h and g, for some $0 < \nu(\bar u) \le \delta(\bar u)$ we have

$$\begin{cases} \forall\,(t,x) \in B_{\nu(\bar u)}(\bar t,\bar x)\ \forall\, u \in B_{\nu(\bar u)}(\bar u), \\ \big\|h_{I(\bar t,\bar x,\bar u)}(t,x,u)\big\| + \|g(t,x,u)\| < \rho(\bar u)\delta(\bar u)/2, \\ \forall\, i \notin I(\bar t,\bar x,\bar u),\ h_i(t,x,u) < 0. \end{cases} \tag{26}$$

Since $Z(\bar{t},\bar{x})$ is compact, there exist $u_i \in Z(\bar{t},\bar{x}),\ i=1,\dots,s$, satisfying

$$Z(\bar{t},\bar{x}) \subset \mathcal{O} := \bigcup_{i=1}^{s} \overset{o}{B}(u_i, \nu(u_i)/8).$$

Furthermore, for some $\mu > 0$ and all $u \in G(\bar{t},\bar{x}) \setminus \mathcal{O}, h(\bar{t},\bar{x},u) \le -2\mu$. The last inequality means that $h_j(\bar{t},\bar{x},u) \le -2\mu\, \forall j$.

Using that $\frac{\partial g}{\partial u}$ is continuous, by (25) we can find $\hat{\varepsilon} > 0$ and $\rho_0 > 0$ such that

$$\forall\ (t,x) \in B_{\hat{\varepsilon}}(\bar{t},\bar{x})\ \forall\ u \in G(\bar{t},\bar{x}) + \hat{\varepsilon}B,\ \ \rho_0 B \subset \frac{\partial g}{\partial u}(t,x,u)(B).$$

Moreover, for some $0 < \bar{\varepsilon} < \hat{\varepsilon}$

$$\forall\ (t,x) \in B_{\bar{\varepsilon}}(\bar{t},\bar{x})\ \forall\ u \in G(\bar{t},\bar{x}) + \bar{\varepsilon}B,\ \ \|g(t,x,u)\| < \rho_0\hat{\varepsilon}/2.$$

Set

$$\bar{\rho} = \min\left\{\rho_0,\ \rho(u_i) \mid i=1,\dots,s\right\},\quad \bar{\nu} = \min\left\{\bar{\varepsilon},\ \nu(u_i) \mid i=1,\dots,s\right\},$$

and let M be a Lipschitz constant (independent of $(t,x) \in B_{\varepsilon}(\bar{t},\bar{x})$) of $h(t,x,\cdot)$ on $G(\bar{t},\bar{x}) + \hat{\varepsilon}B$, which exists by (24.iii)). Since G is upper semicontinuous and h is continuous, we can find $0 < \delta < \frac{\bar{\nu}}{8}$ such that

$$\forall\ (t,x) \in B_{\delta}(\bar{t},\bar{x}),\ \ G(t,x) \subset G(\bar{t},\bar{x}) + B(0,\bar{\nu}/8)$$

and

$$2L\delta < \bar{\rho}\hat{\varepsilon}/2,\ \ L\delta < \bar{\rho}\bar{\nu},\ \ L\delta(1+M/\bar{\rho}) < \mu$$

$$\forall\ u \in G(t,x) \setminus \mathcal{O},\ \ h(t,x,u) \le -\mu.$$

Consider $(t_1,x_1),\ (t_2,x_2) \in B_\delta(\bar{t},\bar{x})$, and $y_1 \in G(t_1,x_1)$, i.e., $g(t_1,x_1,y_1) = 0$. If $y_1 \in G(t_1,x_1) \setminus \mathcal{O}$, then $h(t_1,x_1,y_1) \le -\mu$. From (5) and Theorem 2.3 applied to the map $\varphi = g(t_2,x_2,\cdot),\ \varepsilon = \hat{\varepsilon}/2,\ y=0,\ \bar{x} = y_1$, using that $G(t_2,x_2)$ is compact, we deduce that there exists $y_2 \in U$ such that $g(t_2,x_2,y_2) = 0$ and

$$\|y_2 - y_1\| \le \frac{1}{\bar{\rho}}\|g(t_2,x_2,y_1)\| \le \frac{L}{\bar{\rho}}\left(|t_2-t_1| + \|x_2-x_1\|\right).$$

Therefore

$$h(t_2,x_2,y_2) \le h(t_1,x_1,y_1) + L\delta\left(1 + \frac{M}{\bar{\rho}}\right) \le -\mu + \mu = 0$$

and $y_2 \in G(t_2,x_2)$. Consider next the case $-\mu < h(t_1,x_1,y_1) \le 0$. Then $y_1 \in G(t_1,x_1) \cap \mathcal{O}$. Thus for some i

$$y_1 \in B(u_i,\nu(u_i)/8) + B(0,\nu(u_i)/8) = B(u_i,\nu(u_i)/4),$$

where $u_i \in Z(\bar{t},\bar{x})$. Furthermore

$$\left\|h_{I(\bar{t},\bar{x},u_i)}(t_2,x_2,y_1)\right\| + \|g(t_2,x_2,y_1)\| < \rho(u_i)\delta(u_i)/2.$$

Let k denote the number of elements in $I(\bar{t},\bar{x},u_i)$. By applying Theorem 2.3 to the function $\varphi : U \times \mathbf{R}_+^k \mapsto \mathbf{R}^{k+p}$ defined by

$$\varphi(z,\alpha) := (h_{I(\bar{t},\bar{x},u_i)}, g)(t_2,x_2,z) + (\alpha,0)$$

with

$$\varepsilon = \delta(u_i),\ y = (h_{I(\bar t,\bar x,u_i)}(t_1,x_1,y_1),0),\ \bar x = u_i,\ x = (y_1,0),$$

we deduce from (4) that there exist $y_2 \in U$ and $\alpha \in \mathbf{R}^k_+$ such that

$$(h_{I(\bar t,\bar x,u_i)},g)(t_2,x_2,y_2) = (h_{I(\bar t,\bar x,u_i)}(t_1,x_1,y_1) - \alpha,0)$$

and

$$\begin{aligned}&\|y_2 - y_1\|\\ &\le \rho(u_i)^{-1}\left(\left\|h_{I(\bar t,\bar x,u_i)}(t_2,x_2,y_1) - h_{I(\bar t,\bar x,u_i)}(t_1,x_1,y_1)\right\| + \|g(t_2,x_2,y_1)\|\right)\\ &\le 2L(\bar\rho)^{-1}\left(|t_2-t_1| + \|x_2-x_1\|\right).\end{aligned}$$

In particular, $y_2 \in G(t_2,x_2)$; therefore $G(\cdot,\cdot)$ is $\frac{2L}{\bar\rho}$-Lipschitz on $B_\delta(\bar t,\bar x)$. □

We impose the following assumptions on f:

$$(27)\quad \begin{cases}\text{i.} & \forall\, x,u,\ f(\cdot,x,u) \text{ is measurable,}\\ \text{ii.} & \forall\,\rho>0,\ \exists\, k\in L^1(t_0,T) \text{ such that for almost all } t\in[t_0,T],\\ & \text{the map } f(t,\cdot,\cdot) \text{ is } k(t)\text{-Lipschitz on } B_\rho(0)\times U,\\ \text{iii.} & \exists\,\gamma\in L^1(t_0,T), \text{ such that for a.e. } t\in[t_0,T],\\ & \forall\, x\in\mathbf{R}^n,\ \sup_{u\in U}\|f(t,x,u)\| \le \gamma(t)(1+\|x\|).\end{cases}$$

Consider the case $\mathrm{Dom}(G)\neq\emptyset$. Then, by Theorem 3.2, the sets $F(t,x) = f(t,x,G(t,x))$ are nonempty and compact, the set-valued map F is measurable in t and locally Lipschitz in x in the sense of (14). By subsection 2.3 the control system (15) may be rewritten as the differential inclusion (10).

THEOREM 3.3. *Under assumptions* (23), (24), *and* (27) *the closure of solutions to* (17), (18) *starting at x_0 is compact and equal to the set of solutions to the differential inclusion*

$$\begin{cases} x' & \in & \overline{co}\, f(t,x,G(t,x)),\\ x(t_0) & = & x_0.\end{cases}$$

Notice that if $\mathrm{Dom}(G)=\emptyset$, then the above result holds true. When $\mathrm{Dom}(G)\neq\emptyset$ it follows from Theorems 2.5 and 3.2.

For control problems without constraints one of the approaches to the relaxation theory uses probability measures (i.e., relaxed controls see, e.g., [23]). It has also been used by E. N. Barron and J. Jensen [5] for the problem with inequality constraints. Actually, Theorem 3.3 can be reformulated in terms of relaxed controls, so that it implies, in particular, the result from [5].

More precisely, let $\mathcal{M}(U)$ denote the set of probability measures on U and set

$$\widehat{\mathcal{U}}(t_0,T) = L^\infty([t_0,T];\mathcal{M}(U)).$$

The set $\mathcal{M}(U)$ is endowed with the weak star topology of $\mathcal{C}(U)^*$, the dual of the space $\mathcal{C}(U)$ of continuous real functions defined on U. Generalizing the definition given in [5] to the case involving equality constraints as well, we define the space of admissible relaxed controls as

$$\begin{aligned}\widehat{\mathcal{U}}_{hg}(t_0,T) \ :=\ &\{\mu\in\widehat{\mathcal{U}}(t_0,T) \mid \text{for almost all } t\in[t_0,T],\\ &\mu - \text{ess}\sup_{u\in U} h(t,x(t),u)\le 0,\ \ \mu - \text{ess}\sup_{u\in U}\|g(t,x(t),u)\| = 0\},\end{aligned}$$

where $x(\cdot)$ is the (unique) continuous solution of the integral equation

$$x(t) = x(t_0) + \int_{t_0}^{t} \int_U f(s, x(s), u)\mu(s, du)ds.$$

$x(\cdot)$ is called the *generalized curve of* (17) *corresponding to the relaxed control* μ. From assumptions (27) it follows that x is absolutely continuous.

COROLLARY 3.4. *Under assumptions* (23), (24), (27) *the closure of solutions to* (17), (18) *starting at* x_0 *is compact and equal to the set of generalized curves of the control system* (17) *starting at* x_0 *and corresponding to the admissible relaxed controls* $\widehat{\mathcal{U}}_{hg}(t_0, T)$.

Proof. This result is an immediate consequence of Theorem 3.3. We provide its proof for the sake of completeness. By the very definition of admissible relaxed controls,

$$\mu\left(t, \{u \in U \mid h(t, x(t), u) > 0 \quad \text{or} \quad g(t, x, u) \neq 0\}\right) = 0$$

for almost all $t \in [t_0, T]$. Thus, for almost all $s \in [t_0, T]$,

$$\int_U f(s, x(s), u)\mu(s, du) = \int_{G(s,x(s))} f(s, x(s), u)\mu(s, du).$$

Let $p \in \mathbf{R}^n,\ t \in [t_0, T]$. Then

$$\langle p, x(t) - x(t_0)\rangle \leq \int_{t_0}^{t} \max_{u \in G(s,x(s))} \langle p, f(s, x(s), u)\rangle\, ds$$

and consequently, from the separation theorem,

$$x(t) - x(t_0) \in \int_{t_0}^{t} \overline{co}\, f(s, x(s), G(s, x(s)))ds.$$

Because x is absolutely continuous, we deduce from the mean value theorem (see [2, p. 21]) that $x'(s) \in \overline{co}\, f(s, x(s), G(s, x(s)))$ a.e.

Conversely, consider a solution x to the differential inclusion

$$x'(t) \in \overline{co}\, f(t, x(t), G(t, x(t))), \quad x(t_0) = x_0.$$

There exists then measurable functions $\lambda_i : [t_0, T] \mapsto [0, 1]$ and measurable selections $v_i(s) \in f(s, x(s), G(s, x(s))),\ i = 0, \dots, n$, such that

$$\sum_{i=0}^{n} \lambda_i(s) = 1,\ \sum_{i=0}^{n} \lambda_i(s)v_i(s) = x'(s) \quad \text{a.e. in} \quad [t_0, T]$$

(see, for instance, [4, Chapter 8]). Consider measurable selections $u_i(s) \in G(s, x(s))$ such that $\forall i = 0, \dots, n,\ v_i(s) = f(s, x(s), u_i(s))$ and set

$$\mu(s, u) = \begin{cases} \lambda_i(s) & \text{if } u = u_i(s),\ i = 0, \dots, n, \\ 0 & \text{otherwise.} \end{cases}$$

Then $\mu \in \widehat{\mathcal{U}}_{hg}(t_0, T)$ and x is the corresponding generalized curve. □

3.2. Control-independent equality constraints. By the results from the beginning of section 3, when the equality constraints are control independent, the problem can be reduced to the one considered in subsection 3.1. We translate the results of subsection 3.1 for such situation. For the sake of simplicity we assume here that there are no inequality constraints, which is equivalent to setting $h = -1$. It is straightforward to extend the following considerations to the case where actual inequality constraints $h(t, x, u) \leq 0$ (and even supplementary equality constraints $g(t, x, u) = 0$) are considered as well.

Consider a continuously differentiable $\varphi : [t_0, T] \times \mathbf{R}^n \mapsto \mathbf{R}^p$ and let $\mathcal{Z}$ be a Banach space, $U = \mathcal{Z}$. We introduce the set-valued map

$$G(t,x) = \left\{ u \in U \mid \frac{\partial \varphi}{\partial t}(t,x) + \frac{\partial \varphi}{\partial x}(t,x) f(t,x,u) = 0 \right\}. \tag{28}$$

Thanks to Proposition 3.1 we know that the system (17) subject to the constraint

$$\varphi(t, x(t)) = 0 \tag{29}$$

is equivalent to (17), (18) with g defined by (22) and $h = -1$.

Thus we may apply the results of subsection 3.1. For this aim assume that

$$\left\{ \begin{array}{l} (\frac{\partial \varphi}{\partial t}, \frac{\partial \varphi}{\partial x}) \;\text{ is locally Lipschitz,} \\ f \;\text{ is differentiable with respect to } u, \; \frac{\partial f}{\partial u} \;\text{ is continuous, and} \\ \forall\, u \in G(t,x), \quad \frac{\partial \varphi}{\partial x}(t,x) \frac{\partial f}{\partial u}(t,x,u) \;\text{ is surjective,} \end{array} \right. \tag{30}$$

$$\left\{ \begin{array}{l} \forall\, (\bar t, \bar x) \in [t_0, T] \times \mathbf{R}^n, \exists\, \varepsilon > 0, \; r > 0 \; \forall\, (t,x) \in B_\varepsilon(\bar t, \bar x), \text{ such that the set} \\ \{u \in U \mid \| \frac{\partial \varphi}{\partial x}(t,x) f(t,x,u) \| \leq r - \| \frac{\partial \varphi}{\partial t}(t,x) \| \} \text{ is compact,} \\ \exists\, L > 0 \; \forall\, u \in U, \; f(\cdot, \cdot, u) \;\text{ is L-Lipschitz on } B_\varepsilon(\bar t, \bar x). \end{array} \right. \tag{31}$$

Under these conditions, by Theorem 3.2, the set-valued map $G(\cdot, \cdot)$ is locally Lipschitz and so we have similar relaxation theorems. Applying Theorem 3.3 with $h = -1$ we get the following result.

THEOREM 3.5. *Under assumptions* (27), (30), (31) *the closure of solutions to* (17), (29) *starting at x_0 is compact and equal to the set of solutions to the differential inclusion*

$$\left\{ \begin{array}{rcl} x' & \in & \overline{co}\, f(t, x, G(t,x)), \\ x(t_0) & = & x_0. \end{array} \right.$$

Define the set of admissible relaxed controls $\widehat{\mathcal{U}}_\varphi(t_0, T) = \widehat{\mathcal{U}}_{hg}(t_0, T)$ as in subsection 3.1 with $h = -1$ and g as above. Corollary 3.4 yields the following.

COROLLARY 3.6. *Under assumptions* (27), (30), (31) *the closure of solutions to* (17), (29) *starting at x_0 is compact and equal to the set of generalized curves of the control system* (17) *starting at x_0 and corresponding to the admissible relaxed controls $\widehat{\mathcal{U}}_\varphi(t_0, T)$.*

3.3. Constrained control sets. In this section we assume that U is a closed subset of a finite-dimensional space $\mathcal{Z} = \mathbf{R}^m$. For every $u \in U$, let $C_U(u)$ denote the Clarke tangent cone to U at u. It is well known (see, for instance, [4, Chapter 4]) that

$$\operatorname{Liminf}_{u' \to_U u} T_U(u') = C_U(u) \tag{32}$$

and that $C_U(u)$ is convex.

First, we consider the constrained control system (17), (18). Define $G(t,x)$ by (19) and assume that (24) holds true. The surjectivity assumption (23) has to be replaced by the following one:

$$(33)\begin{cases} \frac{\partial h}{\partial u},\ \frac{\partial g}{\partial u} \text{ are continuous and } \forall\, u \in G(t,x),\ \frac{\partial g}{\partial u}(t,x,u)(C_U(u)) = \mathbf{R}^p \\ \forall\, u \in G(t,x), \exists\, v \in C_U(u),\ \ \frac{\partial h_{I(t,x,u)}}{\partial u}(t,x,u)(v) < 0,\ \ \frac{\partial g}{\partial u}(t,x,u)(v) = 0. \end{cases}$$

Exactly in the same way as in subsection 3.1 we check that G is upper semicontinuous, has compact images, and that its domain is closed.

PROPOSITION 3.7. *Under assumptions* (24) and (33), $G(\cdot,\cdot)$ *is locally Lipschitz on* $[t_0,T]\times\mathbf{R}^n$.

Proof. Observe that (33) and the separation theorem yield

$$(34)\qquad \left(\frac{\partial h_{I(t,x,u)}}{\partial u}, \frac{\partial g}{\partial u}\right)(t,x,u)(C_U(u)) + \mathbf{R}^k_+ \times 0 = \mathbf{R}^{k+p},$$

where k is the number of elements in $I(t,x,u)$. Let $Z(\bar t,\bar x)$ have the same meaning as in the proof of Theorem 3.2, and fix $\bar u \in Z(\bar t,\bar x)$. Using the separation theorem it is not difficult to show that (34) implies the existence of $\rho(\bar u) > 0$, $v_i \in C_U(\bar u)$, $i = 1,\ldots,s \le m+1$, such that $\|v_i\| \le 1$ and

$$8\rho(\bar u)B \subset \left(\frac{\partial h_{I(\bar t,\bar x,\bar u)}}{\partial u}, \frac{\partial g}{\partial u}\right)(\bar t,\bar x,\bar u)\left(\overline{co}\,\{v_1,\ldots,v_s\}\right) + (\mathbf{R}^k_+ \cap B, 0).$$

From (32) we deduce that for some $\delta(\bar u) > 0$ and $\forall u \in U \cap B_{\delta(\bar u)}(\bar u)$

$$4\rho(\bar u)B \subset \left(\frac{\partial h_{I(\bar t,\bar x,\bar u)}}{\partial u}, \frac{\partial g}{\partial u}\right)(\bar t,\bar x,\bar u)\left(\overline{co}\,(T_U(u)\cap B)\right) + (\mathbf{R}^k_+ \cap B, 0).$$

By continuity of the derivative, we can choose $\delta(\bar u)$ in such a way that $\forall (t,x) \in B_{\delta(\bar u)}(\bar t,\bar x)$ and $u \in G(t,x)\cap B_{\delta(\bar u)}(\bar u)$

$$2\rho(\bar u)B \subset \left(\frac{\partial h_{I(\bar t,\bar x,\bar u)}}{\partial u}, \frac{\partial g}{\partial u}\right)(t,x,u)\left(\overline{co}\,(T_U(u)\cap B)\right) + (\mathbf{R}^k_+ \cap B, 0).$$

Using (4) and applying exactly the same arguments as those in the proof of Theorem 3.2 we conclude. □

Arguing as in Theorem 3.3 we deduce the following theorem.

THEOREM 3.8. *Under assumptions* (24), (27), and (33) *the closure of solutions to* (17), (18) *starting at* x_0 *is compact and equal to the set of solutions to the differential inclusion*

$$\begin{cases} x' &\in& \overline{co}\, f(t,x,G(t,x)), \\ x(t_0) &=& x_0. \end{cases}$$

Observe that Corollary 3.4 is still valid in this new situation. Define the set of admissible relaxed controls $\widehat{\mathcal{U}}_{hg}(t_0,T)$ as in subsection 3.1 with U as above.

COROLLARY 3.9. *Under assumptions* (24), (27), *and* (33) *the closure of solutions to* (17), (18) *starting at* x_0 *is compact and equal to the set of generalized curves of the control system* (17) *starting at* x_0 *and corresponding to the admissible relaxed controls* $\widehat{\mathcal{U}}_{hg}(t_0,T)$.

Consider next equality constraints (29) and the set-valued map (28). Assume that

$$\left\{\begin{array}{l} (\frac{\partial\varphi}{\partial t},\frac{\partial\varphi}{\partial x}) \text{ is locally Lipschitz,} \\ f \text{ is differentiable with respect to } u,\ \frac{\partial f}{\partial u} \text{ is continuous, and} \\ \forall\, u\in G(t,x),\quad \frac{\partial\varphi}{\partial x}(t,x)\frac{\partial f}{\partial u}(t,x,u)(C_U(u)) = \mathbf{R}^p. \end{array}\right. \tag{35}$$

Define g by (22). From Theorem 3.8 we get the following theorem.

THEOREM 3.10. *Under assumptions* (27), (31), *and* (35), *the closure of solutions to* (17), (29) *starting at* x_0 *is compact and equal to the set of solutions to the differential inclusion*

$$\left\{\begin{array}{rcl} x' & \in & \overline{co}\, f(t,x,G(t,x)), \\ x(t_0) & = & x_0. \end{array}\right.$$

Define the set of admissible relaxed controls $\widehat{\mathcal{U}}_\varphi(t_0,T)=\mathcal{U}_{hg}(t_0,T)$ as in subsection 3.1 with U as above, $h=-1$, and g given by (22). Then we have the corollary below.

COROLLARY 3.11. *Under assumptions* (27), (31), *and* (35), *the closure of solutions to* (17), (29) *starting at* x_0 *is compact and equal to the set of generalized curves of the control system* (17) *starting at* x_0 *and corresponding to the admissible relaxed controls* $\widehat{\mathcal{U}}_\varphi(t_0,T)$.

4. Control-independent inequality constraints. In this section we subject the control-trajectories pairs to the constraints

$$x\in\overline{\Theta},\quad h(t,x,u)\le 0,\quad g(t,x,u)=0, \tag{36}$$

where $\Theta\subseteq\mathbf{R}^n$ is an open subset and $\overline{\Theta}$ denotes its closure (in view of subsection 3.2 the case of a constraint of the form $\varphi(t,x)=0$ is already covered by this formulation). Observe that Θ is control independent. For instance, it may be given by

$$\overline{\Theta}:=\{x\in\mathbf{R}^n\ :\ k(x)\le 0\},$$

with $k\in C(\mathbf{R}^n)$. Since $\frac{\partial k}{\partial u}(x)=0\ \forall x\in\mathbf{R}^n$, we cannot apply any longer the techniques of section 3. To relax (17), (36) we use a result from [17] on the existence of admissible trajectories which keep a controlled distance from a given (generally not admissible) trajectory. Consider a set-valued map F from $[t_0,T]\times\mathbf{R}^n$ into the subsets of $\mathbf{R}^n$, and let F verify hypotheses (13) and (14). For each subset $A\subseteq\mathbf{R}^n$ and each constant $\rho\ge 0$, let $B(A,\rho)$ denote the set $\{y\in\mathbf{R}^n\mid d(y,A)\le\rho\}$. Let us consider assumption (H) below on the behavior of F near $\partial\Theta$.

Hypothesis (H).

1. $\Theta=\Theta_1\cap\Theta_2$, where Θ_1 and Θ_2 are open subsets verifying the following conditions, respectively:

2. there exist constants η,q,r and a continuous selection $\alpha(t,x)\in F(t,x)$ defined on $[t_0,T]\times B(\partial\Theta_1,\eta)$ such that at any $(t,x)\in[t_0,T]\times(\Theta_1\cap B(\partial\Theta_1,\eta))$ one has

$$B(x+s\alpha(t,x),sr)\subset\Theta_1$$

$\forall s\in]0,q]$;

3. $\forall(t,x)\in[t_0,T]\times\partial\Theta_2$ one has $0\in F(t,x)$; moreover, there is $\delta>0$ such that $\forall(t,x)\in[t_0,T]\times\partial\Theta_2$ there exists a closed proper cone V and $\varepsilon>0$ such that

$$(x+V_\varepsilon)\cap B(x,\delta)\cap\overline{\Theta}_2=\{x\}\quad\text{and}\quad F(s,y)\subseteq V$$

$\forall (s,y) \in B((t,x);\delta) \cap ([t_0,T] \times \overline{\Theta}_2)$, where V_ε denotes the ε-conical neighborhood of V; i.e.,

$$V_\varepsilon := \left\{ \lambda \left(\frac{v}{\|v\|} + B(0,\varepsilon) \right), \quad \lambda > 0, \quad v \in V \backslash \{0\} \right\}.$$

Remark. Condition H1 is a generalization of a condition originally introduced by M. Soner [20]. For a control system with fixed control set and for a smooth Θ he used such a condition to prove the continuity of the value function of an infinite horizon problem. Later the condition was adapted to nonsmooth constraint sets by H. Ishii and S. Koike [13], while the generalization to differential inclusion is provided in [10]. In the same framework as Soner's, condition H2 was used by P. Loreti [15] for relaxation purposes as well. Finally, condition H3 has been introduced in [17]. □

THEOREM 4.1 (see [17]). *Let the map F be locally Lipschitz (in both variables) with compact values and let hypothesis* (H) *be in force. Then, for any compact subset $Q \subseteq \overline{\Theta}$ there exists a positive constant C such that $\forall x_0 \in Q$, and for any solution $x(\cdot)$ of the Cauchy problem*

$$x'(t) \in F(t,x(t)), \qquad x(t_0) = x_0$$

(*possibly violating the constraint $x(t) \in \overline{\Theta}$*) *there is a solution $z(\cdot)$ of the same Cauchy problem such that*

$$z(t) \in \overline{\Theta} \quad \forall\, t \in [t_0,T] \quad \textit{and} \quad \sup_{t\in[t_0,T]} \|x(t) - z(t)\| \le C \sup_{t\in[t_0,T]} d(x(t),\Theta).$$

As a corollary of Theorem 4.1 we obtain the following analogue of the Filippov–Ważewski theorem for differential inclusion with state constraints. Denote by $\mathcal{S}^{\overline{\Theta}}_{[t_0,T]}(x_0)$ the set of solutions to the differential inclusion (11) that verify the state constraint $x(t) \in \overline{\Theta}$ for every $t \in [t_0,T]$.

THEOREM 4.2. *Let the map F be locally Lipschitz with compact values and let hypothesis* (H) *be in force. Then the closure of $\mathcal{S}^{\overline{\Theta}}_{[t_0,T]}(x_0)$ is compact and equal to the set of solutions to the relaxed inclusion* (12) *that verify the constraint $x(t) \in \overline{\Theta}$ for every $t \in [t_0,T]$.*

Remark. By assuming a smoothness hypothesis on $\partial\Theta$, it is possible to replace the Lipschitz condition on F with the weaker conditions (13), (14) (see [17]). Similarly, when the boundary is smooth, the Lipschitz hypothesis on f in all theorems below can be replaced by the weaker conditions (27). □

Proof. Since $\mathcal{S}^{\overline{\Theta}}_{[t_0,T]}(x_0) \subseteq \mathcal{S}_{[t_0,T]}(x_0)$, from Theorem 2.5 it follows that the closure of $\mathcal{S}^{\overline{\Theta}}_{[t_0,T]}(x_0)$ is compact and is contained in the set of the solutions to the relaxed inclusion (12) that verify the constraint $x(t) \in \overline{\Theta}$ for every $t \in [t_0,T]$. Let us prove the converse inclusion. We may assume that the constant C appearing in Theorem 4.1 is greater than or equal to 1. Let us choose an $\varepsilon > 0$ and a trajectory $y(\cdot)$ of the relaxed inclusion (12) satisfying the constraint $y(t) \in \overline{\Theta}$. By the Filippov–Ważewski theorem (Theorem 2.5) one can find a trajectory $x(\cdot)$ of the differential inclusion (11) such that $\|x(t) - y(t)\| \le \varepsilon/2C$ for every $t \in [t_0,T]$. Observe that in general $x(\cdot)$ violates the constraints. However, in view of Theorem 4.1 a trajectory $z \in \mathcal{S}^{\overline{\Theta}}_{[t_0,T]}(x_0)$ can be found that verifies $\|z(t) - x(t)\| \le Cd$ for every $t \in [t_0,T]$, with $d := \sup\{d(x(t),\Theta) | t \in [t_0,T]\}$. Since $d \le \|x(t) - y(t)\|$ for all $t \in [t_0,T]$, we obtain that $\|z(t) - y(t)\| \le \varepsilon/2 + \varepsilon/2C \le \varepsilon$, so concluding the proof. □

In particular, we obtain the theorem below.

THEOREM 4.3. *Let us assume* (24), (33) *and that for every compact* $K \subset \mathbf{R}^n$*, there exists a constant* $L > 0$ *such that* $\forall u \in U$, $f(\cdot, \cdot, u)$ *is L-Lipschitz on* $[t_0, T] \times K$*, and let* $F(t,x) := f(t, x, G(t,x))$ *have compact values and verify hypothesis* (H)*. Then the closure of solutions to* (17), (36) *is compact and equal to the set of solutions to*

$$\begin{cases} x' & \in & \overline{co}\, f(t, x, G(t,x)), \\ x(t_0) & = & x_0, \ \ x(t) \in \overline{\Theta}. \end{cases}$$

Of course, by means of arguments quite similar to those exploited in the proof of Corollary 3.4 it is possible to reformulate the above theorem in terms of relaxed controls. Let us define the set $\widehat{\mathcal{U}}_{hg}^{\overline{\Theta}}(t_0, T)$ of relaxed controls by setting

$$\widehat{\mathcal{U}}_{hg}^{\overline{\Theta}}(t_0, T) := \{\mu \in \widehat{\mathcal{U}}(t_0, T) \mid x([t_0, T]) \subset \overline{\Theta}, \text{and, for almost all } \ t \in [t_0, T], \\ \mu - \text{ess sup}_{u \in U} \ h(t, x(t), u) \leq 0, \ \ \mu - \text{ess sup}_{u \in U} \ \|g(t, x(t), u)\| = 0\},$$

where x is a generalized curve of (17) corresponding to the relaxed control μ.

COROLLARY 4.4. *Under all the assumptions of Theorem* 4.3 *the closure of solutions to* (17), (36) *is compact and equal to the set of generalized curves of the control system* (17) *starting at* x_0 *and corresponding to the admissible relaxed controls* $\widehat{\mathcal{U}}_{hg}^{\overline{\Theta}}(t_0, T)$.

5. Applications. As an application of the results of the previous sections one finds that under suitable hypotheses the value function of the Mayer problem coincides with the value function of the relaxed Mayer problem. This was in fact the main motivation of the paper by E. N. Barron and J. Jensen [5].

Let us consider a function $\psi : \mathbf{R}^n \mapsto \mathbf{R}$ and the constrained control system

$$x'(t) = f(t, x(t), u(t)), \quad u(t) \in U, \tag{37}$$

$$h(t, x(t), u(t)) \leq 0, \quad g(t, x(t), u(t)) = 0 \quad \text{a.e. in } \ [t_0, T], \tag{38}$$

$$x(t) \in \overline{\Theta} \quad \text{in } \ [t_0, T], \tag{39}$$

where $f : [0, T] \times \mathbf{R}^n \times \mathcal{Z} \mapsto \mathbf{R}^n$ and h, g, $\mathcal{Z}$, U, Θ are as in Theorem 4.3. We consider the following minimization problem:

$$V(t_0, x_0) := \inf \ \{\psi(x(T)) \mid x \ \text{ solves } \ (37), \ (38), \ (39), \ x(t_0) = x_0\}.$$

The map V is called the *value function.* Consider the set-valued map $G(\cdot, \cdot)$ defined by (19) and the relaxed differential inclusion

$$x'(t) \in \overline{co}\, f(t, x, G(t,x)). \tag{40}$$

The value function of the relaxed problem is defined by

$$V^{co}(t_0, x_0) := \inf \ \{\psi(x(T)) \mid x \ \text{solves } \ (39), \ (40), \ x(t_0) = x_0\}. \tag{41}$$

THEOREM 5.1. *Let* ψ *be continuous and let the hypothesis of Theorem* 4.3 *be verified. Then the infimum in* (41) *is attained and* $V = V^{co}$*. Moreover,* V *is continuous.*

Proof. The existence of an optimal trajectory for the relaxed problem (41) is a standard result, which can be obtained via the so-called direct method (of course, for this purpose it is sufficient to assume that ψ is lower bounded and lower semicontinuous). As for the equality $V = V^{co}$, it is a straightforward consequence of Theorem 4.3 and of the continuity of ψ. The continuity of V was proved in [17]. □

Obviously, in view of the results of section 4, this theorem can be reformulated in terms of relaxed controls as well.

REFERENCES

[1] J.-P. Aubin, *Lipschitz behavior of solutions to convex minimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.

[2] J.-P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

[3] J.-P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, Wiley Interscience, New York, 1984.

[4] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, Basel, Berlin, 1990.

[5] E. N. Barron and J. Jensen, *Relaxation of Constrained Control Systems*, SIAM J. Control Optim., 34 (1996), pp. 2077–2091.

[6] L. J. Berkovitz, *Optimal Control Theory*, Springer-Verlag, New York, 1974.

[7] L. Cesari, *Optimization Theory and Applications*, Springer-Verlag, New York, 1983.

[8] L. Cesari, *Closure, lower closure, and semicontinuity theorems in optimal control*, SIAM J. Control Optim., 9 (1971), pp. 287–315.

[9] A.L. Dontchev and T. Zolezzi, *Well-Posed Optimization Problems*, Lecture Notes in Math. 1543, Springer-Verlag, 1993.

[10] F. Forcellini and F. Rampazzo, *On nonconvex differential inclusions whose state is constrained in the closure of an open set. Applications to dynamic programming*, J. Differ. Integral Equations, to appear.

[11] H. Frankowska, *Some inverse mapping theorems*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 183–234.

[12] H. Frankowska, *A priori estimates for operational differential inclusions*, J. Differential Equations, 84 (1990), pp. 100–128.

[13] H. Ishii and S. Koike, *A new formulation of state constraint problems for first-order PDEs*, SIAM J. Control Optim., 34 (1996), pp. 554–571.

[14] A. Lasota and C. Olech, (1966) *On the closedness of the set of trajectories of a control system*, Bull. Acad. Polon. Sci. Ser. SCI. Math. Astronom. Phys., 14 (1966), pp. 615–621.

[15] P. Loreti, *Some properties of constrained viscosity solutions of Hamilton–Jacobi–Bellman equation*, SIAM J. Control Optim., 25 (1987), pp. 1244–1252.

[16] E. J. McShane, *Relaxed controls and variational problems*, SIAM J. Control Optim., 5 (1967), pp. 438–485.

[17] M. Motta and F. Rampazzo, *A sufficient condition for the continuity of the value function of control problems with state constraints*, NoDEA, to appear.

[18] C. Olech, (1969) *Existence theorems for optimal control problems with vector-valued cost functions*, Trans. Amer. Math. Soc., 136 (1969), pp. 157–180.

[19] A. Pliś, *Trajectories and quasi-trajectories of an orientor field*, Bull. Acad. Pol. Sc., 13 (1962), pp. 565–569.

[20] H. M. Soner, *Optimal control problems with state–space constraints* (I), SIAM J. Control Optim., 24 (1987), pp. 551–561.

[21] V. M. Tichomirov, *Fundamental Principles of the Theory of Extremal Problems*, John Wiley & Sons, 1987.

[22] J. Warga, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.

[23] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 4, pp. 1310–1311

ERRATUM: "EXISTENCE OF MARKOV CONTROLS AND CHARACTERIZATION OF OPTIMAL MARKOV CONTROLS"*

THOMAS G. KURTZ† AND RICHARD H. STOCKBRIDGE‡

Abstract. A compactness condition used to establish the equivalence of infinite-dimensional linear programming problems with stochastic control problems in [*SIAM J. Control Optim.*, 36 (1998), pp. 609–653.] was misstated. The correct statement of the condition is given, as well as an example to show that the erroneous condition is not sufficient to obtain equivalence.

PII. S0363012998343033

An important condition in [2] is misstated, resulting in an error that may be seriously misleading to users of the results. Expression (6.1) on p. 644, which reads

$$\psi(u) \le a + bc(x,u) \quad \text{or} \quad \psi(u) \le a + bc(s,x,u), \tag{1}$$

should read

$$\psi(u) \le a + bc(x,u)^\beta \quad \text{or} \quad \psi(u) \le a + bc(s,x,u)^\beta, \tag{2}$$

where $0 < \beta < 1$, or slightly more generally,

$$\Phi(\psi(u)) \le a + bc(x,u) \quad \text{or} \quad \Phi(\psi(u)) \le a + bc(s,x,u), \tag{3}$$

where Φ is a *Young function*; that is, $\Phi : [0,\infty) \to [0,\infty)$, Φ is nondecreasing, and $\lim_{r\to\infty} \frac{\Phi(r)}{r} = \infty$.

Condition (vi) of the paper (p. 612) requires that for each $f \in \mathcal{D}(A)$, there exist constants a_f and b_f such that

$$|Af(x,u)| \le a_f + b_f\psi(u). \tag{4}$$

If $\mu_n \in \mathcal{P}(E \times U)$, $\mu_n \to \mu$ weakly, and $\sup_n \int_{E\times U} c(x,u)\mu_n(dx \times du) < \infty$, then (4), together with either (2) or (3), implies

$$\lim_{n\to\infty} \int_{E\times U} Af(x,u)\mu_n(dx \times du) = \int_{E\times U} Af(x,u)\mu(dx \times du). \tag{5}$$

In general, (5) will not hold if (2) and (3) are replaced by the weaker condition (1) as stated in the paper. The conclusion (5) is used explicitly and implicitly throughout section 6 of the paper.

We are, in fact, in the process of writing a second paper that will cover the situations in which (1) is satisfied but not (2) or (3). Such situations arise naturally in the context of singular control, where the conclusions of section 6 are not valid (or it would not be singular control). As a simple example, consider the classical linear regulator, $U = E = \mathbb{R}$, $\mathcal{D}(A) = C_c^2(\mathbb{R})$,

$$Af(x,u) = \frac{1}{2}f''(x) + uf'(x),$$

*Received by the editors August 5, 1998; accepted for publication September 10, 1998; published electronically May 28, 1999.
http://www.siam.org/journals/sicon/37-4/34303.html

†Departments of Mathematics and Statistics, University of Wisconsin, Madison, WI 53706 (kurtz@math.wisc.edu).

‡Department of Statistics, University of Kentucky, Lexington, KY 40506–0027 (stockb@ms.uky.edu).

with running cost

$$c(x,u) = |x| + |u|.$$

The optimal solution to this problem, say for long run average cost, is reflecting Brownian motion on an interval $[-b, b]$. (See, for example, Karatzas [1].) An asymptotically optimal sequence of regular feedback controls is given by

$$u_n(x) = nI_{(-\infty,-b]}(x) - nI_{[b,\infty)}(x).$$

REFERENCES

[1] I. Karatzas, *A class of singular control problems*, Adv. in Appl. Probab., 15 (1983), pp. 225–254.

[2] T. G. Kurtz and R. H. Stockbridge, *Existence of Markov controls and characterization of optimal Markov controls*, SIAM J. Control Optim., 36 (1998), pp. 609–653.

SIAM J. Control Optim.
Vol. 37, No. 5, pp. 1313–1329

LAGRANGE MULTIPLIERS FOR NONCONVEX GENERALIZED GRADIENTS WITH EQUALITY, INEQUALITY, AND SET CONSTRAINTS[*]

JAY S. TREIMAN[†]

Abstract. A Lagrange multiplier rule for finite dimensional Lipschitz problems that uses a nonconvex generalized gradient is proven. This result uses either both the linear generalized gradient and the generalized gradient of Mordukhovich or the linear generalized gradient and a qualification condition involving the pseudo-Lipschitz behavior of the feasible set under perturbations. The optimization problem includes equality constraints, inequality constraints, and a set constraint. This result extends known nonsmooth results for the Lipschitz case.

Key words. Lagrange multipliers, nonsmooth analysis, generalized gradients, optimality conditions

AMS subject classifications. 90C30, 49J52

PII. S0363012996306595

1. Introduction.

In this paper we derive necessary conditions for a finite dimensional constrained optimization problem. The main differences between this and other work is that a small nonconvex generalized gradient is used in conjunction with the generalized gradient of Mordukhovich or with a geometric condition for problems with equality, inequality, and set constraints.

The basic tools used in this paper are the linear generalized gradient (LGG), [21, 22] and the generalized gradient of Mordukhovich and Ioffe (MGG) [7, 8, 9, 11, 13, 14, 15, 16, 17]. In finite dimensions both of these generalized gradients are defined through proximal subgradients. The LGG is always contained in the gradient of Mordukhovich. Both have a nice calculus, with that for LGG involving mostly Lipschitz functions. These generalized gradients are closely related to well known convex generalized gradients. For Lipschitz functions the closed convex hull of the LGG is the generalized gradient of Michel and Penot (MPGG) [10, 12, 19, 20], and the closed convex hull of MGG is the Clarke generalized gradient (CGG) [3, 4, 5].

Since LGG and MGG are generally not convex, they are smaller, respectively, than the generalized gradients of Michel and Penot and Clarke. In particular, this implies that any multiplier rules using these generalized gradients will be sharper when they apply.

There are several important differences between LGG and MGG. One major difference between these generalized gradients is that MGG is upper-semicontinuous (usc) as a multifunction [14], whereas LGG is not. Another difference is that if a function is Fréchet differentiable, then LGG is a singleton [22]. The corresponding condition for MGG is that MGG is a singleton if and only if a function is strictly differentiable [14].

When comparing the multiplier results in this paper with those involving MPGG or CGG it is often useful to consider situations where one generalized gradient is single valued. In order to do this we will use the facts that CGG is a singleton if

[*] Received by the editors July 15, 1996; accepted for publication (in revised form) May 25, 1998; published electronically June 18, 1999.

http://www.siam.org/journals/sicon/37-5/30659.html

[†] Department of Mathematics and Statistics, Western Michigan University, Kalamazoo, MI 49008 (treiman@math-stat.wmich.edu).

and only if a function is strictly differentiable and MPGG is a singleton if a function is Gâteaux differentiable. This will be used to show that the results in this paper may be sharper than those using only MPGG, MGG, or CGG for finite dimensional Lipschitz problems. The basic problem considered in this paper is

$$(*) \qquad \begin{array}{lll} \min f(x) \quad \text{subject to} & g_i(x) \leq 0, & i = 1, 2, \ldots, m, \\ & h_j(x) = 0, & j = 1, 2, \ldots, k, \quad \text{and} \\ & x \in U. & \end{array}$$

Here all of the functions are Lipschitz functions from $\mathbb{R}^n$ to $\mathbb{R}$ and U is a closed subset of $\mathbb{R}^n$. This is a fairly general problem that includes the problem in [21] and the finite dimensional version of the problem in [10]. It is not as general as the problem considered by Mordukhovich [14] since the functions are Lipschitz.

The main optimality condition presented in this paper is a Lagrange multiplier result that uses both LGG and MGG or LGG and a geometric condition. It is sharper than the results of Mordukhovich for (*), but it is weaker in the sense that lower-semicontinuous (lsc) functions are not included. In one version of the result, the MGG is used for the equality constraints and LGG is used for all other functions. Under a pseudo-Lipschitz condition on the behavior of the feasible sets for the equality constraints, one can replace MGG with LGG for the equality constraints. This condition on the constraints is similar to conditions used by Mordukhovich and others.

The pseudo-Lipschitz condition used in this paper is equivalent to a condition using MGG. Our emphasis on the geometric condition makes it clearer that the results involving only LGG do not require the upper-semicontinuity of MGG.

The first section of this paper is an introduction to the definitions and calculus of LGG and MGG. In the second section several technical results showing how equality constraints relate to generalized gradients are proven. The third section is devoted to the main result and several of its corollaries and to a discussion of the relationship with other results. Finally it is shown how one can use our results in a natural setting, bilevel programming.

2. Basic definitions and results.

In order to prove the Lagrange multiplier results we use some calculus results for the generalized gradients in this paper. The proofs of the multiplier results in this paper use the definitions and most of the results in this section. They are all proven elsewhere. *For the results involving MGG the reference* [14] *is usually cited, although many of the results are also in* [16] *and other papers of Ioffe and Mordukhovich.* The basic objects used to define both the LGG and the MGG are the proximal normal and proximal subgradient. For our purposes the definition is restricted to $\mathbb{R}^n$. There are similar ways to define the MGG in Banach spaces. First we define proximal normals and proximal subgradients.

DEFINITION 2.1. *Let $C \subset \mathbb{R}^n$ be a closed set and let $f : \mathbb{R}^n \to \mathbb{R}$ be lsc. A $v \in \mathbb{R}^n$ is a proximal normal to C at $x \in C$ if, for some $\lambda > 0$,*

$$C \cap \bar{B}(x + \lambda v, \lambda\|v\|) = \{x\}.$$

Here $\bar{B}(x, \rho)$ is the closed ball centered at x with radius ρ. A $w \in \mathbb{R}^n$ is a proximal subgradient to f at x if, for some μ,

$$f(y) \geq f(x) + \langle w, y - x\rangle - \mu\|y - x\|^2$$

on a neighborhood of x.

These definitions have been used to characterize a number of generalized gradients and normal cones [3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22]. They are also used, through normal cone definitions, to define co-derivatives [14, 15, 16, 17, 21, 22]. An element of the normal cone of Mordukhovich is defined as the limit of a sequence of proximal normals [11, 14]. To define the linear normal cone one restricts which sequences of proximal normals are allowed.

Definition 2.2. *A sequence of proximal normals $v^k \to v$ to a closed set $C \subset \mathbb{R}^n$ at $x^k \to \bar{x}$ is linear if either $x^k \neq \bar{x}$ for all k and, for some $\lambda > 0$ and all sufficiently large k,*

$$C \cap \bar{B}\left(x^k + \lambda \|x^k - \bar{x}\| v^k, \lambda \|x^k - \bar{x}\| \, \|v^k\|\right) = \{x^k\},$$

or $x^k = \bar{x}$ for all k.

This simply says that the size of the balls defining the proximal normals can be taken as a constant times the distance from x^k to $\bar{x}$. The effect of this is similar to the restriction on the length of vectors when defining MPGG.

With this definition we now define the normal cones.

Definition 2.3. *Let C be a closed subset of $\mathbb{R}^n$. The linear normal cone LNC to C at $\bar{x}$* [22] *is*

$$N_\ell C\bar{x} := \mathrm{cl}\{v : v \textit{ is the limit of a linear sequence of proximal normals to } C \textit{ for } \bar{x}\}.$$

The Mordukhovich normal cone MNC to C at $\bar{x}$ [14] *is*

$$N_M C\bar{x} := \{v : v \textit{ is the limit of a sequence of proximal normals to } C \textit{at } x^k \to \bar{x}\}.$$

It is clear from this definition that MNC always contains LNC. One may also note that MNC is closed without taking the closure of the set of limits of proximal normals. Now we define the generalized gradients. Here $x^k \to_f x$ means that $x^k \to x$ and $f(x^k) \to f(x)$. In what follows, $\|x - y\|_f = \|x - y\| + |f(x) - f(y)|$.

Definition 2.4. *A sequence of proximal subgradients $v^k \to v$ to a lsc f at $x^k \to \bar{x}$ is linear for $\bar{x}$ or linear if either $x^k = \bar{x}$ for all k or $x^k \to_f \bar{x}$, $x^k \neq \bar{x}$ and there exist $\mu, \delta > 0$ such that*

$$f(x^k + h) \geq f(x^k) + \langle v^k, h \rangle - \frac{\mu}{\|x^k - \bar{x}\|_f} \|h\|^2$$

on $B(x^k, \delta \|x^k - \bar{x}\|_f)$.

As with the normal cone, the definition of the linear generalized gradient uses this restricted convergence. One removes this restriction on sequences of proximal subgradients to get MGG.

Definition 2.5 (see [14, 22]). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be lsc. The LGG to f at $\bar{x}$ is the set*

$$\partial_\ell f(\bar{x}) := \mathrm{cl}\{v : v \textit{ is the limit of a linear sequence of proximal subgradients to } f \textit{ for } \bar{x}\}.$$

The MGG to f at $\bar{x}$ is the set

$$\partial_M f(\bar{x}) := \mathrm{cl}\{v : v \textit{ is the limit of a sequence of proximal subgradients to } f \textit{ at } x^k \to \bar{x}\}.$$

As was stated in the introduction, if a function is Fréchet differentiable at a point, LGG is a singleton. This means that for some problems our results will be finer than those using MGG or CGG.

In addition, this generalized gradient can be smaller than MPGG. A simple example is $f(x) = -|x|$. For this function MPGG is $[-1, 1]$ whereas LGG is $\{-1, 1\}$. This means that results using LGG are different than those using MPGG for equality or inequality constraints. Additionally, since LGG is not convex, it can be used more successfully for set constraints. A simple set is the graph of $f(x) = |x|$. At the corner in the graph of $f(x) = |x|$ the normal cone for MPGG is $\mathbb{R}^2$, whereas the LNC is $\{(x, y) : |y| = |x| \text{ or } y \leq |x|\}$. The results for LGG do not coincide with MPGG multiplier results in finite dimensions.

As one hopes, there is a close relationship between these generalized gradients and the corresponding normal cones. In what follows, $\delta_C(x)$ is the indicator function of C,

$$\delta_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{otherwise,} \end{cases}$$

and $d(C, x)$ is the distance from x to C.

Theorem 2.6 (see [14, 22]). *Let C be a closed subset of $\mathbb{R}^n$. Then*

$$\partial_\ell \delta_C(\bar{x}) = N_\ell C \bar{x}$$

and

$$\partial_M \delta_C(\bar{x}) = N_M C \bar{x}.$$

Theorem 2.7 (see [14, 22]). *If C is a closed set in $\mathbb{R}^n$, then*

$$N_\ell C \bar{x} = \mathrm{cl} \cup_{\lambda \geq 0} \lambda \partial_\ell d(C; \bar{x})$$

and

$$N_M C \bar{x} = \cup_{\lambda \geq 0} \lambda \partial_M d(C; \bar{x}).$$

Even with the restriction to linear convergence, the calculus for LGG is fairly rich. It includes a sum rule, a chain rule, and a variety of other results. Only a few results are stated here. It is important to note that some kind of Lipschitz behavior is used in all of these results. The results for MGG are valid in more general situations than those stated here.

Theorem 2.8 (see [7, 14, 22]). *Let f be a lsc function from $\mathbb{R}^n$ to $\mathbb{R}$ and let g be a Lipschitz function from $\mathbb{R}^n$ to $\mathbb{R}$. Then if $f(\bar{x})$ is finite,*

$$\partial_\ell (f + g)(\bar{x}) \subset \partial_\ell f(\bar{x}) + \partial_\ell g(\bar{x})$$

and

$$\partial_M(f+g)(\bar{x}) \subset \partial_M f(\bar{x}) + \partial_M g(\bar{x}).$$

There are rules for positive multiples of a function.

PROPOSITION 2.9 (see [14, 22]). *Let f be a lsc function from $\mathbb{R}^n$ to $\mathbb{R}$ and let $\alpha \geq 0$. Then if $f(\bar{x})$ is finite,*

$$\partial_\ell(\alpha f)(\bar{x}) = \alpha \partial_\ell f(\bar{x})$$

and

$$\partial_M(\alpha f)(\bar{x}) = \alpha \partial_M f(\bar{x}).$$

The following rule will be used in the next section to get an inclusion for the normal cone to the product of sets.

THEOREM 2.10 (see [14, 22]). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be lsc and let $g : \mathbb{R}^m \to \mathbb{R}$ be locally Lipschitz. If f is finite at $\bar{x}$, then*

$$\partial_\ell(f(\bar{x}) + g(\bar{y})) \subset \partial_\ell f(\bar{x}) \times \partial_\ell g(\bar{y})$$

and

$$\partial_M(f(\bar{x}) + g(\bar{y})) = \partial_M f(\bar{x}) \times \partial_M g(\bar{y}).$$

The most basic optimality condition holds for the LGG. Without this result a generalized gradient is not very useful for optimization. We, of course, need it in the proof of the main result.

PROPOSITION 2.11 (see [14, 22]). *If $f : \mathbb{R}^n \to \mathbb{R}$ is lsc and $\bar{x}$ is a local minimizer of f, then*

$$0 \in \partial_\ell f(\bar{x}) \quad \textit{and} \quad 0 \in \partial_M f(\bar{x}).$$

The next result concerns the LGG for the maximum of a finite collection of functions. This type of result is often used to prove calculus rules and Lagrange multiplier rules [21]. It is important in the proof of the main theorem. A similar result holds for MGG [14].

THEOREM 2.12. [21] *Let $g_1, g_2, \ldots, g_n$ be a finite collection of Lipschitz functions from $\mathbb{R}^m$ to $\mathbb{R}$. Then*

$$\partial_\ell \max_{i=1,2,\ldots,n} g_i(x) \subset \left\{ \sum_{i=1,2,\ldots,n} \lambda_i \partial_\ell g_i(x) : \lambda_i \geq 0,\ \lambda_i = 0 \textit{ if } i \notin I(x) \textit{ and } \sum_{i \in I(x)} \lambda_i = 1 \right\},$$

where $I(x) = \{i : g_i(x) = \max_{j=1,\ldots,n} g_j(x)\}$.

3. Equalities and generalized gradients. The basic technique for handling equality constraints in this paper is to rewrite the equality constraints as indicator functions of level sets for the functions. If one can rewrite the generalized gradients of these indicator functions in terms of the generalized gradients of the original functions, one can prove Lagrange multiplier results. This is what we do.

In this section we relate the normal cone to the level set of a function and the generalized gradient of that function. This result is a generalization of the standard result that a normal to the intersection of the level sets of a finite number of C^1 functions with linearly independent derivatives is a linear combination of their derivatives. As is noted by Rockafellar [18], this type of result is not very useful when a generalized gradient is convex.

The basic result we wish to prove is that the normal cone to a set of the form $\{x : h_j(x) = 0,\ j = 1, 2, \ldots, m\}$ is contained in the positive linear combinations of the generalized gradients of the h_j's. This result is not true without some condition. The condition used here corresponds to the linear independence of the derivatives in classical results.

First we look at the case of a single function. The classical result is that the normal cone is the multiples of the gradient of h if $\nabla h(x) \neq 0$. Since $\partial h(x) \neq -\partial(-h)(x)$, it is not surprising that the classical condition $\nabla h(x) \neq 0$ is replaced by the condition that $0 \notin \partial h(x) \cup \partial(-h)(x)$.

We prove only the results in this section for LGG. The same or similar results hold for MGG with simpler proofs. The results for MGG are stated without proofs.

PROPOSITION 3.1. *Assume h is a Lipschitz function from $\mathbb{R}^n$ to $\mathbb{R}$ with $h(\bar{x}) = 0$. Let $C = \{x : h(x) = 0\}$. If $0 \notin \partial_\ell h(\bar{x}) \cup \partial_\ell(-h)(\bar{x})$, then*

$$N_\ell C\bar{x} \subset \cup_{\alpha\in\mathbb{R}} \partial_\ell(\alpha h)(\bar{x}).$$

If $0 \notin \partial_M h(\bar{x}) \cup \partial_M(-h)(\bar{x})$, then [14]

$$N_M C\bar{x} \subset \cup_{\alpha\in\mathbb{R}} \partial_M(\alpha h)(\bar{x}).$$

Proof. Note that $P = \partial_\ell h(\bar{x}) \cup \partial_\ell(-h)(\bar{x})$ is a closed set since h is Lipschitz and all proximal subgradients must have norm at most L, the Lipschitz constant for h. Since $0 \notin P$, P is bounded away from zero. Thus $P' = \cup_{\alpha\geq 0}\alpha P$ is a closed cone. If the set of all limits of linear sequences of proximal normals to C at $\bar{x}$ is in P', the result follows.

Let $v \notin P'$ be the nonzero limit of a linear sequence of proximal normals to C at $\bar{x}$. Then there are $v^k \to v$ such that either each v^k is a proximal normal to C at $\bar{x}$ or the v^k's form a linear sequence of proximal normals to C at $x^k \to \bar{x}$ with $x^k \neq \bar{x}$. The two cases are similar. Only the latter, more difficult case is considered.

Since $v \neq 0$, one may assume that $v^k \neq 0$ for all k. Assume that for some $\mu > 0$,

$$C \cap \bar{B}(x^k + \mu_k v^k, \mu_k\|v^k\|) = \{x^k\}$$

where $\mu_k = \mu\|x^k - \bar{x}\|$. Since h is continuous and nonzero on $B(x^k + \mu_k v^k, \mu_k\|v^k\|)$, it has constant sign there. Assume h is positive on this set. If it is negative, work with $-h$.

To simplify notation, let $z^k = x^k + \mu_k v^k/2$ and $\rho_k = \mu_k\|v^k\|/2$.

For each k define a function $r^k_{\delta,\omega}$,

$$r^k_{\delta,\omega}(x) = \omega + \begin{cases} 0 & \text{if } 0 \leq \|x - z^k\| < \rho_k, \\ -\frac{(\|x-z^k\|-\rho_k)^2}{\rho_k \delta} & \text{if } \rho_k \leq \|x - z^k\| < (1 + L\delta)\rho_k, \\ -L^2(\|x - z^k\| - (1 + L\delta)\rho_k) - L^2\delta\rho_k & \text{otherwise.} \end{cases}$$

By definition, when $\omega \leq 0$, $r^k_{\delta,0}(x)$ is less than h if $x \in B(z^k, \rho_k)$ or $\|x - z^k\| > (1 + L\delta)\rho_k$. Let U^k_δ be the set

$$U^k_\delta = \bar{B}(z^k, (1 + L\delta)\rho_k)\backslash B(x^k + \mu_k v^k, \mu_k\|v^k\|).$$

Note that for each k the gradient of $r^k_{\delta,\omega}$ at any x is either zero or in the direction from x to z^k. As $\delta \searrow 0$, the gradients of $r^k_{\delta,\omega}$ on U^k_δ are independent of ω and converge uniformly to points on the line segment $[0, 2L] \cdot \frac{v^k}{\|v^k\|}$.

For each k, there is a maximum $\omega_k \leq 0$ such that $r^k_{\delta,\omega_k}(x) \leq h(x)$ for all x. For this ω_k, if $r^k_{\delta,\omega_k}(y^k) = h(y^k)$, then the gradient w^k of r^k_{δ,ω_k} at y^k is a proximal subgradient to h at y^k. In addition, the w^k's form a linear sequence of proximal subgradients to h at $y^k \to \bar{x}$.

For a fixed δ, by passing to a subsequence, one can assume that the w^k's converge to some $w^\delta \in \partial_\ell h(\bar{x})$. Since $0 \notin \partial_\ell h(\bar{x})$, the w^δ's are bounded away from 0. Taking a sequence of δ's converging to zero and passing if necessary to a subsequence, there is a $w \in \partial_\ell h(\bar{x})$ such that $v = \lambda w$ for some $\lambda > 0$.

This completes the proof by contradicting that $v \notin P'$. □

If one assumes the function h is Fréchet differentiable at $\bar{x}$, this result reduces to the conculsion that the LNC to $C = \{x : h(x) = h(\bar{x})\}$ at $\bar{x}$ is a single line if $\nabla h(\bar{x}) \neq 0$.

The following proposition is needed for the main technical result of this section, Proposition 3.4. It relates the normal cone to the product of sets to the product of the normal cones. Unlike MNC, this will not be an equality except for special sets where one can control convergence rates for all sets and sequences at the same time.

PROPOSITION 3.2. *Let $C_1, C_2, \ldots, C_k$ be closed subsets of $\mathbb{R}^n$ and let $C_0 = C_1 \times C_2 \times \cdots \times C_k \subset \mathbb{R}^{nk}$. Then*

$$N_\ell C_0(x, x, \ldots, x) \subset N_\ell C_1 x \times N_\ell C_2 x \times \cdots \times N_\ell C_k x$$

and [14]

$$N_M C_0(x, x, \ldots, x) = N_M C_1 x \times N_M C_2 x \times \cdots \times N_M C_k x.$$

Proof. The first part can be proven by a simple direct argument using the fact that $\|x_i\| \leq \|(x_1, x_2, \ldots, x_n)\|$ to generate linear sequences of proximal normals for each C_i from a linear sequence of proximal normals to C_0. The result for $N_M Cx$ has a simple direct proof in the reference. □

In Proposition 3.4 the normal cone to an intersection of sets is expressed in terms of the sum of the normal cones to the original sets. Simple examples show that some condition on the sets is required for this type of result.

Example 3.1. Let $X = \mathbb{R}^2$ and take $C_1 = \bar{B}((0, 1), 1) \cup \bar{B}((0, -1), 1)$ and $C_2 = \{(x, 0) : x \in \mathbb{R}\}$. Here C_1 is the union of two closed balls that are tangent and intersect at the origin. The tangent to the balls at the origin is C_2. This means that $C = C_1 \cap C_2 = \{(0,0)\}$, $N_\ell C_1(0,0) = \{(0,0)\}$, $N_M C_1(0,0) = N_\ell C_2(0,0) = N_M C_2(0,0) = \{(0,y) : y \in \mathbb{R}\}$, and $N_\ell C(0,0) = N_M C(0,0) = \mathbb{R}^2$. Clearly the sum of the normal cones does not contain the normal cone of the intersection.

The condition used in this paper involves the behavior of all of the level sets near the point of interest. There are different ways of expressing this condition. Each is interesting in its own right.

Take $C_1, C_2, \ldots, C_k$ to be closed subsets of $\mathbb{R}^n$. Let $\Phi(v)$ be the set valued function from $\mathbb{R}^{nk}$ to $\mathbb{R}^n$ defined by

$$\Phi(v) = \cap_{i=1}^k (C_i + v_i).$$

The condition is that Φ is pseudo-Lipschitz at $(0, \bar{x})$. Aubin [2] introduced the concept of a pseudo-Lipschitz multifunction. A multifunction $\Omega(v)$ is pseudo-Lipschitz at (v, z) with $z \in \Omega(v)$ if there are neighborhoods V of v and Z of z and a $c > 0$ such that for any v_1 and v_2 in V,

$$\Omega(v_1) \cap Z \subset \Omega(v_2) + c\|v_1 - v_2\|\, B(0,1).$$

Among other things, this implies that Ω is nonempty on V.

If, in our problem, there is only one equality or set constraint, the multifunction Φ is automatically pseudo-Lipschitz. This is easy to see since $\Phi(v)$ will be a linear translation of a single set. Another situation where Φ is pseudo-Lipschitz is where there is no set constraint, all of the equality constraints are C^1, and their gradients are linearly independent. This is a classical constraint qualification.

This pseudo-Lipschitz condition on Φ can also be expressed as saying the set valued function

$$\Psi(w) = D \cap \big((C_1 + w_1) \times (C_2 + w_2) \times \cdots \times (C_k + w_k)\big),$$

with $D = \{(x, x, \ldots, x) \in \mathbb{R}^{nk} : x \in \mathbb{R}^k\}$, is pseudo-Lipschitz at $\bar{x}^k = \big((0,0,\ldots,0), (\bar{x}, \bar{x}, \ldots, \bar{x})\big)$. We use this version of the pseudo-Lipschitz condition in the proofs of Propositions 3.3 and 3.4. In this setting it is easier to relate the normal cone of the intersection to the normal cones of the individual sets.

This type of condition and its relationship to MGG and MNC are discussed in [15, 17]. Using Theorem 4.1 in [17], one can show that a sufficient condition for $\Phi(v)$ to be pseudo-Lipschitz is

$$N(D, \bar{x}^k) \cap N_M \mathrm{gph} H \bar{x}^k = \{0\},$$

where $H(x) = (h_1(x), h_2(x), \ldots, h_k(x))$. One can also use the following sufficient condition for Φ to be pseudo-Lipschitz:

$$N(D, \bar{x}^k) \cap N_M C_1 \times C_2 \times \cdots \times C_k \bar{x}^k = \{0\},$$

where $\bar{x}^k = (\bar{x}, \bar{x}, \ldots, \bar{x})$. Rewriting this yields

$$\sum\nolimits_{i=1,2,\ldots,k}^{v_i \in N_M C_i \bar{x}} v_i = 0 \quad \text{implies} \quad v_i = 0, \quad i = 1, 2, \ldots, k.$$

These conditions involving MNC are versions of the standard condition used in Mordukhovich's multiplier rule [14]. They are also used by other authors.

It turns out that this is also a necessary condition for $\Phi(v)$ to be pseudo-Lipschitz. We will use the next result throughout the rest of the paper in proofs and to state results in terms of the pseudo-Lipschitz property.

PROPOSITION 3.3. *Let Φ be as above. The multifunction Φ is pseudo-Lipschitz at $\bar{x}$ if and only if*

$$(\mathcal{C}) \qquad N(D,\bar{x}^k)\cap N_M C_1\times C_2\times\cdots\times C_k\bar{x}^k=\{0\}.$$

Proof. The sufficiency follows from Theorem 4.1 of [17] using $F:\mathbb{R}^k\times\mathbb{R}^n\to\mathbb{R}^{nk}$ defined by

$$F(x_1,\dots,x_k,y)=(y-x_1-C_1,\dots,y-x_k-C_k),$$

$\lambda=\{0\}$, and $\Omega=\mathbb{R}^k\times\mathbb{R}^n\times\mathbb{R}^{nk}$. It can also be proven directly by noting that if Ψ is not pseudo-Lipschitz at $\bar{x}$, then there exist $x^p\to\bar{x}$ such that

$$\frac{d(C_1\times\cdots\times C_k,x^p)}{d(D\cap(C_1\times\cdots\times C_k),x^p)}\to 0.$$

Applying a variational argument yields a sequence of proximal normals to $C_1\times\cdots\times C_k$ at $y^p\to\bar{x}$ such that $\|v^p\|=1$ for all p and $v^p|_D\to 0$. The sufficiency follows. To prove the necessity we work with Ψ instead of Φ. Assume there is a

$$v\in N(D,\bar{x}^k)\cap N_M C_1\times C_2\times\cdots\times C_k\bar{x}^k\backslash\{0\}.$$

Then there are sequences v^j and x^j such that $x^j\to\bar{x}^k$, $v^j\to v$, and v^j is a proximal normal to $C_1\times C_2\times\cdots\times C_k$ at x^j. Note that $0\in\Psi(-x^j)$. If the rate of change of $d(0,\Psi(-x^j-\alpha v^j))$ as α increases from 0 can be made arbitrarily large for large j, then Ψ, and hence Φ, are not pseudo-Lipschitz at $\bar{x}$. Since v^j is a proximal normal to $C_1\times C_2\times\cdots\times C_k$ at x^j, $D\cap B(r_jv^j-\alpha v^j,r_j\|v^j\|)\cap\Psi(-x^j-\alpha v^j)=\emptyset$ for some fixed $r_j>0$ and for all $\alpha>0$. This implies, if θ_j is the angle between v^j and D, the rate of change of $d(0,\Psi(-x^j-\alpha x^j))$ as α increases at $\alpha=0$ is at least $\|v^j\|\tan(\theta_j)$. Since $\theta_j\to\pi/2$ as $j\to\infty$, Ψ is not pseudo-Lipschitz at $\bar{x}^k$. □

See Mordukhovich's work [13, 14, 15, 16] for more details and examples of this type of equivalence. The next result is the most important of this section. It and its corollary are used in the next section.

PROPOSITION 3.4. *Let $C_1,C_2,\dots,C_k$ be closed subsets of $\mathbb{R}^n$ and let $C=\cap_{j=1}^k C_j$. If $\Phi(v)$ is pseudo-Lipschitz at $(0,\bar{x})$, then*

$$N_\ell C\bar{x}\subset\sum_{j=1}^k N_\ell C_j\bar{x}$$

and

$$N_M C\bar{x}\subset\sum_{j=1}^k N_M C_j\bar{x}.$$

Proof. The proofs are almost identical for the LNC and MNC. Therefore, as usual, only the proof for the LNC is given. Using Theorem 2.6 we can work with the LGG of δ_{C_0} and δ_C rather than with the LNC. In addition, we work with $C_0=C_1\times C_2\times\cdots\times C_k$

and D. The case when there are proximal subgradients to δ_C at $\bar{x}$ is omitted because it is similar to what follows. Let $v \in N_\ell C\bar{x}$ and let $v^k \to v$ be a linear sequence of proximal subgradients to δ_C at $x^k \to \bar{x}$, $x^k \neq \bar{x}$. Take $y \in \mathbb{R}^{nk}$, $y^k = (x^k, x^k, \dots, x^k)$, $\bar{y} = (\bar{x}, \bar{x}, \dots, \bar{x})$, and $w^k = (v^k, v^k, \dots, v^k)$. Then, for some μ and ρ,

$$\delta_{C_0}(y) \geq \delta_{C_0}(y^k) + \langle w^k, y - y^k\rangle - \rho_k \|y - y^k\|^2$$

if $y \in D$, $\|y - y^k\| \leq \mu_k = \mu\|y^k - \bar{y}\|$, and $\rho_k = \rho/\|y^k - \bar{y}\|$. It is relatively simple to show that if the conditions of the proposition are met, there are $\lambda, \beta > 0$ such that, if $z^k = (z, z, \dots, z) \in C_0$ and $D \cap C_0 \cap (z^k + (0, \beta\mu\|z^k\|] \cdot \bar{B}(w^k, \lambda)) = \emptyset$, then

$$C_0 \cap (z^k + (0, \beta\mu\|z^k\|] \cdot \bar{B}(w^k, \lambda)) = \emptyset$$

for any z in the intersection of C and some neighborhood of $\bar{x}$. Fix $\epsilon > 0$. One may assume that the gradient of $\rho_k\|y - y^k\|^2 < \epsilon$ if $\|y - y^k\| < 2\beta\mu_k(\|v^k\| + \lambda)$. Choose an $\eta > 0$ such that if $d(y, D) = \lambda\mu_k\beta/2$ and $\|y_D - y^k\| \leq \beta\mu_k(\|v^k\| + \lambda)$, then

$$r_k(y) = \langle w^k, y - y^k\rangle - \rho_k\|y_D - y^k\|^2 - \eta_k(d(y, D))^2 < 0, \tag{3.1}$$

where y_D is the orthogonal projection of y onto D and $\eta_k = \eta/\|y^k - \bar{y}\|$.

This means that the function r_k in (3.1) is less than zero on the boundary of $S = B(y^k, \beta\mu_k(\|v^k\| + \lambda) \cap \{y : d(y, D) \leq \lambda\mu_k\beta/2\}$. Thus, for some $\omega < 0$, $r^k(y) + \omega \leq \delta_{C_0}(y)$ on S and equals $\delta_{C_0}(y)$ at some point z^k in the interior of S. The gradients of r^k at $z^k \to \bar{y}$ form a linear sequence of proximal subgradients to δ_{C_0}. Taking a subsequence, if necessary, one may assume that these converge to a v_ϵ such that the orthogonal projection of v_ϵ onto D satisfies $\|v_\epsilon - v\| \leq \epsilon$.

There are two situations. If, as $\epsilon \searrow 0$, a sequence of the v_ϵ's is bounded, then there is a $v_0 \in \partial_\ell\delta_{C_0}(\bar{y})$ whose orthogonal projection onto D is v. On the other hand, if no bounded sequence exists, then, by renorming to unit lengths, there is a $v_1 \in \partial_\ell\delta_{C_0}(\bar{y}) \subset \partial_M\delta_{C_0}(\bar{y})$ such that v_1 is orthogonal to D. This contradicts the condition involving the Mordukhovich cone before the propositions and hence contradicts our assumptions. This means that v_0 exists.

Since $N_\ell C_0\bar{y} \subset N_\ell C_1\bar{x} \times N_\ell C_2\bar{x} \times \cdots \times N_\ell C_k\bar{x}$ and the normal cone to D is $\{(w^1, w^2, \dots, w^k) : \sum_{i=1}^k w^i = 0\}$, the result follows. □

Finally comes the result that yields the multiplier rules in the next section. It relates the normal cone to a set defined by equality constraints to the generalized gradients of the constraints. An arbitrary set constraint is included to match the optimization problem in this paper.

Theorem 3.5. *Let $h_1, h_2, \dots, h_k$ be Lipschitz functions from $\mathbb{R}^n$ to $\mathbb{R}$ and let C_{k+1} be a closed subset of $\mathbb{R}^n$. Assume that $h_j(\bar{x}) = 0$ for $j = 1, 2, \dots, k$. Take $C_j = \{x : h_j(x) = 0\}$ and take $\Phi(v)$ as in Proposition* 3.5. *Assume $\Phi(v)$ is pseudo-Lipschitz at $(0, \bar{x})$. If $0 \notin \partial_\ell h_j(\bar{x}) \cup \partial_\ell(-h_j)(\bar{x})$ for $j = 1, 2, \dots, k$, then*

$$N_\ell C\bar{x} \subset \sum\nolimits_{\alpha_j \in \mathbb{R}}^{j=1,2\dots k} \partial_\ell(\alpha h_j)(\bar{x}) + N_\ell C_{k+1}\bar{x}.$$

If $0 \notin \partial_M h_j(\bar{x}) \cup \partial_M(-h_j)(\bar{x})$ for $j = 1, 2, \dots, k$, then [14]

$$N_M C\bar{x} \subset \sum\nolimits_{\alpha_j \in \mathbb{R}}^{j=1,2\dots k} \partial_M(\alpha h_j)(\bar{x}) + N_M C_{k+1}\bar{x}.$$

Proof. Apply Propositions 3.1 and 3.4 to the sets defined in the theorem. □

4. The main results. In this section Lagrange multiplier rules are given for the LGG that include equality constraints. Recall that the problem being considered is

$$\begin{aligned} &(4.1) \qquad && \min f(x) \quad \text{subject to} \quad && g_i(x) \leq 0, \quad i = 1, 2, \ldots, m, \\ &(*) && && h_j(x) = 0, \quad j = 1, 2, \ldots, k, \\ &(4.2) && && x \in U. \end{aligned}$$

All of the functions are Lipschitz functions from $\mathbb{R}^n$ to $\mathbb{R}$ and U is a closed subset of $\mathbb{R}^n$.

The main result in this section is related to Theorem 6.1.1 of Clarke [5], the theorem of Ioffe [10], and the generalized gradient interpretation of Theorem 7.5 of Mordukhovich [14]. All of those results state that if $\bar{x}$ is a minimizer of (*), then for some constants $\beta \geq 0$, $\lambda_i \geq 0$, $i = 1, 2, \ldots, m$, and α_J, $j = 1, 2, \ldots, k$, not all zero, one has

$$0 \in \partial \left(\beta f_0 + \sum_{i=1}^{m} \lambda_i g_i + \sum_{j=1}^{k} \alpha_J h_j \right) (\bar{x}) + N(U, \bar{x}).$$

Here ∂ represents either the CGG, the MPGG, or the MGG. The results of Clarke and Penot assume that all functions are Lipschitz. Clarke's results are finite dimensional, Ioffe's require that U is convex, and Mordukhovich only assumes that the functions are lsc. Mordukhovich's result is sharper than Clarkes since the CGG always contains the MGG. The results in this section are stronger than Theorem 7.5 of Mordukhovich [14] in that the LGG is used. On the other hand, the results are weaker because only Lipschitz functions are allowed. The results are sharper than the finite dimensional restriction of the theorem of Ioffe [10] in the sense that convexity is not required. For the result that only involves LGG the pseudo-Lipschitz behavior of $\Phi(v)$ is invoked instead of using the upper-semicontinuity of MGG or the convexity of MPGG. This means that the results do not compare directly with previous results. The Lagrange multiplier rule is now stated and a proof given. We again take $C_j = \{x : h_j(x) = 0\}$ for $j = 1, 2, \ldots, k$, $C = \cap_{j=1,2,\ldots,k} C_j \cap U$, and

$$\Phi(v) = \cap_{j=1,2,\ldots,k} (C_j + v_j) \cap (U + v_{k+1}).$$

THEOREM 4.1. *Let $\bar{x}$ be a minimizer of (*). Then there exist $\beta \geq 0$, $\lambda_i \geq 0$ and α_j, not all zero, such that $\lambda_i g_i(\bar{x}) = 0$ for all i and*

$$0 \in \beta \partial_\ell f(\bar{x}) + \sum_{i=1}^{m} \lambda_i \partial_\ell g_i(\bar{x}) + \sum_{j=1}^{k} \partial_M (\alpha_j h_j)(\bar{x}) + N_M U \bar{x}.$$

Proof. Apply Theorems 2.6, 2.8, and 2.12 to the function

$$\max\{f(x) - f(\bar{x}), g_1(x), g_2(x), \ldots, g_m(x)\} + \delta_C(x)$$

to get that

$$\begin{aligned} 0 &\in \beta \partial_\ell f(\bar{x}) + \sum_{i=1}^{m} \lambda_i \partial_\ell g_i(\bar{x}) + N_\ell C \bar{x} \\ &\subset \beta \partial_\ell f(\bar{x}) + \sum_{i=1}^{m} \lambda_i \partial_\ell g_i(\bar{x}) + N_M C \bar{x}, \end{aligned}$$

where either β or one of the λ's is not zero and $\lambda_i g_i(\bar{x}) = 0$ for all i. If $0 \in \partial_M h_j(\bar{x}) \cup \partial_M(-h_j)(\bar{x})$ for some j, the result is true. If

$$0 \notin \sum\nolimits_{(\alpha_1 \ldots \alpha_k) \neq 0}^{j=1^k} \partial_M(\alpha h_j)(\bar{x}) + N_M U\bar{x},$$

Theorem 3.5 applies to $N_M C\bar{x}$ since $\Phi(v)$ must be pseudo-Lipschitz. This also means that the result holds. □

Since $\partial_\ell s(x) \subset \partial_M s(x)$ for any function, one can replace the LGG in Theorem 4.1 with MGG to get the following restriction of Theorem 7.5 of Mordukhovich [14].

COROLLARY 4.2. *Let $\bar{x}$ be a minimizer of (*). Then there exist $\beta \geq 0$, $\lambda_i \geq 0$, and α_j, not all zero, such that $\lambda_i g_i(\bar{x}) = 0$ for all i and*

$$0 \in \beta \partial_M f(\bar{x}) + \sum_{i=1}^{m} \lambda_i \partial_M g_i(\bar{x}) + \sum_{j=1}^{k} \partial_M(\alpha_j h_j)(\bar{x}) + N_M U\bar{x}.$$

If one uses the pseudo-Lipschitz behavior of the multifunction Φ, defined in section 3, one can remove MGG from the result. Professor A. Ioffe has stated that, if one does not use a convex or usc generalized gradient, a constraint qualification is necessary for a Lagrange multiplier result. We use the pseudo-Lipschitz behavior of Φ. This result is tighter than the above result since the pseudo-Lipschitz behavior of Φ is equivalent to $0 \notin \sum_{j=1}^{k} [N_M C_j \bar{x} \backslash \{0\}] + [N_M U\bar{x} \backslash \{0\}]$. The proof is almost identical to that for Theorem 4.1.

THEOREM 4.3. *Let $\bar{x}$ be a minimizer of (*) and assume that $\Phi(v)$ is pseudo-Lipschitz at $(0, \bar{x})$. Then there exist $\beta \geq 0$, $\lambda_i \geq 0$, and α_j, not all zero, such that $\lambda_i g_i(\bar{x}) = 0$ for all i and*

$$0 \in \beta \partial_\ell f(\bar{x}) + \sum_{i=1}^{m} \lambda_i \partial_\ell g_i(\bar{x}) + \sum_{j=1}^{k} \partial_\ell(\alpha_j h_j)(\bar{x}) + N_\ell U\bar{x}.$$

Now we turn our attention to examples showing that these results actually differ from previous results. We first give an example where a nonconvex set constraint excludes using Ioffe's result.

Example 4.1. Consider the problem

$$\min_{x,y \in \mathbb{R}} 2\,|y| - x \qquad \text{subject to} \qquad (x, y) \in U,$$

where $U = \{(x, y) : x \geq y \text{ or } x \leq -y\}$. Ioffe's result doesn't apply since U is not convex. If one uses MGG, CGG, or LGG, the only critical point is $\bar{x} = (0, 0)$, where MGG, CGG, and LGG coincide:

$$\partial f(\bar{x}) = \partial_M f(\bar{x}) = \partial_\ell f(\bar{x}) = \{-1\} \times [-2, 2].$$

Also

$$N_M U\bar{x} = N_\ell U\bar{x} = [0, \infty) \times \{(1, -1), (1, 1)\}.$$

Remark 4.1. It is important to note that multiplier existence results may depend on how the constraint region is represented. That is, the same constraint region can satisfy a constraint qualification using one representation while it does not with another. Unfortunately, it is sometimes difficult, if not impossible, for the modeler to determine whether or not a given representation has the right properties. An example of this phenomenon occurs in Example 4.1. In this example, if the set U is written as a functional constraint, $y - |x| \leq 0$, then Ioffe's result applies.

The next example shows that, even without a set constraint, the results in this paper are sharper than those using convex generalized gradients. Here ∂_P denotes the MPGG and ∂ denotes the CGG.

Example 4.2. Let f be an arbitrary differentiable function from $\mathbb{R}^2 \to \mathbb{R}$ and define g by

$$g(x,y) = \begin{cases} -x + y, & x, y > 0, \\ x + y, & x < 0, y > 0, \\ -x - y, & x > 0, y < 0, \\ x - y, & x, y < 0. \end{cases}$$

The point we consider is $(0,0)$. For this function we have

$$\partial_\ell g(0,0) = \partial_M g(0,0) = \{-1, 1\} \times [-1, 1]$$

and

$$\partial g(0,0) = \partial_P g(0,0) = [-1,1] \times [-1,1].$$

For the problem

$$\min f(x,y) \quad s.t. \quad g(x,y) \leq 0,$$

the feasible set is $\{(x,y) : |x| \geq |y|\}$. For any Lipschitz f the point $(0,0)$ is a critical point for CGG and MPGG. On the other hand, the multiplier conditions for MGG and LGG in this paper are met only if there is a (v,w) in $\partial_M f(0,0)$ or $\partial_\ell f(0,0)$ with $|v| = |w|$.

The fact that the multifunction $\Phi(v)$ is always pseudo-Lipschitz if there is a single equality constraint allows simple examples to demonstrate that Theorem 4.3 is sharper than Theorem 4.1. The basis of the following example is that the constraint function is Fréchet differentiable at the point of interest but it is not strictly differentiable there.

Example 4.3. Let $f(x,y) = (y-1)^2$ and let

$$h(x,y) = \begin{cases} \frac{1}{n} + \frac{1}{(100n)^2} - 10|x - \frac{1}{n}| & \text{if } x \in \left(\frac{1}{n} - \frac{1}{90000n^2}\,,\ \frac{1}{n} + \frac{1}{110000n^2}\right) \\ & \quad \text{for } n \in \mathbb{N}, \\ x & \text{otherwise.} \end{cases}$$

Consider the problem

$$\min f(x,y) \quad s.t. \quad h(x,y) = 0.$$

The feasible set $C = \{(x,y) : h(x,y) = 0\}$ is the y-axis. This means that the optimal point is $(0,1)$. Since h is Fréchet differentiable along the y-axis, $\partial_\ell f(0,y) = (1,0)$ and the only critical point is $(0,1)$ when one applies Theorem 4.3. On the other hand one can calculate that $\partial_M h(0,y) = [-10,10] \times \{0\}$ for any $y \in \mathbb{R}$. This means that all feasible points are critical points for Theorem 4.1. Here the MGG result is not as tight as the result using only LGG.

5. Bilevel optimization. An application of the main results of paper is given in this section. It is a simple bilevel problem that demonstrates how one might use the linear and Mordukhovich generalized gradients in tandem. The problem is

$$\text{(UL)} \qquad \min f(x,y) \quad \text{subject to}$$
$$\text{(LL}_1\text{)} \qquad y \in \operatorname{argmin}\{\min_y g(x,y)\}.$$

Here we assume that $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}$ are Lipschitz functions. There has been a large amount of recent research on bilevel problems, see for example [1, 6, 23, 24]. If one tries to derive necessary conditions for this problem one encounters several problems almost immediately. The first is that the lower level constraint on the upper level problem is not "nice." Another is that a second order derivative is required to directly include a necessary condition for LL_1 as a constraint on the upper level problem. If one uses a good generalized gradient $\partial_\#$, the necessary condition for the lower level problem LL_1 is

$$0 \in \partial_{\#,y} g(x,y).$$

Let $S_1(x) = \{y : 0 \in \partial_{\#,y} g(x,y)\}$. Since one must make strong assumptions to make

$$S_1(x) = \operatorname{argmin}\{\min_y g(x,y)\},$$

we first replace the original problem with

$$\text{(UL)} \qquad \min f(x,y) \quad \text{subject to}$$
$$\text{(LL)} \qquad y \in S_1(x).$$

When considering which generalized gradients to use for a second order object, most people require that the graph of $z \to \partial_\# g(z)$ is a multifunction with closed graph. (A requirement that sets be closed is used in the definitions of almost all normal cones and generalized gradients.) In our situation this means that the object one should use for the first order condition on the lower level problem is either MGG or CGG. Since, as was mentioned earlier, CGG is not appropriate, we use MGG. An additional problem is that using MGG in the lower level necessary condition will only make $S_1(x)$ closed for each x. What is actually desired is that $S_1(x)$ have closed graph. In order to do this, one must expand $S_1(x)$ again. The easiest way is to replace S_1 with

$$S(x) = \{y : (x^*, 0) \in \partial_M g(x,y)\}.$$

Under our assumption that g is Lipschitz, we have $S_1(x) \subset S(x)$. Replacing $S_1(x)$ with $S(x)$, we express the graph of S as the first two components of the intersection of the two sets; $C_1 = \operatorname{gph}\partial_M g(x,y)$ and $C_2 = \{(x,y,x^*,0) : x, x^* \in \mathbb{R}^n,\ y \in \mathbb{R}^m \text{ and } 0 \in \mathbb{R}^m\}$. The second set can be written as an equality constraint $h(x,y,z,w) = w = 0$. The problem now becomes

$$\text{(5.1)} \qquad \min_{x,y,z,w} f(x,y) \quad \text{subject to } (x,y,z,w) \in \operatorname{gph}\partial_M g(x,y)$$
$$\text{and} \quad h(x,y,z,w) = 0.$$

Applying Theorem 4.1 to problem (5.1) gives the following condition. There exist z^* and $\lambda \geq 0$, not both zero, and x^* such that

$$(x^*, 0) \in \partial_M g(x,y)$$

and

$$(0,0,0,0) \in \lambda\partial_\ell f(x,y) \times \{0\} \times \{0\} + N_M \text{gph}\partial_M g(x,y)(x,y,x^*,0) + (0,0,0,z^*).$$

The second condition can be rewritten as there exists $(x_1^*, y_1^*) \in D^*\partial_M g(x,y)(s^*, w^*)$ such that

$$(0,0) \in \partial_\ell f(x,y) + (x_1^*, y_1^*)$$

and

$$(0,0) \in (s^*, w^*) + (0, z^*).$$

Here D^* is the co-derivative of Mordukhovich. This says there exist w^* and $\lambda \geq 0$, not both zero, and

$$(x_1^*, y_1^*) \in D^*\partial_M \big(g(x,y)(x,y,x^*,0)\big)(0,w^*)$$

such that

$$(0,0) \in \lambda\partial_\ell f(x,y) + (x_1^*, y_1^*).$$

The above gives the following result. This result does not make any assumptions about the properties of $\partial_M g(x,y)$ except for the closed graph property. This is different from most results since there are no qualifications on the argmin set of LL.

THEOREM 5.1. *Let $(\bar{x}, \bar{y})$ be a solution to* (5.1). *Then there exist $\lambda \geq 0$ and $w^* \in \mathbb{R}^m$, not both zero, such that*

$$-\lambda\partial_\ell f(\bar{x},\bar{y}) \cap D^*\partial_M \big(g(x,y)(\bar{x},\bar{y},x^*,0)\big)(0,w^*) \neq \emptyset.$$

Many of the qualifications used in necessary conditions for bilevel programming guarantee that the multifunction $S(x)$ is well behaved. Convexity of $g(x,y)$ in y implies that $S(x)$ is a closed convex set for each x and strict convexity reduces S to a function. Both of these give that $S(x)$ is closed valued. Other conditions are used to guarantee that, in addition, $S(x)$ has closed graph. In some sense, these usually come down to giving $S(x)$ some Lipschitz property. The type of condition used in Zhang [24] involves the inverse of the multifunction $(x,y) \to \partial_M g(x,y)$. The condition we use is that this inverse is pseudo-Lipschitz. The similarity with Zhang's condition is discussed in his paper. Since C_2 is a subspace that contains the range space of the inverse, this implies that $\Psi(v)$, the intersection of our two sets under perturbations, is pseudo-Lipschitz. Applying Theorem 4.1. gives the following result.

THEOREM 5.2. *Let $(\bar{x}, \bar{y})$ be a solution to* (5.1). *If the inverse of $(x,y) \to \partial_M g(x,y)$ is pseudo-Lipschitz at $(x^*, 0, \bar{x}, \bar{y})$ for all $(x^*, 0) \in \partial_M g(\bar{x}, \bar{y})$, then there exist $\lambda \geq 0$ and w^*, not both zero, and a z^* such that*

$$(z^*, 0) \in \partial_M g(x,y),$$

and there is $(x_1^*, y_1^*, 0, w^*) \in N_\ell \text{gph}\partial_M g(x,y)(\bar{x},\bar{y},z^*,0)$ with

$$-(x_1^*, y_1^*) \in \lambda\partial_\ell f(\bar{x},\bar{y}).$$

These two results show how the two generalized gradients, in combination, give nice results for a complicated situation. It is very difficult to give examples when the functions under consideration are not smooth. If the functions are all assumed to be C^2, one can rewrite Theorem 5.1.

COROLLARY 5.3. *Let $(\bar{x}, \bar{y})$ be a minimizer of problem (5.1) where f and g are C^2 functions. Then there are $\lambda \geq 0$ and w^*, not both zero, such that*

$$0 = \partial_y g(\bar{x}, \bar{y})$$

and

$$\lambda \nabla f(\bar{x}, \bar{y}) = Jg(\bar{x}, \bar{y})(0, w^*).$$

This is the desired necessary condition for a smooth unconstrained bilevel problem. The following is a simple example to show how this can be used.

Example 5.1. Let $f(x, y) = 4x^2 + y^2$ and

$$g(x, y) = \begin{cases} \cos(\sqrt{x^2 + y^2}) & \text{if } (x, y) \neq (0, 0), \\ 1 & \text{if } (x, y) = (0, 0). \end{cases}$$

The first condition in Corollary 5.3 yields that $\sqrt{x^2 + y^2} = k\pi/2$ for any k or $y = 0$. This includes both local minima and local maxima. Adding on the second condition gives the points $(k\pi/2, 0)$ and $(0, k\pi/2)$ for $k = 0, 1, 2, \ldots$. This is the best that one can do with first order necessary conditions. It is interesting to note that the number of possible minima is extremely large. This is an unfortunate, but very common, situation.

Acknowledgments. I wish to thank A. D. Ioffe, Philip Loewen, and Boris Mordukhovich for conversations concerning the LGG. I would also like to thank a referee whose general comments about revisions of papers were well taken.

REFERENCES

[1] G. ANADALINGAM AND T. L. FRIESZ, eds., *Hierarchical optimization*, Ann. Oper. Res., 34 (1992).

[2] J.-P. AUBIN, *Lipschitz behavior of solutions to convex minimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.

[3] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.

[4] F. H. CLARKE, *Methods of Dynamic and Nonsmooth Analysis*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 57, SIAM, Philadelphia, PA, 1989.

[5] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Classics in Appl. Math. 5, SIAM, Philadelphia, PA, 1990.

[6] S. DEMPE, *A necessary and a sufficient optimality condition for bilevel programming problems*, Optimization, 25 (1992), pp. 485–501.

[7] A. D. IOFFE, *Sous-différentielles approchées de fonctions numériques*, C. R. Acad. Sci. Paris Sér. I Math. 292 (1981), pp. 675–678.

[8] A. D. IOFFE, *Approximate subdifferentials and applications.* I. *The finite dimensional theory*, Trans. Amer. Math. Soc., 281 (1984), pp. 389–416.

[9] A. D. IOFFE, *Approximate subdifferentials and applications.* II, Mathematika, 33 (1986), pp. 111–128.

[10] A. D. IOFFE, *A Lagrange multiplier rule with small convex-valued subdifferentials for nonsmooth problems of mathematical programming involving equality and nonfunctional constraints*, Math. Programming, Ser. A, 58 (1993), pp. 137–145.

[11] A. Ja. Kruger and B. Sh. Mordukhovich, *Minimization of nonsmooth functionals in optimal control problems*, Izv. Akad. Nauk. SSSR Tekhn. Kibernet, (1978), pp. 176–183 (in Russian); Engrg. Cybernetics, 16 (1978), pp. 126–133 (in English).

[12] Michel and J.-P. Penot, *Subdifferential calculus for Lipschitzian and non-Lipschitzian functions*, C. R. Acad. Sci. Paris Sér. I Math., 298 (1984), pp. 269–272.

[13] B. Sh. Mordukhovich, *Maximum principle in the problem of time optimal response with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.

[14] B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988 (in Russian).

[15] B. S. Mordukhovich, *Complete characterization of openness, metric regularity and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–35.

[16] B. S. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.

[17] B. S. Mordukhovich, *Lipschitzian stability of constraint systems and generalized equations*, Nonlinear Anal., 222 (1994), pp. 173–206.

[18] R. T. Rockafellar, *Lipschitzian properties of multifunctions*, Nonlinear Anal., 9 (1985), pp. 665–698.

[19] J. S. Treiman, *Shrinking generalized gradients*, Nonlinear Anal., 12 (1988), pp. 1429–1450.

[20] J. S. Treiman, *Finite dimensional optimality conditions: B-gradients*, J. Optim. Theory Appl., 62 (1989), pp. 139–150.

[21] J. S. Treiman, *The linear nonconvex generalized gradient and Lagrange multipliers*, SIAM J. Optim., 5 (1995), pp. 670–680.

[22] J. S. Treiman, *The linear nonconvex generalized gradient*, in Proceedings of the First International Conference on Nonlinear Analysis, de Gruyten, Berlin, 1996.

[23] J. J. Ye, D. L. Zhu, and Q. Zhu, *Generalized bilevel programming problems*, Optimization 33 (1997), pp. 361–366.

[24] Rouxin Zhang, *Problems of hierarchical optimization in finite dimensions*, SIAM J. Optim., 4 (1994), pp. 521–536.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 5, pp. 1330–1345

A SPECTRAL TEST FOR OBSERVABILITY AND REACHABILITY OF TIME-VARYING SYSTEMS*

M. A. PETERS[†] AND P. A. IGLESIAS[†]

Abstract. A spectral test for the observability and reachability of linear time-invariant systems—the Popov–Belevitch–Hautus test—is well known and serves as a powerful characterization of these properties. In this paper it is shown that similar tests exist for linear time-varying systems. The test presented here involves a check over a subset of the spectrum of the weighted block shift known as the set of *almost eigenvalues*.

Key words. observability, operator-theoretic methods, controllability, exponential stability, time-varying systems, discrete-time systems

AMS subject classifications. 93B05, 93B07, 93B28, 93C50, 93C55, 93D20

PII. S0363012997321826

1. Introduction.

For linear time-invariant (LTI) systems described by a state-space realization, spectral tests exist to determine fundamental properties of the realization such as stability, reachability, and observability.

It has long been known that, for linear time-varying (LTV) systems, the location of eigenvalues of the "A" matrix in the realization does not furnish meaningful information regarding the system's stability. However, it was shown in [5, 8] that for discrete-time systems, stability could be characterized by the spectral radius of an operator arising from the state-space realization. In this paper we show how only part of the spectrum, the set of *almost eigenvalues*, is required to characterize stability.

For reachability and observability regarding LTI systems, the PBH (for Popov–Belevitch–Hautus; see [9]) test allows one to determine these properties based on the eigenvalues/eigenvectors of the A matrix. For LTV systems, however, no corresponding test in known. In this paper we show that a test analogous to the PBH test for discrete-time LTV systems can be formulated. Moreover, it will be reminiscent of the spectral test for stability developed in [5, 8].

The rest of the paper is organized as follows. In section 2 we describe the operator-theoretic setting that will be used throughout the paper. This setting is used in section 3 to review the characterization of stability developed in [5, 8]. Spectral tests for observability and detectability are considered in sections 4 and 5, respectively, whereas the dual results for reachability and stabilizability are found in section 6. Finally, concluding remarks are presented in section 7.

2. Preliminaries.

In this section we introduce the class of systems that will be considered, as well as the notation and some basic concepts used throughout the paper.

The notation is standard. All matrices and vectors are assumed real. For a matrix $M \in \mathbb{R}^{m\times n}$, M^T denotes its transpose. We consider the set of sequences from $\mathbb{Z} \to \mathbb{R}^n$. The subset of square summable sequences is denoted ℓ_2^n. This is a Hilbert

[*] Received by the editors May 23, 1997; accepted for publication (in revised form) June 8, 1998; published electronically June 18, 1999. This research was supported in part by NSF grant ECS-9309387.

http://www.siam.org/journals/sicon/37-5/32182.html

[†] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218 (pi@jhu.edu).

space with inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\ell_2^n} := \sum_{k=-\infty}^{\infty} x_k^T y_k$$

and norm

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{k=-\infty}^{\infty} |x_k|^2} < \infty\,.$$

Where the dimension of the underlying space is immaterial this will be omitted.

In the space of linear operators we will consider only the subset of bounded operators mapping ℓ_2^m to ℓ_2^p. A linear operator $\boldsymbol{G}$, mapping ℓ_2^m to ℓ_2^p, has an infinite-dimensional matrix representation $(\boldsymbol{G})_{k,j} = G_{k,j}$, where $G_{k,j} \in \mathbb{R}^{p\times m}$ for every $k, j \in \mathbb{Z}$. The subspace of operators which are memoryless is denoted by $\mathcal{M}$. These are the operators with block-diagonal matrix representations, and in general we will use the shorter notation $\boldsymbol{G} = \mathrm{diag}\{G_k\}$.

By the adjoint of an operator $\boldsymbol{G} : \ell_2^m \to \ell_2^p$ we denote the unique operator $\boldsymbol{G}^* : \ell_2^p \to \ell_2^m$ satisfying

$$\langle \boldsymbol{x}, \boldsymbol{G}\boldsymbol{y} \rangle_{\ell_2^p} = \langle \boldsymbol{G}^*\boldsymbol{x}, \boldsymbol{y} \rangle_{\ell_2^m}$$

$\forall\ \boldsymbol{x} \in \ell_2^m$ and $\boldsymbol{y} \in \ell_2^p$. Note that, for the operators that we consider, the adjoint is just the transpose of the infinite-dimensional matrix $\boldsymbol{G}$.

By the norm of an operator we mean the ℓ_2-induced operator norm

$$\|\boldsymbol{G}\| = \sup_{\boldsymbol{0}\neq \boldsymbol{w}\in\ell_2^m} \frac{\|\boldsymbol{G}\boldsymbol{w}\|_2}{\|\boldsymbol{w}\|_2}.$$

An operator $\boldsymbol{T} : \ell_2^n \to \ell_2^n$ is called invertible if there exists $\boldsymbol{S} : \ell_2^n \to \ell_2^n$ such that $\boldsymbol{TS} = \boldsymbol{ST} = \boldsymbol{I}$. The operator $\boldsymbol{S}$ is the inverse of $\boldsymbol{T}$, denoted $\boldsymbol{S} = \boldsymbol{T}^{-1}$.

We say that an operator $\boldsymbol{W} : \ell_2^n \to \ell_2^n$ is positive definite ($\boldsymbol{W} > \boldsymbol{0}$) if $\boldsymbol{W} = \boldsymbol{W}^*$ and if there exists an $\epsilon > 0$ such that $\langle \boldsymbol{x}, \boldsymbol{W}\boldsymbol{x} \rangle \geq \epsilon \langle \boldsymbol{x}, \boldsymbol{x} \rangle$ for all $\boldsymbol{x} \in \ell_2^n$. It can be checked that if $\boldsymbol{W} > \boldsymbol{0}$, then $\boldsymbol{W}$ is invertible and $\boldsymbol{W}^{-1} > \boldsymbol{0}$.

In the following discussion we will consider systems of the form

$$\begin{aligned} x_{k+1} &= A_k x_k + B_k u_k, \\ y_k &= C_k x_k. \end{aligned} \tag{2.1}$$

To simplify the notation in the sequel, we form block-diagonal operators using the state-space matrix sequences $\{A_k\}$, $\{B_k\}$, and $\{C_k\}$,

$$\boldsymbol{A} := \mathrm{diag}\{A_k\}, \qquad \boldsymbol{B} := \mathrm{diag}\{B_k\}, \qquad \boldsymbol{C} := \mathrm{diag}\{C_k\},$$

which are all memoryless operators. We will always assume that these sequences are uniformly bounded. One further operator is needed. Let $\boldsymbol{Z}$ be the usual unit advance operator, satisfying $(\boldsymbol{Z}\boldsymbol{x})_k = x_{k+1}$. Note that $\boldsymbol{Z}^*\boldsymbol{Z} = \boldsymbol{Z}\boldsymbol{Z}^* = \boldsymbol{I}$. Using these operators we can describe the system (2.1) as

$$\begin{aligned} \boldsymbol{Z}\boldsymbol{x} &= \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}, \\ \boldsymbol{y} &= \boldsymbol{C}\boldsymbol{x}. \end{aligned} \tag{2.2}$$

The operator $\boldsymbol{G}$, mapping $\boldsymbol{u} \in \ell_2^m$ to $\boldsymbol{y} \in \ell_2^p$, will often be called the system $\boldsymbol{G}$.

Finally, to abbreviate notation we will use the weighted shifts $\mathcal{A} := \boldsymbol{Z}^*\boldsymbol{A}$ and $\mathcal{B} := \boldsymbol{Z}^*\boldsymbol{B}$, and we introduce the spectral radius of an operator $\boldsymbol{T} : \ell_2^n \to \ell_2^n$, which is defined as

$$\rho(\boldsymbol{T}) := \lim_{n\to\infty} \|\boldsymbol{T}^n\|^{1/n}. \tag{2.3}$$

In the next section we will relate stability properties of the system (2.2) to conditions on the spectrum of the shifted operator $\mathcal{A}$.

3. Uniform exponential stability. We say that the operator $\boldsymbol{A}$ is uniformly exponentially stable (UES) if there exist constants $c > 0$ and $\beta \in [0, 1)$ such that $\forall\ k \in \mathbb{Z}$ and $l \in \mathbb{N}$

$$\|A_{k+l-1}A_{k+l-2}\cdots A_k\| \leq c\beta^l.$$

An equivalent expression is that $\forall\ k_0$ and x_{k_0}, the inequality

$$|x_k| \leq c\beta^{k-k_0}|x_{k_0}|$$

is satisfied $\forall\ k > k_0$.

If $\boldsymbol{A}$ is UES, we say that the corresponding system (2.2) is stable. The notion of uniform exponential stability can be characterized in terms of the spectral radius of the block-weighted shift $\mathcal{A}$, in a way which is reminiscent of the well-known result for LTI systems.

Lemma 3.1 (see [5, 8]). *Suppose that the operator $\boldsymbol{A} \in \mathcal{M}$. Then $\boldsymbol{A}$ is UES if and only if $\rho(\mathcal{A}) < 1$.*

For an LTI system (i.e., $A_k = A\ \forall\ k$) it can easily be shown that $\rho(\mathcal{A}) = \rho(A)$. Hence, an LTI system is stable if and only if all the eigenvalues of A have magnitude less than 1.

For time-varying systems, however, we will be checking for points λ in the spectrum of the block-weighted shift $\mathcal{A}$. Because this is an infinite-dimensional operator, which may not have eigenvalues, we will work with the slightly more general sets of *almost eigenvalues* and *almost eigenvectors* [2]. A number $\lambda \in \mathbb{C}$ is an almost eigenvalue of an operator $\boldsymbol{T}$ if there exists a sequence $\{\boldsymbol{x}_n\}_{n\geq 0}$, where each of the $\boldsymbol{x}_n$ is an element of ℓ_2 with $\|\boldsymbol{x}_n\|_2 = 1$, such that

$$\lim_{n\to\infty} \|\boldsymbol{T}\boldsymbol{x}_n - \lambda\boldsymbol{x}_n\|_2 = 0.$$

The sequence $\{\boldsymbol{x}_n\}$ is called an almost eigenvector corresponding to λ. We find the following.

Lemma 3.2. *If $\boldsymbol{A} \in \mathcal{M}$ is UES, then all almost eigenvalues of $\mathcal{A}$ have magnitude less than* 1.

Proof. The spectral radius of an operator is the maximum modulus of all elements of the spectrum of the operator. The set of almost eigenvalues of an operator is a subset of the spectrum [2]. Hence, if $\boldsymbol{A}$ is UES, from Lemma 3.1 all elements of the spectrum have magnitude less than 1, and hence so do all almost eigenvalues of $\mathcal{A}$. □

The result of this lemma follows directly from those of [5, 8]. The converse, however, is new. We now show that instead of checking for the maximum magnitude over all elements of the spectrum, we need only check the subset of the spectrum consisting of the almost eigenvalues.

LEMMA 3.3. *For $\boldsymbol{A} \in \mathcal{M}$, if all almost eigenvalues of $\mathcal{A}$ have magnitude less than* 1, *then $\boldsymbol{A}$ is UES.*

Proof. Suppose that $\boldsymbol{A}$ is not UES. Then

$$(\forall c > 0)(\forall \beta \in [0,1))(\exists k_0 \in \mathbb{Z})(\exists l \in \mathbb{N}) \qquad \|A_{k_0+l-1}A_{k_0+l-2}\cdots A_{k_0}\| > c\beta^l.$$

Now, for $n \in \mathbb{N}$ take $c = n$ and $\beta = 1 - 1/n$. Then there exists $k_0 \in \mathbb{Z}$, $l \in \mathbb{N}$, and $|x_{k_0}| = 1$ such that

$$|A_{k_0+l-1}\cdots A_{k_0}x_{k_0}| > n(1-1/n)^l. \tag{3.1}$$

The reason for taking $c = n$ is that it forces l to go to infinity in case n goes to infinity, which we will use later on. Define the number

$$r_n := |A_{k_0+l-1}\cdots A_{k_0}x_{k_0}|^{-1/l},$$

which is well defined, with $1/r_n > (1-1/n)n^{1/l} \geq 1 - 1/n$. With this notation we define the sequence

$$(\boldsymbol{x}_n)_k := \begin{cases} x_{k_0} & \text{for } k = k_0, \\ r_n^{k-k_0}A_{k-1}\cdots A_{k_0}x_{k_0} & \text{for } k = k_0+1,\ldots,k_0+l-1, \\ 0 & \text{elsewhere} \end{cases}$$

and $\bar{\boldsymbol{x}}_n := \boldsymbol{x}_n/\|\boldsymbol{x}_n\|_2$. Obviously $\|\bar{\boldsymbol{x}}_n\|_2 = 1$. For the sequence $\bar{\boldsymbol{x}}_n$ we can compute

$$\begin{aligned} \left\|\mathcal{A}\bar{\boldsymbol{x}}_n - \frac{1}{r_n}\bar{\boldsymbol{x}}_n\right\|_2^2 &= \frac{1}{r_n^2}\frac{|x_{k_0}|^2}{\|\boldsymbol{x}_n\|_2^2} + \frac{r_n^{2(l-1)}}{\|\boldsymbol{x}_n\|_2^2}|A_{k_0+l-1}\cdots A_{k_0}x_{k_0}|^2 \\ &= \frac{1}{r_n^2\|\boldsymbol{x}_n\|_2^2}\left(1 + r_n^{2l}|A_{k_0+l-1}\cdots A_{k_0}x_{k_0}|^2\right) \\ &= \frac{2}{r_n^2\|\boldsymbol{x}_n\|_2^2}, \end{aligned}$$

where we used the definition of r_n.

Since $\boldsymbol{A}$ is uniformly bounded, there exists an α such that $\|A_k\| < \alpha\ \forall\ k$. (Note that $\alpha \geq 1$; otherwise (3.1) cannot be satisfied for n large enough such that $1 - 1/n > \alpha$.) It is easy to check that $r_n > 1/\alpha$. Hence,

$$\left\|\mathcal{A}\bar{\boldsymbol{x}}_n - \frac{1}{r_n}\bar{\boldsymbol{x}}_n\right\|_2^2 < \frac{2\alpha^2}{\|\boldsymbol{x}_n\|_2^2}. \tag{3.2}$$

Now we see the importance of taking $c = n$. Namely, if n becomes larger, then also l is forced to be large in order to satisfy (3.1) since $\boldsymbol{A}$ is uniformly bounded, and it is easy to show that $l \to \infty$ as $n \to \infty$. Since $|x_{k_0}| = |x_{k_0+l}| = 1$ (by definition of r_n) and $\boldsymbol{A}$ is uniformly bounded, it can be shown that $\|\boldsymbol{x}_n\|_2 \to \infty$ as $n \to \infty$. Hence, the right-hand side of (3.2) goes to zero as n goes to infinity. Since

$$1/\alpha < r_n < 1 - 1/n$$

there exists a subsequence of $n \in \mathbb{N}$ for which r_n converges, say $r_n \to r$ $(n \to \infty)$, and we see that $1/r$ is an almost eigenvalue of $\mathcal{A}$. Since $1/r_n > 1 - 1/n$, we have $1/r \geq 1$. Therefore $\mathcal{A}$ has an almost eigenvalue with magnitude greater than or equal to 1. □

Note that the construction in the proof of Lemma 3.3 leads to an almost eigenvalue of $\mathcal{A}$ which is real and nonnegative (namely $\lambda = 1/r$). In fact, we can show that it is sufficient only to consider real nonnegative almost eigenvalues.

LEMMA 3.4. *If λ is an almost eigenvalue of $\mathcal{A}$, then $|\lambda|$ is also an almost eigenvalue of $\mathcal{A}$, and it has a corresponding real almost eigenvector.*

Proof. Assume $\|\mathcal{A}\boldsymbol{x}_n - \lambda \boldsymbol{x}_n\|_2 \to 0$ as $n \to \infty$ for $\|\boldsymbol{x}_n\|_2 = 1$. Hence, with $\lambda = |\lambda| e^{i\theta}$, we have

$$\sum_{k=-\infty}^{\infty} \left| A_k x_{n_k} - |\lambda| e^{i\theta} x_{n_{k+1}} \right|^2 \to 0 \quad (n \to \infty).$$

This is equivalent to

$$\sum_{k=-\infty}^{\infty} \left| A_k e^{ik\theta} x_{n_k} - |\lambda| e^{i(k+1)\theta} x_{n_{k+1}} \right|^2 \to 0 \quad (n \to \infty), \tag{3.3}$$

which shows that $\tilde{\boldsymbol{x}}_n$, given by $\tilde{x}_{n_k} = e^{ik\theta} x_{n_k}$, is an almost eigenvector for the almost eigenvalue $|\lambda|$. Note that $\|\tilde{\boldsymbol{x}}_n\|_2 = \|\boldsymbol{x}_n\|_2 = 1$. To construct a real almost eigenvector, we can separate the real and the imaginary parts in (3.3), which gives

$$\sum_{k=-\infty}^{\infty} \left| A_k \mathrm{Re}(\tilde{x}_{n_k}) - |\lambda| \mathrm{Re}(\tilde{x}_{n_{k+1}}) \right|^2 + \left| A_k \mathrm{Im}(\tilde{x}_{n_k}) - |\lambda| \mathrm{Im}(\tilde{x}_{n_{k+1}}) \right|^2 \to 0$$

as n goes to infinity. Hence, we can take either

$$\boldsymbol{z}_{1n} := \frac{\mathrm{Re}(\tilde{\boldsymbol{x}}_n)}{\|\mathrm{Re}(\tilde{\boldsymbol{x}}_n)\|_2} \quad \text{or} \quad \boldsymbol{z}_{2n} := \frac{\mathrm{Im}(\tilde{\boldsymbol{x}}_n)}{\|\mathrm{Im}(\tilde{\boldsymbol{x}}_n)\|_2}$$

in order to have a real almost eigenvector. Note that, since $\|\mathrm{Re}(\tilde{\boldsymbol{x}}_n)\|_2^2 + \|\mathrm{Im}(\tilde{\boldsymbol{x}}_n)\|_2^2 = 1$, at least one of them is well defined. □

Remark. In a similar way we can show that $|\lambda| e^{i\psi}$ is an almost eigenvalue for every $\psi \in [0, 2\pi)$. This generalizes the well-known result that for the shift operator $\boldsymbol{Z}^*$ (i.e., $\boldsymbol{A} = \boldsymbol{I}$), the whole unit circle is part of the spectrum.

Example 3.5. We will illustrate the results of Lemma 3.4. In fact, by choosing an LTI system as our example, we show how the result considered is new even for the LTI setting. Consider the LTI system with

$$A_k = A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

In this case $\lambda = i$ is an eigenvalue of the matrix A with eigenvector $x = \left[\begin{smallmatrix} 1 \\ i \end{smallmatrix}\right]$, and it is also an almost eigenvalue of $\mathcal{A}$ with almost eigenvector $\boldsymbol{x}_n$ given by

$$x_{n_k} = \begin{cases} \dfrac{1}{\sqrt{2(2n+1)}} \begin{bmatrix} 1 \\ i \end{bmatrix} & \text{for } k = -n, \dots, n-1, n, \\ \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \text{for } k \neq -n, \dots, n-1, n. \end{cases}$$

Following the construction in the lemma we find a real almost eigenvector $\boldsymbol{z}_n$, corresponding to $|\lambda| = 1$, given by

$$z_{n_0} = \frac{1}{\sqrt{(2n+1)}} \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad z_{n_1} = \frac{1}{\sqrt{(2n+1)}} \begin{bmatrix} 0 \\ -1 \end{bmatrix},$$

$$z_{n_2} = \frac{1}{\sqrt{(2n+1)}} \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \qquad z_{n_3} = \frac{1}{\sqrt{(2n+1)}} \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

$$z_{n_k} = \begin{cases} z_{n_{k \bmod 4}} & \text{for } k = -n, \dots, n-1, n, \\ \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \text{for } k \neq -n, \dots, n-1, n. \end{cases}$$

Remark. The computation of the generalized eigenvalues/eigenvectors will in general not be trivial. Some general comments on the computation of the spectrum of weighted shifts such as those considered here can be found in [10].

Summarizing the results, we have the following theorem.

THEOREM 3.6. *The following statements are equivalent:*

(i) $\boldsymbol{A}$ *is UES.*

(ii) $\rho(\mathcal{A}) < 1$.

(iii) *All almost eigenvalues of* $\mathcal{A}$ *have magnitude less than* 1.

These statements are the time-varying analogues of the well-known LTI results. In the next section we will outline the concept of uniform observability and discuss a PBH-type test which checks whether a system satisfies this property.

4. Uniform observability. Recall that an LTI system is said to be observable if different initial conditions give rise to different outputs. Specifically, for the unforced LTI equation

$$\begin{aligned} x_{k+1} &= Ax_k, \\ y_k &= Cx_k, \end{aligned}$$

the pair (C, A) is observable if $CA^l x_0 = CA^l \bar{x}_0$ for $l = 0, 1, \dots, n-1$ implies that $x_0 = \bar{x}_0$.

As is well known, n is the maximum number of steps needed to recover the initial state from the output, although it might be possible in fewer.

LEMMA 4.1. *An LTI system is observable if and only if*

$$\left.\begin{aligned} Ax &= \lambda x \\ Cx &= 0 \end{aligned}\right\} \Longrightarrow x = 0.$$

That is, A has no unobservable eigenvalues.

Note that it is necessary to perform the PBH test only at eigenvalue/eigenvector pairs of the matrix A.

In the time-varying case, we consider systems of the form

$$\begin{aligned} x_{k+1} &= A_k x_k, \\ y_k &= C_k x_k. \end{aligned}$$

A corresponding notion of observability for time-varying systems can be formulated based on the idea that for an observable system, different initial conditions should lead to different outputs. This leads to the concept of uniform l-step observability [9].

DEFINITION 4.2. *The pair $(\boldsymbol{C}, \boldsymbol{A})$ is called uniformly l-step observable if there exists $\epsilon > 0$ such that $\forall\ k_0$, x_{k_0}, and $\bar{x}_{k_0}$, the inequality*

$$\sum_{k=k_0}^{k_0+l-1} |y_k - \bar{y}_k|^2 \geq \epsilon |x_{k_0} - \bar{x}_{k_0}|^2$$

is satisfied.

The uniform requirement ensures that the same ϵ will be suitable for all starting points k_0.

Example 4.3. Consider a time-varying system with $A_k = 0 \ \forall \ k$ and

$$C_k = \begin{cases} 1 & \text{for } k \leq 0, \\ 1/k & \text{for } k > 0. \end{cases}$$

Based on the observation of the output, at any time k, the initial state, can be recovered in one step, regardless of the particular value of k. Nevertheless, for $k_0 > 0$ we have that

$$y_k = \begin{cases} 1/k_0 x_{k_0} & \text{for } k = k_0, \\ 0 & \text{elsewhere.} \end{cases}$$

It follows that the difference in output $y_k - \bar{y}_k$ can be made arbitrarily small (in an ℓ_2 sense) compared to $x_{k_0} - \bar{x}_{k_0}$ by choosing k_0 large.

For LTI systems, we can always choose $l = n$. For LTV systems, this is not true.

Definition 4.4. *The pair $(\boldsymbol{C}, \boldsymbol{A})$ is called uniformly observable if there exists a positive integer l for which the system is uniformly l-step observable.*

Remark. It is important that l be finite. For example if $A_k = 1 \ \forall \ k$ and

$$C_k = \begin{cases} -1/k & \text{for } k < 0, \\ 1 & \text{for } k \geq 0, \end{cases}$$

it is easy to check that the pair $(\boldsymbol{C}, \boldsymbol{A})$ is not uniformly l-step observable for any finite l, but it is for $l = \infty$.

Since the system is linear, it is easy to check that the definition of uniform l-step observability is dependent only on the difference between the initial states; thus we can assume without loss of generality that $\bar{x}_{k_0} = 0$ and $\bar{y}_k = 0 \ \forall \ k$.

The l-step observability Gramian operator is the block-diagonal operator

$$\boldsymbol{M}_l := \sum_{j=0}^{l-1} (\mathcal{A}^*)^j \boldsymbol{C}^* \boldsymbol{C} \mathcal{A}^j. \tag{4.1}$$

Using this operator we find the following.

Lemma 4.5. *The pair $(\boldsymbol{C}, \boldsymbol{A})$ is uniformly l-step observable if and only if $\boldsymbol{M}_l > \boldsymbol{0}$.*

Proof. Since

$$\sum_{k=k_0}^{k_0+l-1} |y_k|^2 = \sum_{k=k_0}^{k_0+l-1} |C_k A_{k-1} \cdots A_{k_0} x_{k_0}|^2 = x_{k_0}^T (\boldsymbol{M}_l)_{k_0} x_{k_0}$$

for arbitrary k_0 and initial condition x_{k_0}, the result is immediate. □

If the system is time-invariant, all the block diagonal elements of $\boldsymbol{M}_l$ are the same. These elements have the form

$$(\boldsymbol{M}_l)_k = \sum_{r=k}^{k+l-1} (A^{r-k})^T C^T C A^{r-k} = \sum_{p=0}^{l-1} (A^p)^T C^T C A^p.$$

For an LTI system it is well known that this matrix is positive definite for a finite l if and only if the system is observable. Hence, we have the following corollary.

COROLLARY 4.6. *For an LTI system, the pair* $(\boldsymbol{C}, \boldsymbol{A})$ *is uniformly observable if and only if the pair* (C, A) *is observable.*

Hence, the LTI results can be recovered directly from results we find for time-varying systems.

Now, we are interested in testing whether a given pair $(\boldsymbol{C}, \boldsymbol{A})$ is uniformly observable, in terms of a PBH test analogous to Lemma 4.1 for LTI systems. This can be done by considering the set of almost eigenvalues corresponding to the weighted shift operator $\mathcal{A}$.

LEMMA 4.7. *If the pair* $(\boldsymbol{C}, \boldsymbol{A})$ *is uniformly observable, then there exists no almost eigenvalue of* $\mathcal{A}$ *for which the corresponding almost eigenvector* $\boldsymbol{x}_n$ *satisfies* $\|\boldsymbol{C}\boldsymbol{x}_n\|_2 \to 0$ *as* $n \to 0$.

Proof. If the pair $(\boldsymbol{C}, \boldsymbol{A})$ is uniformly observable, there exists $l \in \mathbb{N}$ for which $\boldsymbol{M}_l > \boldsymbol{0}$. Now, suppose that λ is an almost eigenvalue of $\mathcal{A}$ with corresponding almost eigenvector $\boldsymbol{x}_n$. Hence, $\|\boldsymbol{x}_n\|_2 = 1$, and we can assume without loss of generality that $\boldsymbol{x}_n$ is such that

$$\mathcal{A}\boldsymbol{x}_n = \lambda \boldsymbol{x}_n + \boldsymbol{q}_n \quad \text{where } \|\boldsymbol{q}_n\|_2 \le 1/n.$$

Also assume that $\|\boldsymbol{C}\boldsymbol{x}_n\|_2 \to 0$ as $n \to 0$, in particular, that $\|\boldsymbol{C}\boldsymbol{x}_n\|_2 \le 1/n$. For $j \ge 1$ we find that

$$\begin{aligned}\|\boldsymbol{C}\mathcal{A}^j\boldsymbol{x}_n\|_2 &= \|\boldsymbol{C}\mathcal{A}^{j-1}(\lambda\boldsymbol{x}_n + \boldsymbol{q}_n)\|_2 \\ &\le |\lambda|\,\|\boldsymbol{C}\mathcal{A}^{j-1}\boldsymbol{x}_n\|_2 + \|\boldsymbol{C}(\boldsymbol{Z}^*\boldsymbol{A})^{j-1}\|\,\|\boldsymbol{q}_n\|_2 \\ &\le |\lambda|\,\|\boldsymbol{C}\mathcal{A}^{j-1}\boldsymbol{x}_n\|_2 + \|\boldsymbol{C}\|\,\|\boldsymbol{Z}^*\boldsymbol{A}\|^{j-1}\,\frac{1}{n}.\end{aligned}$$

By recursion, it follows that $\forall\ j \ge 1$

$$\|\boldsymbol{C}\mathcal{A}^j\boldsymbol{x}_n\|_2 \le \frac{|\lambda|^j}{n} + \frac{\|\boldsymbol{C}\|}{n}\sum_{r=0}^{j-1}\|\mathcal{A}\|^{j-r-1}|\lambda|^r.$$

Now, by defining $\gamma := \max\{1, \|\mathcal{A}\|^{l-1}\}$, we see that for $1 \le j \le l-1$ this is bounded above by

$$\left\|\boldsymbol{C}\mathcal{A}^j\boldsymbol{x}_n\right\|_2 \le \frac{|\lambda|^j}{n} + \frac{\gamma\,\|\boldsymbol{C}\|}{n}\sum_{r=0}^{j-1}|\lambda|^r.$$

This yields

$$\begin{aligned}\langle \boldsymbol{x}_n, \boldsymbol{M}_l\boldsymbol{x}_n\rangle &= \sum_{j=0}^{l-1}\|\boldsymbol{C}\mathcal{A}^j\boldsymbol{x}_n\|_2 \\ &\le \frac{1}{n}\sum_{j=0}^{l-1}|\lambda|^j + \frac{\gamma\,\|\boldsymbol{C}\|}{n}\sum_{j=1}^{l-1}\sum_{r=0}^{j-1}|\lambda|^r \\ &= \frac{1}{n}\frac{1-|\lambda|^l}{1-|\lambda|} + \frac{\gamma\,\|\boldsymbol{C}\|}{n}\frac{1-2|\lambda|+|\lambda|^l}{(1-|\lambda|)^2}\end{aligned}$$

whenever $|\lambda| \ne 1$. For $|\lambda| = 1$ it is easy to compute

$$\langle \boldsymbol{x}_n, \boldsymbol{M}_l\boldsymbol{x}_n\rangle = \sum_{j=0}^{l-1}\|\boldsymbol{C}\mathcal{A}^j\boldsymbol{x}_n\|_2 \le \frac{l}{n} + \frac{\|\boldsymbol{C}\|\,\gamma}{n}\frac{l(l+1)}{2}.$$

It follows that

$$\langle \boldsymbol{x}_n, \boldsymbol{M}_l \boldsymbol{x}_n \rangle \to 0 \quad (n \to \infty)$$

and since $\boldsymbol{M}_l > \boldsymbol{0}$ we see that $\|\boldsymbol{x}_n\|_2 \to 0$, which contradicts the assumption that $\boldsymbol{x}_n$ is an almost eigenvector. □

The next result shows that the converse is also true.

LEMMA 4.8. *If there are no almost eigenvalues of $\mathcal{A}$ for which the corresponding almost eigenvector $\boldsymbol{x}_n$ satisfies $\|\boldsymbol{C}\boldsymbol{x}_n\|_2 \to 0$ as $n \to \infty$, then the pair $(\boldsymbol{C}, \boldsymbol{A})$ is uniformly observable.*

Proof. Suppose that $\mathcal{A}$ has no almost eigenvalues for which the corresponding eigenvector satisfies $\|\boldsymbol{C}\boldsymbol{x}_n\|_2 \to 0$ as $n \to \infty$. Then

$$(\exists \gamma > 0)(\forall \lambda \in \mathbb{C})(\forall \|\boldsymbol{x}\|_2 = 1) \qquad \|\mathcal{A}\boldsymbol{x} - \lambda \boldsymbol{x}\|_2^2 + \|\boldsymbol{C}\boldsymbol{x}\|_2^2 \geq \gamma.$$

This implies that the same statement is also true for all $\boldsymbol{x}$ with $\|\boldsymbol{x}\|_2 \geq 1$. Therefore

$$\|\mathcal{A}\boldsymbol{x} - \lambda \boldsymbol{x}\|_2^2 + \|\boldsymbol{C}\boldsymbol{x}\|_2^2 < \gamma \Longrightarrow \|\boldsymbol{x}\|_2 < 1.$$

Define $\delta := \gamma/2 > 0$. Then for every λ

$$\left.\begin{array}{rcl} \|\mathcal{A}\boldsymbol{x} - \lambda \boldsymbol{x}\|_2^2 & < & \delta \\ \|\boldsymbol{C}\boldsymbol{x}\|_2^2 & < & \delta \end{array}\right\} \Rightarrow \|\boldsymbol{x}\|_2^2 < 1. \tag{4.2}$$

Now, assume that the pair $(\boldsymbol{C}, \boldsymbol{A})$ is not uniformly observable. Then

$$(\forall l \in \mathbb{N})(\forall \epsilon > 0)(\exists k_0, |x_{k_0}| = 1) \qquad \sum_{k=k_0}^{k_0+l-1} |y_k|^2 < \epsilon. \tag{4.3}$$

For $l \in \mathbb{N}$, define $\epsilon_l := \delta l^{-2(l-1)}$. According to (4.3) we find a sequence $k_0(l)$ and $|x_{k_0(l)}| = 1$. Similar to the proof of Lemma 3.3, we define the sequence

$$r_l := \min \left\{ |A_{k_0(l)+l-1} \cdots A_{k_0(l)} x_{k_0(l)}|^{-1/l}, l \right\},$$

where $r_l := l$ if the inverse is not well defined. With this notation we define the sequence

$$(\boldsymbol{x}_l)_k := \left\{ \begin{array}{cl} x_{k_0(l)} & \text{for } k = k_0(l), \\ r_l^{k-k_0(l)} A_{k-1} \cdots A_{k_0(l)} x_{k_0(l)} & \text{for } k = k_0(l)+1, \cdots, k_0(l)+l-1, \\ 0 & \text{elsewhere} \end{array} \right.$$

and $\bar{\boldsymbol{x}}_l := \boldsymbol{x}_l / \|\boldsymbol{x}_l\|_2$. For the sequence $\bar{\boldsymbol{x}}_l$ we will compute the norms in (4.2). Obviously $r_l > 0 \ \forall \ l \in \mathbb{N}$, and it is easy to get

$$\begin{aligned} \left\| \mathcal{A}\bar{\boldsymbol{x}}_l - \frac{1}{r_l}\bar{\boldsymbol{x}}_l \right\|_2^2 &= \frac{1}{r_l^2} \frac{|x_{k_0(l)}|^2}{\|\boldsymbol{x}_l\|_2^2} + \frac{r_l^{2(l-1)}}{\|\boldsymbol{x}_l\|_2^2} |A_{k_0(l)+l-1} \cdots A_{k_0(l)} x_{k_0(l)}|^2 \\ &= \frac{1}{r_l^2 \|\boldsymbol{x}_l\|_2^2} \left(1 + r_l^{2l} |A_{k_0(l)+l-1} \cdots A_{k_0(l)} x_{k_0(l)}|^2\right), \\ &\leq \frac{2}{r_l^2 \|\boldsymbol{x}_l\|_2^2}, \end{aligned} \tag{4.4}$$

where the inequality follows by definition of r_l.

We consider two distinct possibilities. Suppose that there exists an $l^2 > 2/\delta$ for which $r_l = l$. It follows that the bound in (4.4) yields

$$\left\| \mathcal{A}\bar{\boldsymbol{x}}_l - \frac{1}{r_l}\bar{\boldsymbol{x}}_l \right\|_2^2 < \delta$$

since $\|\boldsymbol{x}_l\|_2^2 \geq |x_{k_0(l)}|^2 = 1$.

Now, suppose that no such l exists equivalently, that

$$r_l = |A_{k_0(l)+l-1} \cdots A_{k_0(l)} x_{k_0(l)}|^{-1/l} < l$$

$\forall$ l large enough. As in the proof of Lemma 3.3 we get that $r_l > 1/\alpha$, where $\|A_k\| < \alpha \; \forall \; k$. Hence, in (4.4) we have

$$\left\| \mathcal{A}\bar{\boldsymbol{x}}_l - \frac{1}{r_l}\bar{\boldsymbol{x}}_l \right\|_2^2 < \frac{2\alpha^2}{\|\boldsymbol{x}_l\|_2^2}.$$

However, with this solution for r_l, it follows by construction of the sequence $\boldsymbol{x}_l$ and the boundedness of $\boldsymbol{A}$ that $\|\boldsymbol{x}_l\|_2 \to \infty$ as $l \to \infty$ (see the proof of Lemma 3.3). Thus there exists an l such that $\|\boldsymbol{x}_l\|_2^2 > 2\alpha^2/\delta$, ensuring that (4.4) is bounded above by δ. Furthermore,

$$\begin{aligned} \|\boldsymbol{C}\bar{\boldsymbol{x}}_l\|_2^2 &= \sum_{k=k_0(l)}^{k_0(l)+l-1} \frac{r_l^{2(k-k_0(l))}}{\|\boldsymbol{x}_l\|_2^2} |y_k|^2 \\ &\leq \frac{\max\{1, r_l^{2(l-1)}\}}{\|\boldsymbol{x}_l\|_2^2} \sum_{k=k_0(l)}^{k_0(l)+l-1} |y_k|^2. \end{aligned}$$

Since $|x_{k_0(l)}| = 1$, we have that $\|\boldsymbol{x}_l\|_2^2 \geq 1$, and using the construction of ϵ_l we get

$$\begin{aligned} \|\boldsymbol{C}\bar{\boldsymbol{x}}_l\|_2^2 &< \max\left\{1, r_l^{2(l-1)}\right\} \epsilon_l \\ &= \max\left\{1, r_l^{2(l-1)}\right\} \delta l^{-2(l-1)} \\ &\leq \delta, \end{aligned}$$

where the last step follows since $r_l \leq l$. Thus we have found $l \in \mathbb{N}$ such that $\|\mathcal{A}\bar{\boldsymbol{x}}_l - \frac{1}{r_l}\bar{\boldsymbol{x}}_l\|_2^2 < \delta$ and $\|\boldsymbol{C}\bar{\boldsymbol{x}}_l\|_2^2 < \delta$. On the other hand, however, we know that $\|\bar{\boldsymbol{x}}_l\|_2^2 = 1$, contradicting (4.2). □

Remark. We see that, by letting δ tend to zero, the construction in the proof of Lemma 4.8 leads to an almost eigenvalue of $\mathcal{A}$ which is real and nonnegative; namely $\lambda = 1/r$, where $r_l \to r$ for a subsequence of l for which this limit exists. (Note that if $r_l = l$ for l large enough, this corresponds to an almost eigenvalue $\lambda = 0$.) We have seen this in Lemma 3.4, where we showed that we have to consider only real nonnegative almost eigenvalues of $\mathcal{A}$ for checking uniform exponential stability of $\boldsymbol{A}$. Following the construction of real almost eigenvectors in the proof of Lemma 3.4, we see that $\|\mathrm{Re}(\tilde{\boldsymbol{x}}_n)\|_2^2 + \|\mathrm{Im}(\tilde{\boldsymbol{x}}_n)\|_2^2 = 1$. Hence, one of these two has a squared norm larger than or equal to 1/2. By choosing this one, it is immediate that the resulting $\boldsymbol{z}_n$ satisfies

$$\|\boldsymbol{C}\boldsymbol{z}_n\|_2^2 \leq 2\|\boldsymbol{C}\boldsymbol{x}_n\|_2^2.$$

This shows that, in the PBH test stated in the previous lemmas, it is sufficient to consider real nonnegative almost eigenvalues only.

Summarizing, we have the following theorem.

THEOREM 4.9. *The following statements are equivalent:*

(i) *The pair $(\boldsymbol{C}, \boldsymbol{A})$ is uniformly observable.*

(ii) *There exists $l \in \mathbb{N}$ such that $\boldsymbol{M}_l > \boldsymbol{0}$.*

(iii) *There exists no almost eigenvalue of $\mathcal{A}$ for which the corresponding almost eigenvector $\boldsymbol{x}_n$ satisfies $\|\boldsymbol{C}\boldsymbol{x}_n\|_2 \to 0$ as $n \to \infty$.*

It can be shown that if the pair $(\boldsymbol{C}, \boldsymbol{A})$ is uniformly observable, there exists a bounded memoryless operator $\boldsymbol{H}$ such that $\boldsymbol{A} + \boldsymbol{H}\boldsymbol{C}$ is UES.

LEMMA 4.10 (see [6]). *If the pair $(\boldsymbol{C}, \boldsymbol{A})$ is uniformly observable, then the operator*

$$\boldsymbol{H} := -\boldsymbol{A}\mathcal{A}^l \boldsymbol{M}_{l+1}^{-1} (\mathcal{A}^*)^l \boldsymbol{C}^*,$$

where l is such that (ii) *in Theorem* 4.9 *is satisfied, guarantees that $\boldsymbol{A}_H := \boldsymbol{A} + \boldsymbol{H}\boldsymbol{C}$ is UES.*

Next we will discuss the concept of uniform detectability, which is less restrictive than uniform observability.

5. Uniform detectability. For uniform detectability, the requirement that different inputs result in significantly different outputs will be restricted to those initial states for which the state itself does not decay fast enough. Formally, this yields the definition as introduced in [1, 3].

DEFINITION 5.1. *The pair $(\boldsymbol{C}, \boldsymbol{A})$ is called uniformly detectable if there exist $l, p \in \mathbb{N}$ and constants $\gamma \in [0, 1)$ and $\epsilon > 0$ such that, whenever*

$$|x_{k_0+p}| \geq \gamma |x_{k_0}|$$

for some k_0 and x_{k_0}, then

$$\sum_{k=k_0}^{k_0+l-1} |y_k|^2 \geq \epsilon |x_{k_0}|^2.$$

The idea is that, when a zero-input trajectory starting at x_{k_0} fails to converge much toward the origin, then x_{k_0} should be observable to a minimum level.

In terms of the observability Gramian (4.1), this definition can be rewritten as

$$x_{k_0}^T \left((\mathcal{A}^*)^p \mathcal{A}^p \right)_{k_0} x_{k_0} \geq \gamma |x_{k_0}|^2 \quad \Longrightarrow \quad x_{k_0}^T (\boldsymbol{M}_l)_{k_0} x_{k_0} \geq \epsilon |x_{k_0}|^2.$$

If the pair $(\boldsymbol{C}, \boldsymbol{A})$ is uniformly observable, then $\boldsymbol{M}_l > \boldsymbol{0}$ for some $l \in \mathbb{N}$ (see Lemma 4.5). Hence, we immediately have the following lemma.

LEMMA 5.2. *If the pair $(\boldsymbol{C}, \boldsymbol{A})$ is uniformly observable, then the pair $(\boldsymbol{C}, \boldsymbol{A})$ is also uniformly detectable.*

The concept of uniform detectability has been well investigated [1, 3]. A result that will prove to be useful is the following.

LEMMA 5.3 (see [1]). *The pair $(\boldsymbol{C}, \boldsymbol{A})$ is uniformly detectable if and only if there exists an operator $\boldsymbol{H} \in \mathcal{M}$ for which $\boldsymbol{A} + \boldsymbol{H}\boldsymbol{C}$ is UES.*

Remark. In case the pair $(\boldsymbol{C}, \boldsymbol{A})$ is uniformly observable, it is possible to define an operator $\boldsymbol{H}$ guaranteeing that $\boldsymbol{A} + \boldsymbol{H}\boldsymbol{C}$ is UES in terms of the observability Gramian (see Lemma 4.10). In the case of uniform detectability, however, a Riccati operator equation has to be solved to find a stabilizing feedback (see [3]), which is much less appealing of course.

For LTI systems, the matrix pair (C, A) is defined to be detectable if there exists a matrix H such that $A + HC$ is stable. Hence, the following is an immediate result from Lemma 5.3.

COROLLARY 5.4. *For an LTI system the pair* (C, A) *is detectable if and only if the corresponding pair* $(\boldsymbol{C}, \boldsymbol{A})$ *is uniformly detectable.*

Now, we are interested in testing uniform detectability of a system in terms of the almost eigenvalues of the block-weighted shift $\mathcal{A}$. For LTI systems the PBH test is given by the following result.

LEMMA 5.5. *An LTI system is detectable if and only if* $\forall\, |\lambda| \geq 1$

$$\left.\begin{array}{ll} Ax & = \lambda x \\ Cx & = 0 \end{array}\right\} \Longrightarrow x = 0.$$

That is, A has no unobservable eigenvalues with magnitude greater than or equal to 1.

For time-varying systems we find similar results.

LEMMA 5.6. *If the pair* $(\boldsymbol{C}, \boldsymbol{A})$ *is uniformly detectable, then* $\mathcal{A}$ *has no almost eigenvalues with magnitude greater than or equal to* 1, *for which the corresponding eigenvector* $\boldsymbol{x}_n$ *satisfies* $\|\boldsymbol{C}\boldsymbol{x}_n\|_2 \to 0$ *as* $n \to \infty$.

Proof. Suppose that the pair $(\boldsymbol{C}, \boldsymbol{A})$ is uniformly detectable. According to Lemma 5.3 there exists an operator $\boldsymbol{H}$ such that $\boldsymbol{A}_H := \boldsymbol{A} + \boldsymbol{H}\boldsymbol{C}$ is UES. Now, let λ be any almost eigenvalue of $\mathcal{A}$, with corresponding eigenvector $\boldsymbol{x}_n$ with $\|\boldsymbol{x}_n\|_2 = 1$. Suppose that $\|\boldsymbol{C}\boldsymbol{x}_n\|_2 \to 0$ as $n \to \infty$. Then

$$\begin{aligned} \|\mathcal{A}_H \boldsymbol{x}_n - \lambda \boldsymbol{x}_n\|_2 &= \|\mathcal{A}\boldsymbol{x}_n - \lambda \boldsymbol{x}_n + \boldsymbol{Z}^* \boldsymbol{H}\boldsymbol{C}\boldsymbol{x}_n\|_2 \\ &\leq \|\mathcal{A}\boldsymbol{x}_n - \lambda \boldsymbol{x}_n\|_2 + \|\boldsymbol{Z}^* \boldsymbol{H}\boldsymbol{C}\boldsymbol{x}_n\|_2 \\ &\leq \|\mathcal{A}\boldsymbol{x}_n - \lambda \boldsymbol{x}_n\|_2 + \|\boldsymbol{H}\|\|\boldsymbol{C}\boldsymbol{x}_n\|_2, \end{aligned}$$

since $\|\boldsymbol{Z}^*\| = 1$. We know that λ is an almost eigenvector of $\mathcal{A}$, $\boldsymbol{H}$ is bounded and $\|\boldsymbol{C}\boldsymbol{x}_n\|_2 \to 0$ as $n \to \infty$; hence, it follows that this expression goes to zero as n goes to infinity. Thus λ is an almost eigenvalue of $\mathcal{A}_H$, and since $\boldsymbol{A}_H$ is UES, it follows from Lemma 3.2 that λ has magnitude less than 1. □

LEMMA 5.7. *If* $\mathcal{A}$ *has no almost eigenvalues with magnitude greater than or equal to* 1 *for which the corresponding eigenvector* $\boldsymbol{x}_n$ *satisfies* $\|\boldsymbol{C}\boldsymbol{x}_n\|_2 \to 0$ *as* $n \to \infty$, *then the pair* $(\boldsymbol{C}, \boldsymbol{A})$ *is uniformly detectable.*

Proof. The proof is a combination of the proofs of Lemmas 3.3 and 4.8. Suppose that $\mathcal{A}$ has no almost eigenvalues with magnitude larger than or equal to 1 for which the corresponding eigenvector satisfies $\|\boldsymbol{C}\boldsymbol{x}_n\|_2 \to 0$ as $n \to \infty$. Then, as in Lemma 4.8, there exists $\delta > 0$ such that for every $|\lambda| \geq 1$

$$\left.\begin{array}{rcl} \|\mathcal{A}\boldsymbol{x} - \lambda \boldsymbol{x}\|_2^2 & < & \delta \\ \|\boldsymbol{C}\boldsymbol{x}\|_2^2 & < & \delta \end{array}\right\} \Longrightarrow \|\boldsymbol{x}\|_2^2 < 1. \tag{5.1}$$

Now, assume that the pair $(\boldsymbol{C}, \boldsymbol{A})$ is not uniformly detectable. Then

(5.2)

$$(\forall l, p \in \mathbb{N})(\forall \gamma \in [0, 1))(\forall \epsilon > 0)(\exists k_0, |x_{k_0}| = 1) \qquad |x_{k_0+p}| \geq \gamma \quad \text{and} \quad \sum_{k=k_0}^{k_0+l-1} |y_k|^2 < \epsilon.$$

For $l \in \mathbb{N}$, define $\epsilon_l := \delta(1-1/l)^{2(l-1)}$ and $\gamma_l := (1-1/l)^l$ and take $p = l$. According to (5.2) we find a sequence $k_0(l)$ and $|x_{k_0(l)}| = 1$. Again we define the sequence

$$r_l := |A_{k_0(l)+l-1} \cdots A_{k_0(l)} x_{k_0(l)}|^{-1/l},$$

which is well defined with $r_l < 1/(1-1/l)$. With this notation we define the sequence

$$(\boldsymbol{x}_l)_k := \begin{cases} x_{k_0(l)} & \text{for } k = k_0(l), \\ r_l^{k-k_0(l)} A_{k-1} \cdots A_{k_0(l)} x_{k_0(l)} & \text{for } k = k_0(l)+1, \ldots, k_0(l)+l-1, \\ 0 & \text{elsewhere} \end{cases}$$

and $\bar{\boldsymbol{x}}_l := \boldsymbol{x}_l / \|\boldsymbol{x}_l\|_2$. For the sequence $\bar{\boldsymbol{x}}_l$ we will compute the norms in (5.1). As in the proof of Lemma 3.3 we can compute

$$\left\| \mathcal{A}\bar{\boldsymbol{x}}_l - \frac{1}{r_l}\bar{\boldsymbol{x}}_l \right\|_2^2 = \frac{2}{r_l^2 \|\boldsymbol{x}_l\|_2^2} < \frac{2\alpha^2}{\|\boldsymbol{x}_l\|_2^2}, \tag{5.3}$$

where the inequality follows by definition of r_l, where $\|A_k\| < \alpha \ \forall\ k$.

Using construction of the sequence $\boldsymbol{x}_l$ and the boundedness of $\boldsymbol{A}$ it can be shown that $\|\boldsymbol{x}_l\|_2 \to \infty$ as $l \to \infty$ (see the proof of Lemma 3.3). Thus there exists an $L \in \mathbb{N}$ such that $\forall\ l > L$ the right-hand side of (5.3) is bounded above by δ. Furthermore, as in the proof of Lemma 4.8, we can show that

$$\|\boldsymbol{C}\bar{\boldsymbol{x}}_l\|_2^2 \le \frac{\max\{1, r_l^{2(l-1)}\}}{\|\boldsymbol{x}_l\|_2^2} \sum_{k=k_0(l)}^{k_0(l)+l-1} |y_k|^2.$$

Since $|x_{k_0(l)}| = 1$, we have that $\|\boldsymbol{x}_l\|_2^2 \ge 1$, and using the construction of ϵ_l, we get

$$\begin{aligned} \|\boldsymbol{C}\bar{\boldsymbol{x}}_l\|_2^2 &< \max\left\{1, r_l^{2(l-1)}\right\} \epsilon_l \\ &= \max\left\{1, r_l^{2(l-1)}\right\} \delta(1-1/l)^{2(l-1)} \\ &\le \delta, \end{aligned}$$

where the last step follows since $r_l < (1-1/l)^{-1}$. Thus we have found $L \in \mathbb{N}$ such that $\forall\ l > L$ we have that $\|\mathcal{A}\bar{\boldsymbol{x}}_l - \frac{1}{r_l}\bar{\boldsymbol{x}}_l\|_2^2 < \delta$ and $\|\boldsymbol{C}\bar{\boldsymbol{x}}_l\|_2^2 < \delta$. Since r_l is bounded, we can take a subsequence of $l \in \mathbb{N}$ for which r_l converges, say $r_l \to r$ as $l \to \infty$. Since $1/r_l < 1 - 1/l$, we have $1/r \ge 1$. Hence we have

$$\begin{aligned} \|\mathcal{A}\boldsymbol{x}_l - \tfrac{1}{r}\boldsymbol{x}_l\|_2^2 &< \delta, \\ \|\boldsymbol{C}\boldsymbol{x}_l\|_2^2 &< \delta \end{aligned}$$

for l large enough. But on the other hand we know that $\|\bar{\boldsymbol{x}}_l\|_2^2 = 1$, contradicting (5.1). □

Notice that again it suffices to only consider real nonnegative almost eigenvalues. Summarizing, we have the following theorem.

THEOREM 5.8. *The following statements are equivalent*:

(i) *The pair* $(\boldsymbol{C}, \boldsymbol{A})$ *is uniformly detectable.*

(ii) *There exists an operator* $\boldsymbol{H} \in \mathcal{M}$ *for which* $\boldsymbol{A} + \boldsymbol{HC}$ *is UES.*

(iii) *There exists no almost eigenvalue of* $\mathcal{A}$ *with magnitude greater than or equal to* 1 *for which the corresponding almost eigenvector* $\boldsymbol{x}_n$ *satisfies* $\|\boldsymbol{C}\boldsymbol{x}_n\|_2 \to 0$ *as* $n \to \infty$.

In the next section we will shortly outline the results for the dual properties of uniform observability (resp., uniform detectability), namely uniform reachability (resp., uniform stabilizability).

6. Uniform reachability and uniform stabilizability. In this section, we consider systems of the form

$$x_{k+1} = A_k x_k + B_k u_k.$$

For the concept of uniform reachability we use the definition in [9].

DEFINITION 6.1. *The pair $(\boldsymbol{A}, \boldsymbol{B})$ is called uniformly reachable if there exists $l \in \mathbb{N}$ such that for every k_0, ξ, there exists a uniformly bounded (with respect to $|\xi|$) input $u_{k_0}, \dots, u_{k_0+l-1}$ such that $x_{k_0+l} = \xi$ where we start at $x_{k_0} = 0$.*

Since $x_{k_0} = 0$, we have

$$x_{k_0+l} = \sum_{j=k_0}^{k_0+l-1} A_{k_0+l-1} \cdots A_{j+1} B_j u_j.$$

It is easy to see that the matrix

$$R(k_0, l) := \left[\begin{array}{cccc} B_{k_0+l-1} & A_{k_0+l-1} B_{k_0+l-2} & \cdots & A_{k_0+l-1} \cdots A_{k_0+1} B_{k_0} \end{array} \right]$$

has to be right-invertible, and this right-inverse has to be uniformly bounded in k_0.

Define the l-step reachability Gramian operator as the block-diagonal operator

$$\boldsymbol{W}_l := \sum_{j=0}^{l-1} \mathcal{A}^j \mathcal{B} \mathcal{B}^* (\mathcal{A}^*)^j.$$

Since

$$R(k_0, l) R(k_0, l)^T = (\boldsymbol{W}_l)_{k_0+l}$$

we have the following.

COROLLARY 6.2. *The pair $(\boldsymbol{A}, \boldsymbol{B})$ is uniformly (l-step) reachable if and only if $\boldsymbol{W}_l > \boldsymbol{0}$.*

Define the time-reverse operator Ω as [4]

$$(\Omega \boldsymbol{x})_k = x_{-k}.$$

It is straightforward to check that $(\Omega \boldsymbol{T} \Omega)_{i,j} = T^T_{-j,-i}$ for any operator $\boldsymbol{T}$. It is also easy to see that $\Omega = \Omega^*$, $\Omega^2 = \boldsymbol{I}$ and that the shift operator satisfies $\Omega \boldsymbol{Z} = \boldsymbol{Z}^* \Omega$. An important property of this time-reverse operator is that $\boldsymbol{A}$ is UES if and only if $\Omega \boldsymbol{A}^* \Omega$ is UES [4, 7]. Since

$$\boldsymbol{W}_l = \boldsymbol{Z}^* \Omega \left(\sum_{j=0}^{l-1} (\Omega \boldsymbol{A} \Omega \boldsymbol{Z})^j (\Omega \boldsymbol{B} \Omega)(\Omega \boldsymbol{B}^* \Omega)(\boldsymbol{Z}^* \Omega \boldsymbol{A}^* \Omega)^j \right) \Omega \boldsymbol{Z},$$

we can compare this with the observability Gramian (4.1). With $\Omega^2 = \boldsymbol{Z}^* \boldsymbol{Z} = \boldsymbol{I}$ it is immediate that the pair $(\boldsymbol{A}, \boldsymbol{B})$ is uniformly reachable if and only if the pair $(\Omega \boldsymbol{B}^* \Omega, \Omega \boldsymbol{A}^* \Omega)$ is uniformly observable. Going through the details we can show the following equivalences.

THEOREM 6.3. *The following statements are equivalent:*

(i) *The pair $(\boldsymbol{A}, \boldsymbol{B})$ is uniformly reachable.*

(ii) *There exists $l \in \mathbb{N}$ such that $\boldsymbol{W}_l > \boldsymbol{0}$.*

(iii) *There exists no almost eigenvalue of* $\boldsymbol{Z}\boldsymbol{A}^*$ *for which the corresponding almost eigenvector* $\boldsymbol{x}_n$ *satisfies* $\|\boldsymbol{B}^*\boldsymbol{x}_n\|_2 \to 0$ *as* $n \to \infty$.

Moreover, if the pair $(\boldsymbol{A}, \boldsymbol{B})$ *is uniformly reachable, then the operator*

$$\boldsymbol{F} := -\mathcal{B}^*(\mathcal{A}^*)^l \boldsymbol{W}_{l+1}^{-1} \mathcal{A}^{l+1},$$

where l *is such that* (ii) *is satisfied, guarantees that* $\boldsymbol{A} + \boldsymbol{B}\boldsymbol{F}$ *is UES.*

For the concept of uniform stabilizability we use the definition in [1, 3].

Definition 6.4. *The pair* $(\boldsymbol{A}, \boldsymbol{B})$ *is called uniformly stabilizable if there exist* $l, p \in \mathbb{N}$ *and constants* $\gamma \in [0, 1)$ *and* $\epsilon > 0$ *such that, whenever*

$$|x_{k_0+p}| \geq \gamma |x_{k_0}|$$

for some k_0 *and* x_{k_0}, *then*

$$x_{k_0}^T (\boldsymbol{W}_l)_{k_0+l} x_{k_0} \geq \epsilon |x_{k_0}|^2.$$

The idea is that, when a zero-input trajectory starting at x_{k_0} fails to converge much toward the origin, there exists a bounded input steering the state toward the origin. Note that this mimics the LTI case, where stabilizability is equivalent to the requirement that any uncontrollable mode be asymptotically stable. Also, it is easily seen that uniform reachability implies uniform stabilizability.

Again, using the time-reverse operator Ω, it can be shown that the pair $(\boldsymbol{A}, \boldsymbol{B})$ is uniformly stabilizable if and only if the pair $(\Omega \boldsymbol{B}^* \Omega, \Omega \boldsymbol{A}^* \Omega)$ is uniformly detectable (see also [1]). Hence we find the following equivalences.

Theorem 6.5. *The following statements are equivalent:*

(i) *The pair* $(\boldsymbol{A}, \boldsymbol{B})$ *is uniformly stabilizable.*

(ii) *There exists an operator* $\boldsymbol{F} \in \mathcal{M}$ *for which* $\boldsymbol{A} + \boldsymbol{B}\boldsymbol{F}$ *is UES.*

(iii) *There exists no almost eigenvalue of* $\boldsymbol{Z}\boldsymbol{A}^*$ *with magnitude greater than or equal to* 1 *for which the corresponding almost eigenvector* $\boldsymbol{x}_n$ *satisfies* $\|\boldsymbol{B}^*\boldsymbol{x}_n\|_2 \to 0$ *as* $n \to \infty$.

7. Conclusions.

The work of [5, 8] was the first to show spectral characterizations of key properties of LTV systems could be formulated, which were the direct analogues of the LTI characterizations, provided that one works with the block-weighted shift $\mathcal{A}$. In this paper we have shown that one need work only with the subset of this spectrum, consisting of the almost eigenvalues of $\mathcal{A}$.

We have also shown that this allows one to consider spectral tests for observability and detectability and their duals, reachability and stabilizability, for linear discrete-time time-varying systems. These tests are the natural generalization of the PBH tests known in the LTI setting.

As was mentioned in the text, computation of the almost eigenvalues/eigenvectors will be difficult in most general cases. Nevertheless, we believe that the real use of the PBH test, even for time-invariant systems, lies not in the numerical determination of controllability/observability, but in its use in theoretical derivations. The characterizations of observability and detectability found in this paper should prove to be similarly useful.

REFERENCES

[1] B. Anderson and J. Moore, *Detectability and stabilizability of time-varying discrete-time linear systems*, SIAM J. Control Optim., 19 (1981), pp. 20–32.

[2] B. Beauzamy, *Introduction to Operator Theory and Invariant Subspaces*, North–Holland, Amsterdam, The Netherlands, 1988.

[3] W. Hager and L. Horotwitz, *Convergence and stability properties of the discrete riccati operator equation and the associated optimal control and filtering problems*, SIAM J. Control Optim., 14 (1976), pp. 295–312.

[4] A. Halanay and V. Ionescu, *Time-Varying Discrete Linear Systems*, Oper. Theory Adv. Appl. 68, Birkhäuser, Basel, Switzerland, 1994.

[5] E. Kamen, P. Khargonekar, and K. Poolla, *A transfer-function approach to linear time-varying discrete-time systems*, SIAM J. Control Optim., 23 (1985), pp. 550–565.

[6] J. Moore and B. Anderson, *Coping with singular transition matrices in estimation and control stability theory*, Internat. J. Control, 31 (1980), pp. 571–586.

[7] M. Peters and P. Iglesias, *Minimum Entropy Control for Time-Varying Systems*, Birkhäuser, Boston, MA, 1997.

[8] K. Przyłuski and S. Rolewicz, *On the stability of linear time-varying infinite-dimensional discrete-time systems*, Systems Control Lett., 4 (1984), pp. 307–315.

[9] W. Rugh, *Linear System Theory*, 2nd ed., Prentice–Hall, Englewood Cliffs, NJ, 1996.

[10] A. Shields, *Weighted shift operators and analytic function theory*, in Topics in Operator Theory, C. Pearcy, ed., Math. Surveys Monogr. 13, American Mathematical Society, Providence, RI, 1974, pp. 49–128.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 5, pp. 1346–1371

PARAMETRIZED FAMILIES OF EXTREMALS AND SINGULARITIES IN SOLUTIONS TO THE HAMILTON–JACOBI–BELLMAN EQUATION*

MATTHEW KIEFER† AND HEINZ SCHÄTTLER†

Abstract. We analyze the effect which a fold and simple cusp singularity in the flow of a parametrized family of extremal trajectories of an optimal control problem has on the corresponding parametrized cost or value function. A fold singularity in the flow of extremals generates an edge of regression of the value implying the well-known results that trajectories stay strongly locally optimal until the fold-locus is reached, but lose optimality beyond. Thus fold points correspond to conjugate points. A simple cusp point in the parametrized flow of extremals generates a swallowtail in the parametrized value. More specifically, there exists a region in the state space which is covered 3:1 with both locally minimizing and maximizing branches. The changes from the locally minimizing to the maximizing branch occur at the fold-loci and there trajectories lose strong local optimality. However, the branches intersect and generate a cut-locus which limits the optimality of close-by trajectories and eliminates these trajectories from optimality near the cusp point *prior* to the conjugate point. In the language of partial differential equations, a simple cusp point generates a shock in the solutions to the Hamilton–Jacobi–Bellman equation while fold points will not be part of the synthesis of optimal controls near the simple cusp point.

Key words. Hamilton–Jacobi–Bellman equation, method of characteristics, fold and simple cusp singularities, conjugate points

AMS subject classifications. Primary, 49K15, 49L05; Secondary, 35B37, 35L67

PII. S0363012997319139

1. Introduction.

We study singularities in solutions to the Hamilton–Jacobi–Bellman equation for the value-function of an optimal control problem for ordinary differential equations. It is well known (see, for instance, [3]) that the necessary conditions for optimality given in the Pontryagin maximum principle [28] also give the characteristic equations for the Hamilton–Jacobi–Bellman equation. Thus, if the flow of extremals (trajectories which satisfy the necessary conditions of the Pontryagin maximum principle) covers an open region R of the state-space diffeomorphically, then a smooth solution to the Hamilton–Jacobi–Bellman equation can be constructed on R by the method of characteristics. In general, however, except for special problems like the linear-quadratic regulator, the value function is typically not smooth. The difficulties in finding solutions to optimal control problems or, equivalently, in finding solutions to the Hamilton–Jacobi–Bellman equation, precisely lie in analyzing the singularities.

There is a large body of literature on singularities of solutions to the Hamilton–Jacobi–Bellman equation. Most of these papers deal with general topological properties of the singular set or try to establish smoothness of solutions. In [17] Fleming proves that the singular set is closed and of Hausdorff dimension at most equal to the dimension of the state space. Dafermos [16] analyzes singular sets for more general hyperbolic conservation laws, but only in dimension one. Cannarsa and Soner [11] analyze the local structure of the singular set for viscosity solutions which satisfy cer-

*Received by the editors March 31, 1997; accepted for publication (in revised form) July 8, 1998; published electronically June 29, 1999. This research was supported in part by NSF grant DMS-9503356.

http://www.siam.org/journals/sicon/37-5/31913.html

†Department of Systems Science and Mathematics, Washington University, St. Louis, MO, 63130-4899 (Matthew.Kiefer@AUC.TRW.com, hms@cec.wustl.edu).

tain Lipschitz-type conditions. This literature typically addresses the problem from a PDE point of view. Few papers deal with the optimal control problems directly. Cannarsa and Frankowska [11] show that smoothness of the value-function at a point of a trajectory implies smoothness at all later times along the trajectory. Byrnes and Frankowska [9] and Byrnes and Jhemi [10] use convexity properties of the Hamiltonian of the optimal control problem to obtain special structures for the Hamilton–Jacobi–Bellman equation as a Riccati-type PDE and they give conditions for the absence of shocks, i.e., for smooth solutions. In most of these papers, establishing the differentiability of the value-function or of solutions to the Hamilton–Jacobi–Bellman equation is the main interest; it is not to analyze the structure of singularities in more detail.

A different approach to solve optimal control problems is the theory of regular synthesis. Here the value function is only derived indirectly. Instead the optimal control is synthesized as a feedback control. The original paper by Boltyanskii [5] gives conditions which allow to construct the value-function despite singularities and to prove the optimality of the corresponding trajectories. These conditions have been weakened considerably since then by Brunovsky [7] and Sussmann [33] and by Sussmann and Piccoli [27]. General results about the existence of a regular synthesis have been proven by Brunovsky [7]. Detailed results which establish regular syntheses exist in low dimensions. Piccoli [26] analyzed the generic singularities for time-optimal control for two-dimensional systems based on Sussmann's results [34, 35, 36]. In dimension three, Krener and Schättler [22] and Schättler and Jankovic [**?**] construct a regular synthesis for time-optimal control problems under codimension 0 and 1 conditions. In these results the precise structure of the singularities of the value-function is established under generic assumptions. Naturally it is quite difficult if not impossible to obtain precise results under general conditions. Kupka [23] analyzes the least degenerate singularities for extremals in the cotangent bundle, i.e., together with the corresponding multipliers. However, projections into the state space still need to be taken to construct a regular synthesis.

In this paper we describe a framework which aims to analyze and explain singularities in solutions to the Hamilton–Jacobi–Bellman equation through the behavior of the extremal trajectories of the optimal control problem. The starting point for our analysis is *smoothly parametrized families of extremals* (see section 3). In our view this is a natural concept which simply formalizes the notion of integrating extremals of the control problem backward from the terminal manifold without imposing that the flow is a diffeomorphism, i.e., is a field. Yet the relevant identities for the corresponding parametrized cost can be established ("shadow prices") which, when coupled with injectivity of the flow, allow us to prove that the cost is the value function and hence that controls are optimal. While the global existence of these smooth parametrizations depends on the nature of the problem, they generically exist locally around a reference extremal and thus this concept is well suited to study strong local optimality. These parametrizations can typically be obtained in a rather straightforward way near reference extremals which are sufficiently regular finite concatenations of continuous controls (like bang-bang extremals or more general concatenations with singular arcs) by taking the terminal point(s) and/or switching times along the arcs as parameters [24].

Given a smoothly parametrized family of extremals, well-known results from the singularity theory of functions [20] can be applied. In this paper we analyze the effect which the occurrence of the two least degenerate singularities, the so-called fold and simple cusp catastrophes [37, 38], in the flow of extremals have on the structure of the value-function. For the fold singularity these results, or rather their formulations which give the implications on the structure of optimal controls, are classical. Fold points in smooth parametrizations of the extremals generate "edges of regression" [2] in the corresponding value function. It follows that the fold generates the well-known behavior of trajectories which stay strongly locally optimal until the fold-locus is reached but lose optimality beyond. Hence fold singularities correspond to conjugate points. Numerous formulations of related statements exist in the literature ranging from early engineering texts [8] to modern formulations [14]. For instance, Caroff and Frankowska [14] analyze the characteristics of the Hamilton–Jacobi–Bellman equation and formulate necessary conditions for a weak minimum and sufficient conditions for a strong minimum in terms of conjugate points. The closest formulation to ours is the one given by Agrachev and Gamkrelidze in terms of concepts from symplectic geometry. It has been shown [1, Theorem 3.1] that extremals lose optimality as the corresponding extremal lifts (which form Lagrangian manifolds in the cotangent bundle) pass a fold. The details, however, are beyond the scope of this article and we strongly recommend the reader consult the expository article [1]. For the sake of completion, and to set the stage, in this paper we include a brief section on the fold singularity, which states the result in the framework of parametrized families of extremals; we refer the reader to [21] for the proof.

Our focus in this paper is on the effect which a simple cusp singularity in the flow of extremals has on the value function. Indeed, the behavior of optimal trajectories implied by the fold singularity is seldom seen in a regular synthesis of optimal trajectories, i.e., although these trajectories stay locally optimal up to the conjugate point, they are no longer optimal globally prior to the conjugate point. This is explained by the behavior of extremals near a simple cusp point. The simple cusp point generates a region in the state space which is covered 3:1 by both locally minimizing and locally maximizing branches of the corresponding value function. Away from the simple cusp point the changes from the locally minimizing to the maximizing branches occur at the fold-loci and there trajectories lose strong local optimality. However, the two minimizing (maximizing) branches intersect and generate a cut-locus which limits the optimality of the close-by trajectories and indeed eliminates these trajectories from optimality near the cusp point prior to the conjugate point. An analogous behavior of optimal trajectories can already be seen in the classical calculus of variations. For the problem of finding surfaces of revolution of minimum surface area, smooth solutions correspond to the catenaries and cease to be locally optimal at the conjugate point which is the fold point in the parametrization [4, 15]. However, the Goldschmidt solutions (which correspond to broken extremals) are better already for points prior to the conjugate point in the catenaries and the optimal solution is determined by the cut-locus between these two families of trajectories [15, pp. 143–148]. Exactly the same behavior is described in [22] for the three-dimensional time-optimal control problem to an equilibrium point for the generic nonlinear system.

Our result can also be interpreted in terms of the classification of Lagrangian singularities. If the parametrized flow of extremals undergoes a simple cusp singularity, then the corresponding value exhibits a swallow-tail singularity [2]; respectively, in the terminology of [20], its graph looks like the singular set of a swallow-tail singularity. This connection, which implies all the statements about the optimal trajectories made above, will be proven by elementary means in this paper. In particular, verifiable conditions for the existence of a simple cusp in the parametrization of the extremals allow us to determine the occurrence of the swallow-tail singularity in the parametrized value. In summary, our main result implies that *the simple cusp catastrophe generates a shock* (a point for which there exist multiple characteristics) *in the solutions to the Hamilton–Jacobi–Bellman equation while fold catastrophes near a simple cusp will not be part of the solution or synthesis of optimal controls near the simple cusp point.*

In section 2 we formulate the optimal control problem and review the conditions of the maximum principle and the Hamilton–Jacobi–Bellman equation. Section 3 contains a formulation of the method of characteristics for the Hamilton–Jacobi–Bellman equation. This material is classical and is based on ideas from field theory [13, 39, 3, 25]. These results are included since we are not aware of a published reference which would treat the subject in the form as we need it in the sense that our presentation emphasizes the parametrization aspect. The proofs which are based on [30] are only indicated. In section 4 we discuss the normal forms for local coordinates for the fold and simple cusp points and the corresponding changes of coordinates. Normal coordinates will then be used in sections 5 and 6 to analyze the local behavior of the parametrized cost when a parametrized family of extremals undergoes a fold or a simple cusp singularity.

2. Problem formulation. Let U be a subset of $\mathbb{R}^m$, the *control set*, and denote by $\mathcal{U}$ the class of all locally bounded Lebesgue measurable maps defined on some interval $I \subset \mathbb{R}$ with values in U, $u : I \to U$, the *space of (admissible) controls.* Suppose

$$f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n, \quad (t, x, u) \mapsto f(t, x, u),$$

the dynamics of the control system, is a continuous map which for fixed $t \in \mathbb{R}$ is r-times continuously differentiable in (x, u). Let N be a k-dimensional embedded C^{r+1}-submanifold of (t, x)-space $\mathbb{R} \times \mathbb{R}^n$, the *terminal manifold.* Thus, near every point $q \in N$, there exists an open set $\Omega \subset \mathbb{R} \times \mathbb{R}^n$ containing q and a C^{r+1}-map $\Psi : \Omega \to \mathbb{R}\times\mathbb{R}^n$ with components ψ_i, $i = 0, \dots , n-k$, which have linearly independent gradients $\nabla\psi_i$ and satisfy $N \cap \Omega = \{(t, x) \in \Omega : \psi_i(t, x) = 0, \quad i = 0, \dots , n-k\}$. Also let $\varphi : N \to \mathbb{R}$ be a C^{r+1}-function. We consider the problem to minimize over $\mathcal{U}$ a cost functional given in Bolza form as

$$\mathcal{J}(u; \tau, \xi) = \int_\tau^T L(s, x, u)ds + \varphi(T, x(T)) \tag{2.1}$$

subject to the dynamics $\dot{x} = f(t, x, u)$ with initial condition $x(\tau) = \xi$ and terminal condition $(T, x(T)) \in N$. The terminal time T is free. (A fixed terminal time would be included as a constraint in N.)

The maximum principle gives necessary conditions for a controlled trajectory (x, u) to be optimal. In our notation we distinguish between tangent vectors which we write as column vectors (such as x, $f(t, x, u)$ etc.) and cotangent vectors which we write as row vectors (like the multipliers λ and ν in the statement of the Maximum principle below). We denote the space of n-dimensional row vectors by $(\mathbb{R}^n)^*$. Define the Hamiltonian function H,

$$H : \mathbb{R} \times [0, \infty) \times (\mathbb{R}^n)^* \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$$

as

$$H(t, \lambda_0, \lambda, x, u) = \lambda_0 L(t, x, u) + \lambda f(t, x, u) \tag{2.2}$$

Theorem 2.1 (Pontryagin maximum principle [28]). *Suppose the controlled trajectory (x,u) defined over the interval $[\tau,T]$ is optimal. Then there exist a constant $\lambda_0 \geq 0$, a covector $\nu \in (\mathbb{R}^{n+1-k})^*$, and an absolutely continuous function $\lambda : [\tau,T] \to (\mathbb{R}^n)^*$, the adjoint variable, such that $(\lambda_0,\lambda(t)) \neq 0$ for all $t \in [\tau,T]$ and the following conditions are satisfied:*

1. *Adjoint equation: almost everywhere in* $[\tau,T]$

$$\dot{\lambda}(t) = -\lambda_0 L_x(t,x(t),u(t)) - \lambda(t) f_x(t,x(t),u(t)), \tag{2.3}$$

2. *Minimum condition: almost everywhere in* $[\tau,T]$

$$H(t,\lambda_0,\lambda(t),x(t),u(t)) = \min_{v\in U} H(t,\lambda_0,\lambda,x,v) \tag{2.4}$$

3. *Transversality condition: the vector $(H+\lambda_0\varphi_t, -\lambda+\lambda_0\varphi_x)$ is orthogonal to the terminal constraint in the endpoint, i.e., at the terminal time T we have that*

$$0 = \lambda_0\varphi_t + \nu D_t\Psi + H, \qquad \lambda = \lambda_0\varphi_x + \nu D_x\Psi. \tag{2.5}$$

We call controlled trajectories (x,u) for which there exist multipliers λ_0, λ, and ν such that the conditions of the maximum principle are satisfied *extremals* and sometimes we refer to the triple (λ,x,u) as an *extremal lift.* Note that the conditions are linear in the multipliers (λ_0,λ,ν) and thus it is possible to normalize this vector. In particular, if $\lambda_0 > 0$, then we can divide by λ_0 and thus assume $\lambda_0 = 1$. These kinds of extremals are called *normal* while extremals with $\lambda_0 = 0$ are called *abnormal.* The existence of optimal abnormal extremals can in general not be ruled out.

Sufficient conditions for optimality center around feedback controls, Hamilton–Jacobi theory, and the related notions of fields of extremals and the value-function. Define the *value-function* V on some open subset R of the state-space, $V : R \to \mathbb{R}$, as

$$V(t,x) = \inf_{u\in\mathcal{U}} \mathcal{J}(u;t,x), \tag{2.6}$$

i.e., as the infimum over all $u \in \mathcal{U}$ of the values of the cost functional $\mathcal{J}(u;t,x)$ if the initial conditions of the dynamics are given by (t,x). It is elementary to see [19] that if the value-function V is differentiable at (t,x), then V satisfies the inequality

$$V_t(t,x) + V_x(t,x)f(t,x,u) + L(t,x,u) \geq 0 \tag{2.7}$$

for all $u \in U$ and equality holds for a (sufficiently regular) optimal control. Conversely, if V is any continuously differentiable function and $u^* = u^*(t,x)$ is an admissible feedback control which, together, satisfy the so-called Hamilton–Jacobi–Bellman equation on an open connected set R containing N, i.e.,

$$V_t(t,x) + \min_{u\in U}\{V_x(t,x)f(t,x,u) + L(t,x,u)\} \equiv 0. \tag{2.8}$$

$$V(t,x) = \varphi(t,x) \quad \text{for } (t,x) \in N, \tag{2.9}$$

then it is well known (and completely elementary under these differentiability conditions) that the feedback control u^* is optimal on R (i.e., with respect to any other controlled trajectory which entirely lies in R) and that V is the corresponding value-function. In particular, each of the extremals in the associated field is a strong local minimum over the set R.

3. The method of characteristics in optimal control. In this section we show how a solution to the Hamilton–Jacobi–Bellman equation can be constructed from a parametrized family of normal extremals which satisfy the maximum principle and cover a set R injectively. The construction below formulates the method of characteristics adapted to the optimal control problem and is a refinement of arguments in [30] which are based on notes by Knobloch. In sections 4 and 5 we use this particular local construction to relate corank 1 singularities in the parametrization to singularities in solutions to the Hamilton–Jacobi–Bellman equation.

Definition 3.1. *A C^r-parametrized family $\mathcal{E}$ of extremals is an 8-tuple $(P; \mathcal{T} = (t_0, t_f); \xi, \nu; x, u, \lambda_0, \lambda)$ consisting of*

- *an open set P in some n-dimensional manifold and a pair $\mathcal{T} = (t_0, t_f)$ of r times continuously differentiable functions $t_0 : P \to \mathbb{R}$ and $t_f : P \to \mathbb{R}$ defined on P which satisfy $t_0(p) < t_f(p)$ for all $p \in P$. They define the domain of the parametrization as $D = \{(t,p) : p \in P,\ t \in I_p = [t_0(p), t_f(p)]\}$. The functions t_0 and t_f define the (compact) intervals of definition for the controlled trajectories with t_f denoting the terminal time.*
- *an r-times continuously differentiable function $\xi : P \to N$ which parametrizes the terminal conditions for the states.*
- *extremal lifts consisting of controlled trajectories $(x,u) : D \to \mathbb{R}^n \times U$, corresponding adjoint vectors $\lambda_0 : P \to [0,\infty)$ and $\lambda : D \to (\mathbb{R}^n)^*$, and a row vector $\nu : P \to (\mathbb{R}^{n+1-k})^*$ which parametrizes the terminal conditions for the costates. Specifically, we assume*
  1. *the multipliers $(\lambda_0(p), \lambda(t,p))$ are nontrivial for all $t \in I_p$,*
  2. *the controls $u = u(\cdot,p)$, $p \in P$, parametrize admissible controls which are continuous in (t,p) and for t fixed depend r-times continuously differentiable on p with the derivatives continuous in (t,p),*
  3. *the trajectories $x = x(t,p)$ solve the terminal value problems for the dynamics*

     $$\dot{x}(t,p) = f(t, x(t,p), u(t,p)), \quad x(t_f(p), p) = \xi(p), \tag{3.1}$$

  4. *the costate $\lambda = \lambda(t,p)$ solves the corresponding adjoint equation*

     $$\dot{\lambda}(t,p) = -\lambda_0(p) L_x(t, x(t,p), u(t,p)) - \lambda(t,p) f_x(t, x(t,p), u(t,p)), \tag{3.2}$$

     with terminal conditions

     $$\lambda(t_f(p), p) = \lambda_0(p)\varphi_x(t_f(p), \xi(p)) + \nu(p) D_x \Psi(t_f(p), \xi(p)), \tag{3.3}$$

  5. *the controls solve the minimization problem*

     $$H(t, \lambda_0(p), \lambda(t,p), x(t,p), u(t,p)) = \min_{v \in U} H(t, \lambda_0(p), \lambda(t,p), x(t,p), v), \tag{3.4}$$

  6. *the transversality condition on the terminal time,*

     $$\begin{aligned} &H(t_f(p), \lambda_0(p), \lambda(t_f(p),p), x(t_f(p),p), u(t_f(p),p)) \\ &\qquad + \lambda_0(p)\varphi_t(t_f(p), \xi(p)) + \nu(p) D_t \Psi(t_f(p), \xi(p)) = 0, \end{aligned} \tag{3.5}$$

     holds.

This notion of a C^r-parametrized family of extremals is a natural concept to formalize an approach which tries to construct a field of extremals by integrating the system and adjoint equations backward from the terminal manifold while maintaining the minimum condition (3.4) without requiring that the flow is a diffeomorphism. No injectivity assumptions are made.

It follows from standard results about differentiable dependence on parameters of solutions to ordinary differential equations that the trajectories $x = x(t,p)$ (and their time-derivatives $\dot{x}(t,p)$) are r-times continuously differentiable in p for fixed t and that these derivatives are continuous jointly in (t,p). We denote this class of functions by $C^{1,r}$. These partial derivatives can be calculated as solutions to the corresponding variational equations which are obtained by interchanging the time-derivative with the p-derivatives. At the moment we do not impose regularity conditions on λ_0 and ν. Consequently $\lambda(t,p)$ need not be differentiable in p. However, if λ_0 and ν are r-times continuously differentiable, then the costate λ has the same smoothness properties as the state x. In this case we call $\mathcal{E}$ a *nicely* C^r-parametrized family of extremals.

Modifications of the definitions for the time-invariant case are straightforward with the one exception that we then want the control to lie in $C^{r-1,r}$. This guarantees that $x(t,p)$ has the desired differentiability properties.

Example. Consider the optimal control problem to minimize

$$J(u) = \frac{1}{2}\int_0^T ||u||_2^2 dt + \varphi(x(T)) \tag{3.6}$$

subject to

$$\dot{x} = f(x) + \sum_{i=1}^{m} g_i(x)u_i, \qquad u_i \in \mathbb{R}, \tag{3.7}$$

where f and the g_i are C^r vector fields on $\mathbb{R}^n$, φ is C^{r+1}, and the terminal time T is fixed. It is easy to see that extremals for this problem are normal and thus, without loss of generality, we may normalize $\lambda_0 \equiv 1$. The Hamiltonian

$$H = \frac{1}{2}\sum_{i=1}^{m} u_i^2 + \lambda\left(f(x) + \sum_{i=1}^{m} g_i(x)u_i\right) \tag{3.8}$$

is strictly convex in u with the unique minimum given by $u_i = -\lambda g_i(x)$. Substituting this relation into the dynamics and adjoint equation gives the following system of $2n$ ordinary differential equations

$$\dot{x} = f(x) - \sum_{i=1}^{m} g_i(x)g_i(x)^T\lambda^T \tag{3.9}$$

$$\dot{\lambda} = -\lambda\left(Df(x) - \sum_{i=1}^{m} Dg_i(x)g_i(x)^T\lambda^T\right). \tag{3.10}$$

If we take as parameter space $P = \mathbb{R}^n$ and integrate these differential equations backward with terminal conditions

$$x(T,p) = p, \qquad \lambda(T,p) = \varphi_x(T,p), \tag{3.11}$$

then the solutions $x = x(t,p)$ and $\lambda = \lambda(t,p)$ exist on a maximal interval $I_p = (t_0(p), T]$ and these solutions are C^r functions of t and p. In particular, the controls

$$u_i(t,p) = -\lambda(t,p)g_i(x(t,p)) \tag{3.12}$$

are also $C^{r,r}$. Thus this generates a nicely C^r parametrized family of extremals.

More generally, if a terminal manifold N is given with free terminal time, then extremals will be parametrized through their endpoints in the terminal manifold N (k-dimensional) and the vector ν in the transversality condition which gives the terminal condition for the multiplier λ ($(n+1-k)$-dimensional). However, we also need to enforce the transversality condition (3.5) on H which pins down the terminal time. Hence the parameter space is n-dimensional.

While the procedure is straightforward, clearly the smoothness conditions required in the parametrizations need not be satisfied in general. But the definition of a C^r-parametrized family of extremals and the shadow price lemma given below generalize to parametrized families of *broken* extremals which have discontinuities on manifolds described by C^r functions $t = t(p)$. For instance, a family of bang-bang trajectories can easily be parametrized using the switching times as parameters. In this paper, however, we consider only parametrized families where the control is continuous (respectively, $C^{r-1,r}$ in the time-invariant case) since we want to study singularities as they arise in smooth parametrizations and study their effect on the parametrized value-function. In our view these parametrizations are no artefact, but arise naturally. Particularly this holds if the parameter set P is restricted to some small neighborhood of some reference value respective to reference trajectory. Then our results have immediate implications on the strong local optimality of the reference trajectory.

On D we define the cost $C : D \to \mathbb{R}$ along a C^r-parametrized family $\mathcal{E}$ of extremals as

$$C(t,p) = \int_t^{t_f(p)} L(s, x(s,p), u(s,p))ds + \varphi(t_f(p), \xi(p)), \tag{3.13}$$

i.e., $C(t,p)$ denotes the cost for the optimal control problem with initial condition $(t,x) = (t, x(t,p))$ corresponding to the control $u = u(t,p)$. It follows from our assumptions and the above smoothness properties that C is continuously differentiable in t with time-derivative

$$\frac{\partial C}{\partial t} = -L(t, x(t,p), u(t,p)) \tag{3.14}$$

and that both C and its time-derivative $\frac{\partial C}{\partial t}$ are r-times continuously differentiable in p.

Notation. For a function like C we denote the gradient with respect to p (which we consider a row vector) by $\frac{\partial C}{\partial p}$. Consequently, for a column vector like $x = (x_1, \ldots, x_n)^T$ we denote by $\frac{\partial x}{\partial p}$ the matrix whose rows are given by the gradients of the components of x, i.e., $\frac{\partial x}{\partial p} = \left(\frac{\partial x_i}{\partial p_j}\right)_{1 \le i,j \le n}$ with row index i and column index j. However, to be consistent, for a row-vector like $\lambda = (\lambda_1, \ldots, \lambda_n)$ we denote the matrix of the partial derivatives $\left(\frac{\partial \lambda j}{\partial p_i}\right)_{1 \le i,j \le n}$ with row index i and column index j by $\frac{\partial \lambda}{\partial p}$. In this sense, we have $\frac{\partial \lambda}{\partial p} = \left(\frac{\partial \lambda^T}{\partial p}\right)^T$. This will allow us to write most formulas without having to use transposes. Finally we denote by $\frac{\partial^2 C}{\partial p^2}$ the Hessian matrix of the second derivatives of a function C.

The following relation is crucial to the whole construction:

LEMMA 3.1 (shadow prices). *Let $\mathcal{E}$ be a C^1-parametrized family of extremals. Then we have that*

$$\lambda_0(p)\frac{\partial C}{\partial p}(t,p) = \lambda(t,p)\frac{\partial x}{\partial p}(t,p). \tag{3.15}$$

Proof (*sketch*). For p fixed both sides of (3.15) are continuously differentiable functions in t. It therefore suffices to show that both sides have the same t-derivative with identical values at the terminal time $t_f(p)$. The latter can be shown by adjoining the terminal condition with multiplier $\nu = \nu(p)$ to C and then using the transversality conditions (2.5) after differentiating in p. Furthermore, using the adjoint equation and the variational equation for $\frac{\partial x}{\partial p}$, it follows that

$$\frac{d}{dt}\left\{\lambda(t,p)\frac{\partial x}{\partial p}(t,p)\right\} = \lambda_0(p)\frac{\partial^2 C}{\partial t \partial p}(t,p) + H_u\frac{\partial u}{\partial p}(t,p),$$

where H_u is evaluated along the parametrized extremal. But

$$H_u(t,\lambda_0(p),\lambda(t,p),x(t,p),u(t,p))\frac{\partial u}{\partial p}(t,p) \equiv 0 \qquad \text{on } D \tag{3.16}$$

follows from the minimization property of the extremal control $u(t,p)$. □

If the parametrization of the extremals covers the state-space diffeomorphically, then the shadow price lemma provides the bridge between the necessary conditions for optimality of the maximum principle and the sufficient conditions of the Hamilton–Jacobi theory. Theorem 3.1 below formalizes this. In its essential contents the statement is classical and can be found in many books on optimal control such as [3, 18]. It is included here since it also gives a means to calculate the singularities in a parametrized flow of extremals.

THEOREM 3.1. *Let $\mathcal{E}$ be a C^r-parametrized family of normal extremals, $r \geq 1$, and suppose the map*

$$\sigma : D \to \mathbb{R}\times\mathbb{R}^n, \qquad (t,p) \mapsto (t,x(t,p)) \tag{3.17}$$

is a C^1-diffeomorphism from some open subset $O \subset \operatorname{int} D$ onto an open subset $R \subset \mathbb{R}\times\mathbb{R}^n$ of the state-space (i.e., the map is bijective and has an everywhere nonvanishing Jacobian). Then the function $V : R \to \mathbb{R}, V = C\circ\sigma^{-1}$, is continuously differentiable in t and r-times continuously differentiable in x on S. The function $u^* : R \to \mathbb{R}, u^* = u\circ\sigma^{-1}$ is an admissible feedback-control which is continuous in t and r-times continuously differentiable in x. Together the pair (V,u^*) solves the Hamilton–Jacobi–Bellman equation

$$V_t(t,x) + \min_{u\in U}\{V_x(t,x)f(t,x,u) + L(t,x,u)\} \equiv 0 \tag{3.18}$$

on R. Furthermore, the following identities hold in the parameter space on O:

$$V_t(t,x(t,p)) = -H(t,\lambda(t,p),x(t,p),u(t,p)) \tag{3.19}$$

$$V_x(t,x(t,p)) = \lambda(t,p). \tag{3.20}$$

If $\mathcal{E}$ is nicely C^r-parametrized, then V is $(r+1)$-times continuously differentiable in x on R and we also have

$$V_{xx}(t,x(t,p)) = \frac{\partial \lambda^T}{\partial p}(t,p)\left(\frac{\partial x}{\partial p}(t,p)\right)^{-1}. \tag{3.21}$$

Proof (*sketch*). By assumption σ is injective with $C^{1,r}$ inverse. Thus V and u^* are well defined and a priori $V \in C^{1,r}$. Since $C = V \circ \sigma$, we have that

$$\frac{\partial C}{\partial p}(t,p) = V_x(t,x(t,p))\frac{\partial x}{\partial p}(t,p),$$

and thus, in view of Lemma 3.1 and the fact that $\frac{\partial x}{\partial p}$ is nonsingular (3.20) follows. Equations (3.18) and (3.19) then follow directly from the conditions of the maximum principle. If $\mathcal{E}$ is nicely C^r parametrized, then λ is C^r in p, and thus we still have $V_x \in C^{1,r}$ since $V_x = \lambda \circ \sigma^{-1}$ on R. In particular, differentiating (3.20) with respect to p and observing that we need to take a transpose in λ to keep the notation consistent, we get (3.21). □

The relation (3.21) can be used to calculate singularities in the map x. Under the assumptions of Theorem 3.1 the function

$$Q(t,p) = V_{xx}(t,x(t,p)) = \frac{\partial \lambda^T}{\partial p}(t,p)\left(\frac{\partial x}{\partial p}(t,p)\right)^{-1} \tag{3.22}$$

satisfies the differential equation

$$\dot{Q} = -Qf_x - f_x^T Q - H_{xx} - (Qf_u + H_{xu})\frac{\partial u}{\partial p}\left(\frac{\partial x}{\partial p}\right)^{-1}, \tag{3.23}$$

where the partial derivatives of f and H are evaluated along the extremal corresponding to the parameter p. This follows by direct differentiation. In particular, if the gradient H_u vanishes identically and the matrix H_{uu} is positive definite along (λ, x, u) on I, then

$$\frac{\partial u}{\partial p} = -H_{uu}^{-1}\left(H_{ux}\frac{\partial x}{\partial p} + f_u^T\frac{\partial \lambda^T}{\partial p}\right), \tag{3.24}$$

and we can eliminate the control-term from (3.23) to get the customary Riccati-equation for the second derivatives V_{xx} [8]:

$$\dot{Q} = -Qf_x - f_x^T Q - H_{xx} + (Qf_u + H_{xu})H_{uu}^{-1}\left(H_{ux} + f_u^T Q\right). \tag{3.25}$$

In this case singularities in $\frac{\partial x}{\partial p}$ therefore relate to the explosion times of the Riccati-equation. It is clear from (3.21) that explosion times in the Riccati-equation can only occur as $\frac{\partial x}{\partial p}$ becomes singular and conversely

COROLLARY 3.1. *Suppose*

$$\frac{\partial x}{\partial p}(t_0,p_0)v_0 = 0 \qquad \text{and} \qquad w_0 = v_0^T\frac{\partial \lambda}{\partial p}(t_0,p_0) \neq 0.$$

Then the solution $Q(\cdot,p_0)$ to the Riccati-equation (3.25) *has a finite explosion time at $t = t_0$.*

Proof. We have $\lim_{t\to t_0} Q(t,p_0)\frac{\partial x}{\partial p}(t,p_0)v_0 = \lim_{t\to t_0} \frac{\partial \lambda^T}{\partial p}(t,p_0)v_0 = \frac{\partial \lambda^T}{\partial p}(t_0,p_0)v_0 = w_0^T \neq 0$. Since also $\lim_{t\to t_0} \frac{\partial x}{\partial p}(t,p_0)v_0 = \frac{\partial x}{\partial p}(t_0,p_0)v_0 = 0$, it follows that $Q(\cdot,p_0)$ has a finite explosion time at t_0. □

In particular, it is well known, and commercial software exists on how to calculate these points numerically. The next corollary, which follows by differentiating (3.15) with respect to p, relates left- and right-eigenvectors at singular points of the flow map. This will be needed to interpret the transversality conditions from singularity theory for the parametrized flow.

COROLLARY 3.2. *Let $\mathcal{E}$ be a C^2-parametrized family of normal extremals. Then the matrix*

$$\Xi(t,p) = \frac{\partial \lambda}{\partial p}(t,p)\frac{\partial x}{\partial p}(t,p) \tag{3.26}$$

is symmetric.

COROLLARY 3.3. *Let $\mathcal{E}$ be a C^2-parametrized family of normal extremals. Suppose*

$$\frac{\partial x}{\partial p}(t_0,p_0)v_0 = 0 \qquad \text{and} \qquad w_0 = v_0^T\frac{\partial \lambda}{\partial p}(t_0,p_0) \neq 0.$$

Then w_0 is a left-eigenvector for eigenvalue 0 to $\frac{\partial x}{\partial p}(t_0,p_0)$,

$$w_0\frac{\partial x}{\partial p}(t_0,p_0) = 0. \tag{3.27}$$

Proof. We have for all $z \in \mathbb{R}^n$ that

$$w_0\frac{\partial x}{\partial p}(t_0,p_0)z = v_0^T\frac{\partial \lambda}{\partial p}(t_0,p_0)\frac{\partial x}{\partial p}(t_0,p_0)z = z^T\frac{\partial \lambda}{\partial p}(t_0,p_0)\frac{\partial x}{\partial p}(t_0,p_0)v_0 = 0$$

and thus $w_0\frac{\partial x}{\partial p}(t_0,p_0) = 0$. □

4. Normal forms for functions near singularities.

In this section we recall the normal forms for folds and cusps for C^r maps $F : \mathbb{R}^n \mapsto \mathbb{R}^n$ and present the required equations for the changes of coordinates. The normal forms are classical and go back to fundamental papers by Whitney [38] and Thom [37]. We include a brief exposition following the presentation in [20]. However, in view of our applications to the optimal control problem, we formulate the results already for the map $\sigma : (t,p) \mapsto (t,x(t,p))$, where t plays the role of a bifurcation parameter. The singular set, S, of σ or equivalently of x, is given as $S = \{(t,p) \in D : \text{rank}\frac{\partial x}{\partial p}(t,p) < n\}$ and the corank 1 singularities, S_1, are where $\frac{\partial x}{\partial p}$ has rank $n-1$. If $(t_0,p_0) \in S_1$, then $\frac{\partial x}{\partial p}(t_0,p_0)$ has a simple eigenvalue zero. We will denote the corresponding left- and right-eigenvectors by w_0 and v_0, respectively. By choosing a sufficiently small open neighborhood D of (p_0,t_0) we may assume that σ has only corank 1 singularities on D. Furthermore, it is a mere consequence of the rank assumption [6, Chap. II.7] that there exist local changes of coordinates in the parameter space $\Phi : D \to D', (t,p) \mapsto (\tau,\xi) = \Phi(t,p) = (t-t_0,\ \xi(t,p))$ and in the state-space $\Psi : R \to R', (t,x) \mapsto (\tau,\eta) = \Psi(t,x) = (t-t_0,\ \eta(t,x))$, such that the map σ has the following form in the new coordinates:

$$k: \quad D' \to R', \qquad (\tau,\xi) = \begin{pmatrix} \tau \\ \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} \mapsto k(\tau,\xi) = \begin{pmatrix} \tau \\ \xi_1 \\ \vdots \\ \xi_{n-1} \\ h(\tau,\xi) \end{pmatrix} \tag{4.1}$$

for some function $h \in C^{1,r}$ (i.e., h is continuously differentiable in τ and r times continuously differentiable in ξ.) Furthermore, $h(0,0) = 0$ and $\frac{\partial h}{\partial \xi_n}(0,0) = 0$. In particular, if the gradient of $\frac{\partial h}{\partial \xi_n}$ does not vanish at (t_0, p_0), then S_1 is an $(n-1)$-dimensional embedded submanifold near (t_0, p_0). The transversality conditions which determine the normal forms at the singular points are best expressed in the coordinates for k and we first give the required formulas for this change of coordinates. However, having the change of coordinates into normal form in mind we already allow more general changes of coordinates Φ and Ψ in which the time variable may be changed according to $\tau = \tau(t,p)$ and $\bar{\tau} = \bar{\tau}(t,x)$. The structure of the map k is retained.

In the new coordinates the singular set is given by

$$S' = \left\{ (\tau,\xi) \in D' : \frac{\partial h}{\partial \xi_n}(\tau,\xi) = 0 \right\} \tag{4.2}$$

and the right-eigenvector to eigenvalue 0 is given by the coordinate vector field $(0, e_n')^T = (0, \frac{\partial}{\partial \xi_n})^T$. Let $(\rho, v)^T$ be the image of $(0, \frac{\partial}{\partial \xi_n})^T$ under this coordinate change, i.e.,

$$\begin{pmatrix} 0 \\ e_n' \end{pmatrix} = D\Phi(t,p) \begin{pmatrix} \rho(t,p) \\ v(t,p) \end{pmatrix}. \tag{4.3}$$

Since $\Psi \circ \sigma = k \circ \Phi$, it follows that

$$D\Psi(t,x) D\sigma(t,p) \begin{pmatrix} \rho(t,p) \\ v(t,p) \end{pmatrix} = \frac{\partial h}{\partial \xi_n}(\tau,\xi) \begin{pmatrix} 0 \\ e_n' \end{pmatrix} \tag{4.4}$$

and thus for $(t,p) \in S$, we have

$$0 = D\sigma(t,p) \begin{pmatrix} \rho(t,p) \\ v(t,p) \end{pmatrix} = \begin{pmatrix} \rho(t,p) \\ \frac{\partial x}{\partial p}(t,p) v(t,p) \end{pmatrix}. \tag{4.5}$$

Hence, v defines a $C^{1,r}$ time-varying vector field on D which gives the right-eigenvector to eigenvalue 0 in the singular set S. Analogously, we can define a $C^{1,r}$ time-varying covector field on D which gives the left-eigenvector w through the change of coordinates in the range applied to the left-eigenvector

$$\varpi(\tau,\xi) = \left(-\frac{\partial h}{\partial \tau}(\tau,\xi), \dots, -\frac{\partial h}{\partial \xi_{n-1}}(\tau,\xi), 1 \right) \tag{4.6}$$

for $Dk(\tau,\xi)$. Here we define

$$(\pi(t,p), w(t,p)) = \varpi(\bar{\tau},\eta) D\Psi(t, x(t,p)) \tag{4.7}$$

and obtain

$$w(t,p)\frac{\partial x}{\partial p}(t,p) = \frac{\partial h}{\partial \xi_n}(\tau,\xi)\frac{\partial \xi_n}{\partial p}(t,p). \tag{4.8}$$

Using these changes of coordinates it is possible to relate the transversality conditions in the different coordinate systems.

LEMMA 4.1. *If $\tau = t - t_0$, then for $(\tau,\xi) = \Phi(t,p) \in S'$ we have that*

$$\frac{\partial^2 h}{\partial \xi_n^2}(\tau,\xi) = w(t,p)\frac{\partial^2 x}{\partial p^2}(t,p)(v(t,p),v(t,p)), \tag{4.9}$$

$$\begin{aligned}\frac{\partial^3 h}{\partial \xi_n^3}(\tau,\xi) = w(t,p)\bigg(&\frac{\partial^3 x}{\partial p^3}(t,p)(v(t,p),v(t,p),v(t,p))\\ &+3\frac{\partial^2 x}{\partial p^2}(t,p)\left(\frac{\partial v}{\partial p}(t,p)v(t,p),v(t,p)\right)\bigg),\end{aligned} \tag{4.10}$$

$$\frac{\partial^2 h}{\partial \tau \partial \xi_n}(\tau,\xi) = w(t,p)\frac{\partial^2 x}{\partial t \partial p}(t,p)v(t,p). \tag{4.11}$$

This lemma is proved by direct calculations which are given in detail in [21]. Note that the assumption about the translation $\tau = t - t_0$ is sufficient to formulate the transversality conditions which then allow the transformation into the normal form. However, this transformation will require more general changes of coordinates in τ.

We now describe the normal forms for *fold* and *cusp* points, respectively. The map σ, respectively, x, has a *fold*-singularity at (t_0,p_0) if and only if the vector $(0,v_0)$ does not lie in the tangent space to S_1 at p_0, $(0,v_0)\notin T_{(t_0,p_0)}S_1$. Equivalently, S_1 and the kernel of $\frac{\partial x}{\partial p}(t_0,p_0)$ are transversal. In the coordinates on D' this is characterized by

$$\frac{\partial^2 h}{\partial \xi_n^2}(0,0) = \left(\frac{\partial^2 h}{\partial \tau \partial \xi_n}(0,0), \frac{\partial^2 h}{\partial \xi \partial \xi_n}(0,0)\right)\cdot\begin{pmatrix} 0 \\ e_n' \end{pmatrix} \neq 0 \tag{4.12}$$

or by

$$w_0\frac{\partial^2 x}{\partial p^2}(t_0,p_0)(v_0,v_0) \neq 0 \tag{4.13}$$

in the original coordinates. It follows from Whitney's results (see, for instance [20, Chap. III, Thm. 4.5]) that we can change the last coordinate on R' so that σ is given by the normal form

$$n : (\tau,\xi_1,\dots,\xi_n) \mapsto (\tau,\xi_1,\dots,\xi_{n-1},\xi_n^2), \tag{4.14}$$

i.e., $h(\tau,\xi) = \xi_n^2$.

The map σ has a *cusp*-singularity at (t_0,p_0) if and only if the eigenvector $(0,v_0)$ lies in the tangent space to S_1 at p_0, $(0,v_0) \in T_{(t_0,p_0)}S_1$, which is equivalent to the geometric fact that the (one-dimensional) kernel of $\frac{\partial x}{\partial p}(t_0,p_0)$ is tangent to S_1. Thus,

$$\frac{\partial^2 h}{\partial \xi_n^2}(0,0) = w_0\frac{\partial^2 x}{\partial p^2}(t_0,p_0)(v_0,v_0) = 0. \tag{4.15}$$

In this case the precise normal form of the function h is determined by the order of contact between the kernel of $\frac{\partial x}{\partial p}(t_0,p_0)$ and S_1. This leads to a further classification

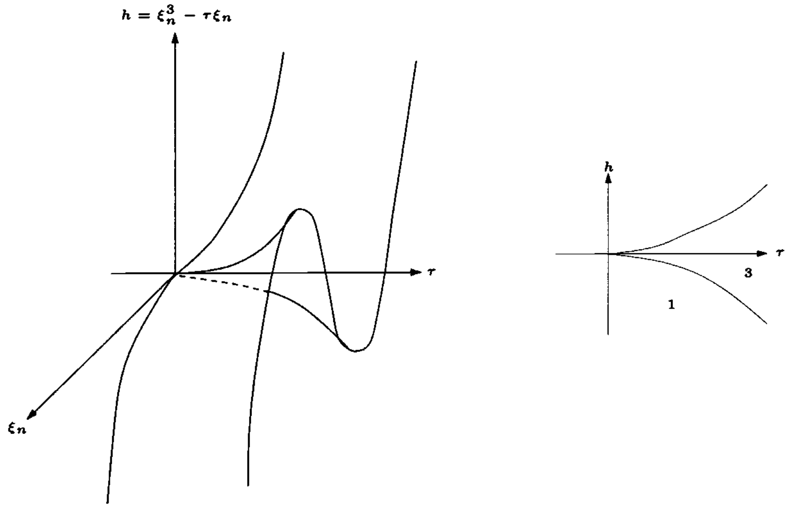


FIG. 4.1. *Simple cusp.*

of these singularities of S_1 into subclasses S_{1_k}, called *Morin*-singularities. Here we are interested only in the case when this order of contact is one, the so-called simple cusp. Even more specifically, we assume that the map σ takes on the singularity transversely in the following sense:

$$\frac{\partial^2 h}{\partial\tau\partial\xi_n}(0,0) \neq 0 \qquad \text{and} \qquad \frac{\partial^3 h}{\partial\xi_n^3}(0,0) \neq 0, \tag{4.16}$$

i.e.,

$$0 \neq w_0 \frac{\partial^2 x}{\partial t \partial p}(t_0, p_0)(v_0, v_0), \tag{4.17}$$

$$0 \neq w_0 \left(\frac{\partial^3 x}{\partial p^3}(t_0, p_0)(v_0, v_0, v_0) + 3\frac{\partial^2 x}{\partial p^2}(t_0, p_0)\left(\frac{\partial v}{\partial p}(t_0, p_0)v_0, v_0 \right) \right). \tag{4.18}$$

Condition (4.17) implies that S_1 is an $(n-1)$-dimensional manifold near (t_0, p_0), while (4.18) states that the order of contact between S_1 and $\ker\frac{\partial x}{\partial p}(t_0, p_0)$ is one. Here [20, Chap. VII, Thm. 4.1.] one can change the last coordinate on R' so that σ is indeed given by the normal form with function $h(\tau, \xi) = \xi_n^3 - \tau\xi_n$:

$$n : (\tau, \xi_1, \dots, \xi_n) \mapsto (\tau, \xi_1, \dots, \xi_{n-1}, \xi_n^3 - \tau\xi_n)). \tag{4.19}$$

Figure 4.1 gives the well-known picture of the simple cusp. It consists of a family of cubic polynomials in ξ_n parametrized by τ. These polynomials have only the real root $\xi_n = 0$ for $\tau < 0$, but three real roots for $\tau > 0$ and, correspondingly, a local minimum and maximum which generate a region where the polynomial is 3:1. Figure 4.1(a) shows the graph of the function h projected into the (τ, ξ_n, h) subspace while Figure 4.1(b) identifies the multiplicities of the images in the (ξ_n, h)-plane. Notice that the image of the singular set in the (ξ_n, h)-plane is a cusp which separates a region where the map is 1:1 from the region where the map is 3:1. On the singular set itself the map is 2:1.

5. The fold singularity in optimal control: Conjugate points. We now describe the structure of the value-function in the state-space for a parametrized flow of extremals near a fold point and its implications on the structures of optimal trajectories. The shadow price lemma allows us to relate the singularity in the flow of extremals to a singularity in the value-function for the corresponding parametrized flow. These considerations are in the sense that we analyze only the given parametrized flow, but do not analyze possible other flows of extremals which can overlap in the same region in the state-space. The theoretical considerations here provide what corresponds to small pieces of a puzzle which gives the required local (in a neighborhood of a reference *point*) structures which then can be pieced together to effect a *regular synthesis.*

We assume that

(A) The matrix $\frac{\partial x}{\partial p}(t_0, p_0)$ has a simple eigenvalue 0 with right-eigen-vector v_0 and left-eigenvector $w_0 = v_0^T \frac{\partial \lambda}{\partial p}(t_0, p_0) \neq 0$.

(F) The transversality condition

$$w_0 \frac{\partial^2 x}{\partial p^2}(t_0, p_0)(v_0, v_0) \neq 0 \tag{5.1}$$

holds.

THEOREM 5.1. *Suppose conditions* (A) *and* (F) *are met at* (t_0, p_0) *for a* $C^{1,3}$ *parametrized flow of normal extremals. Then there exist (open) neighborhoods* D *of* (t_0, p_0) *and* R *of* $(t_0, x_0) = (t_0, x(t_0, p_0))$ *with the following properties:*

(a) *The singular set* S *restricted to* D *is an embedded* n*-dimensional submanifold of fold points which splits* D *into two connected components* D_+ *and* D_-, $D = D_- \cup S \cup D_+$.

(b) *The map* $\sigma : D \to R$, $(t, p) \longmapsto (t, x(t, p))$, *restricted to* D_+ *or* D_- *is a* $C^{1,3}$ *diffeomorphism and both restrictions map* D_+, *respectively,* D_- *onto a region* $R_+ \subset R$,

$$\sigma(D_+) = R_+ = \sigma(D_-). \tag{5.2}$$

Thus, away from the singular set the map σ *is* 2:1 *on* D.

(c) *Let* $\sigma_\pm^{-1} : R_+ \to D_\pm$, *denote the inverses to the restriction of* σ *to* $D_\pm$ *and define the corresponding sections of the value-function for the parametrized flow by*

$$V_\pm : R_+ \to \mathbb{R}, \qquad V_\pm = C \circ \sigma_\pm^{-1}. \tag{5.3}$$

These functions and their gradients can be extended continuously to the fold $F = \sigma(S)$. *The graphs of* V_+ *and* V_- *are tangent on* F *but do not intersect otherwise over* R_+.

Thus the corresponding graphs of V^+ and V^- over R_+ do not intersect near (t_0, x_0), $x_0 = x(t_0, p_0)$, other than on the fold-locus. In other words, V^+ entirely lies to one side of V^-. The structure of the parametrized flow of extremals and the corresponding value-function is summarized in Figure 5.1.

Theorem 5.1 gives a geometric interpretation for what is essentially a classical result. For if the flow of the parametrized extremals is a diffeomorphism until the fold points are encountered, this structure implies the well-known results about strong local extrema of extremal trajectories and conjugate points. In this case a smooth solution to the Hamilton–Jacobi–Bellman equation is obtained in the state-space up to

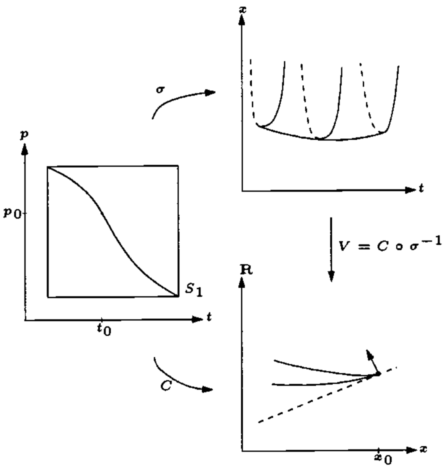


FIG. 5.1. *Value near a fold.*

the hypersurface F defined by the image of the fold-locus. Each of the parametrized extremals is optimal with respect to any other trajectory which lies in the region covered by the parametrized flow of extremals, hence is a *strong local extremum* in the sense of the calculus of variations. At the fold-points optimality ceases. Thus *the fold-points correspond to conjugate points* in the sense of the classical results for strong local optimality. This is exactly the structure as it is already encountered in the calculus of variations in the problem of minimum surfaces of revolution for smooth extremals, the catenaries [4]. As discussed in the introduction, in optimal control numerous formulations of similar results are well known, ranging from engineering textbooks [8] to modern geometric formulations expressed in terms of symplectic geometry [1]. Theorem 5.1 is included here for the sake of completeness since it describes the behavior for the least degenerate singularity and thus the most common degenerate behavior. Its proof within our framework of parametrized extremals is elementary, yet cleverly uses direct calculations in normal form. Its proof, which can be found in [21], is omitted here since the result itself is classical. Instead we illustrate Theorem 5.1 for a simple but instructive example.

Example. Consider the one-dimensional control problem to minimize

$$J(u) = \frac{1}{2}\int_0^T u^2 dt + \frac{1}{3}x(T)^3, \tag{5.4}$$

subject to $\dot{x} = u$, $u \in \mathbb{R}$. The minimum condition implies that $u = -\lambda$ and thus, taking $p = x(T)$, a smoothly parametrized family of extremals is given by

$$x(t,p) = p + (T-t)p^2, \quad u(t,p) = -p^2, \quad \lambda(t,p) = p^2. \tag{5.5}$$

Hence

$$S = \{(t,p) \in [0,T] \times \mathbb{R} : 2(T-t)p + 1 = 0\} \tag{5.6}$$

and all points in S are fold points.

6. The simple cusp singularity in optimal control: Cut-loci. In a regular synthesis, however, the picture of trajectories which stay optimal up to a surface of conjugate points or to a fold-singularity is not a familiar one. The reason lies in the presence of simple-cusp points which, as will be seen now, determine the local behavior of optimal trajectories in a neighborhood of the cusp point.

As above, we make the corank 1 singular point assumption (A). In particular, it follows from the discussions in section 4 that there exists a $C^{1,r}$ time-varying vector field v defined near (t_0, p_0) with the property that v is a right-eigenvector to eigenvalue 0 in the singular set. We now assume that

(C) If v_0 denotes the value at (t_0, p_0) of v, then

$$w_0 \frac{\partial^2 x}{\partial p^2}(t_0, p_0)(v_0, v_0) = 0 \tag{6.1}$$

and the following transversality conditions hold:

$$0 \neq w_0 \frac{\partial^2 x}{\partial t \partial p}(t_0, p_0) v_0, \tag{6.2}$$

$$0 \neq w_0 \left(\frac{\partial^3 x}{\partial p^3}(t_0, p_0)(v_0, v_0, v_0) + 3 \frac{\partial^2 x}{\partial p^2}(t_0, p_0)(\frac{\partial v}{\partial p}(t_0, p_0) v_0, v_0) \right). \tag{6.3}$$

We first summarize the mapping properties of the map σ near (t_0, p_0). The notion of stratifications and compatible maps provides a precise formulation to describe the multiplicities of the map σ near (t_0, p_0).

Definition 6.1. *Let M be a C^r manifold. A C^r stratification $\mathcal{S}$ of M is a locally finite decomposition of M into pairwise disjoint connected embedded C^r manifolds S_i, $i \in I$, which satisfies the so-called frontier-axiom, i.e., if S is an element of $\mathcal{S}$, then the frontier of S, Fron $S = (Clos\, S) \backslash S$, is a union of other elements of $\mathcal{S}$ which have lower dimension. The elements of $\mathcal{S}$ are called strata.*

Definition 6.2. *Let M be a C^r manifold and let N be an embedded C^r submanifold. A C^r stratification $\mathcal{S}$ of M is said to be compatible with N if N is a union of strata.*

Definition 6.3. *Let M and N be C^r manifold and let $F : M \to N$ be a C^r map. Let $\mathcal{S}$ and $\mathcal{T}$ be C^r stratifications of M and N, respectively. We say the stratifications $\mathcal{S}$ and $\mathcal{T}$ are compatible with the map F if for every $S \in \mathcal{S}$ there exists a $T \in \mathcal{T}$ such that the restriction of F to S is a C^r diffeomorphism onto T.*

We summarize the crucial mapping properties of σ near a simple cusp point in terms of these definitions.

Proposition 6.1. *Suppose conditions* (A) *and* (C) *hold at (t_0, p_0) for a $C^{1,4}$ parametrized flow of normal extremals. Then there exist (open) neighborhoods D of (t_0, p_0) and R of $(t_0, x_0) = (t_0, x(t_0, p_0))$ such that there exist stratifications $\mathcal{D}$ of D compatible with the singular set S, and $\mathcal{R}$ of R which are compatible with the map σ. Precisely, on D we have that*

(a) *The singular set S restricted to D is an n-dimensional embedded submanifold of corank 1 singular points which splits D into two connected components D_0 and $\widetilde{D_1}$. There exists an $(n-1)$-dimensional submanifold S_0 embedded into S which consists of simple cusp points and splits S into two connected components $S_\pm$ which consist of fold points.*

(b) *There exists an n-dimensional embedded submanifold T which is tangent to S at S_0 such that S_0 splits T into two connected components T_+ and T_- which are contained in $\widetilde{D_1}$. The map σ is* 1:1 *on the submanifolds S_0, $S_\pm$ and $T_\pm$ with images*

$$\sigma(S_0) =: J_0 \qquad \sigma(S_\pm) = \sigma(T_\pm) =: J_\pm. \tag{6.4}$$

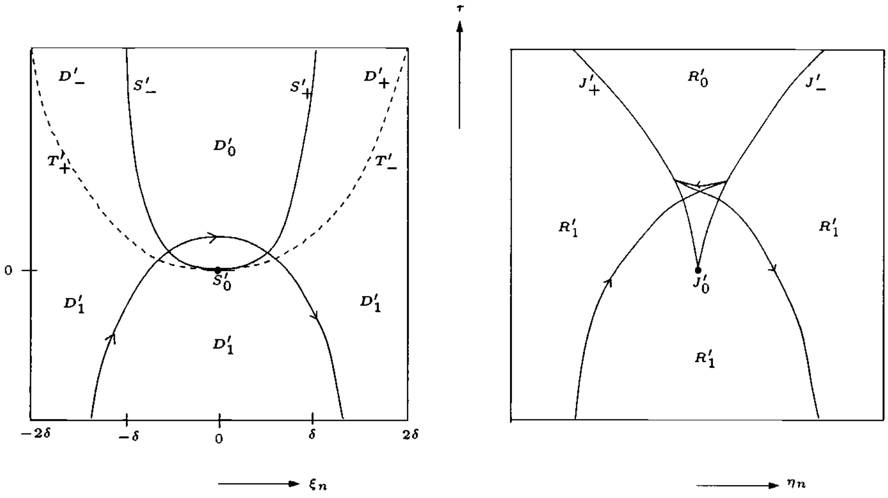


FIG. 6.1. *Compatible stratification for the simple-cusp.*

The submanifold T stratifies $\widetilde{D_1}$ into three connected components D_1, D_+, and D_-; σ restricted to each of the open submanifolds D_0, D_1, D_+, and D_- is a $C^{1,4}$ diffeomorphism with the following images

$$(6.5) \qquad \sigma(D_+) = \sigma(D_0) = \sigma(D_-) =: R_0 \qquad \sigma(D_1) =: R_1.$$

The sets J_0, $J_\pm$, R_0, and R_1 define the stratification $\mathcal{R}$ of R.

Thus the map σ is 1:1 on D_1 and S_0 onto R_1 and J_0, 2:1 on $T_\pm \cup S_\pm$ onto $J_\pm$, and 3:1 from $D_- \cup D_0 \cup D_+$ onto R_0. The structure of the stratifications for the domain and range for the map σ are illustrated in Figure 6.1.

The mapping properties of σ near a simple-cusp point are well known and hence the proof is omitted. The nontrivial aspect of the construction lies in relating them to the mapping properties of the parametrized value-function near the simple-cusp.

THEOREM 6.1. *Suppose conditions* (A) *and* (C) *hold at* (t_0,p_0) *for a* $C^{1,4}$ *parametrized flow of normal extremals and let $\mathcal{D}$ and $\mathcal{R}$, respectively, be the stratifications of the (open) neighborhoods D of (t_0,p_0) and R of $(t_0,x_0) = (t_0,x(t_0,p_0))$ constructed in Proposition* 6.1. *Let $\sigma_\kappa^{-1} : R_+ \to D_\kappa$, $\kappa \in \{-1,0,1\}$, denote the inverses to the restriction of σ to D_κ and define the corresponding sections of the value-function for the parametrized flow by*

$$(6.6) \qquad V_\kappa : R_+ \to \mathbb{R}, \qquad V_\kappa = C \circ \sigma_\kappa^{-1}.$$

These functions and their gradients can be extended continuously to the fold submanifolds $J_\pm$ in the respective domains and the graphs of $V_\pm$ and V_0 are tangent on $J_\pm$, but do not intersect otherwise over R_+. However, V_+ and V_- intersect in a cut-locus Γ which is an n-dimensional embedded submanifold with J_0 in its frontier.

The parametrized value-function is shown in Figure 6.2.

Proof: Since the frontier strata $J_\pm$ of R_0 consist of fold points, the statements about continuous extensions of V_κ and the corresponding gradients onto $J_\pm$ follow from Theorem 5.1. We need to analyze the values $V_\kappa(t,x)$ for a point $(t,x) \in R_0$. Let

$$(6.7) \qquad (t,x) = \sigma_\kappa(t,p_\kappa), \qquad (\tau,\xi_\kappa) = \Phi(t,p_\kappa).$$

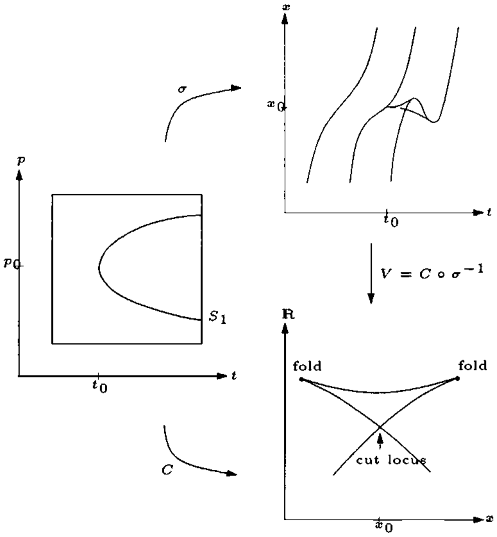


FIG. 6.2. *Value near a simple cusp point.*

Then, for $C' = C \circ \sigma^{-1}$ and $i, j \in \{-1, 0, 1\}$ we have

$$V_i(t,x) - V_j(t,x) = C(t,p_i) - C(t,p_j) = C'(\tau,\xi_i) - C'(\tau,\xi_j) \tag{6.8}$$

and we will analyze the cost in normal coordinates. We first establish the values of the derivatives of $x' = x \circ \sigma^{-1}$ and C' on the cusp surface S_0'. Lemmas 6.1 through 6.3 follow by direct calculations making the necessary change of coordinates into normal form. We only indicate the proofs. Full details can be found in [21].

LEMMA 6.1. *For $(\tau,\xi) \in S'$ we have*

$$\frac{\partial x'}{\partial \xi_n}(\tau,\xi) = 0, \quad \frac{\partial t'}{\partial \xi_n}(\tau,\xi) = 0. \tag{6.9}$$

On the cusp surface S_0' we have

$$\frac{\partial^2 x'}{\partial \xi_n^2}(\tau,\xi) = 0, \quad \frac{\partial^2 t'}{\partial \xi_n^2}(\tau,\xi) = 0, \tag{6.10}$$

$$\frac{\partial^4 x'}{\partial \xi_n^4}(\tau,\xi) = 0, \quad \frac{\partial^4 t'}{\partial \xi_n^4}(\tau,\xi) = 0, \tag{6.11}$$

$$6\frac{\partial^2 x'}{\partial \tau \partial \xi_n}(\tau,\xi) + \frac{\partial^3 x'}{\partial \xi_n^3}(\tau,\xi) = 0, \tag{6.12}$$

$$6\frac{\partial^2 t'}{\partial \tau \partial \xi_n}(\tau,\xi) + \frac{\partial^3 t'}{\partial \xi_n^3}(\tau,\xi) = 0. \tag{6.13}$$

Proof. Let $\sigma' = \sigma \circ \Phi^{-1}$. Then $t' = t(\tau,\xi)$ and $\Psi \circ \sigma' = n$ implies that

$$D\Psi(t,x)\begin{pmatrix} \frac{\partial t'}{\partial \xi_n}(\tau,\xi) \\ \frac{\partial x'}{\partial \xi_n}(\tau,\xi) \end{pmatrix} = \frac{\partial h}{\partial \xi_n}(\tau,\xi)\begin{pmatrix} 0 \\ e_n' \end{pmatrix}, \tag{6.14}$$

which for $(\tau, \xi) \in S'$ gives (6.9). The other relations follow by differentiating (6.14). □

In order to evaluate the derivatives of the cost C' on S'_0 we need to account for the fact that the change into normal form for the simple-cusp also requires that we change the time variable as well, $\tau = \tau(t, p)$. In the new coordinates, the shadow price lemma transforms as follows.

Lemma 6.2.

$$\frac{\partial C'}{\partial \xi_n}(\tau, \xi) = \lambda'(\tau, \xi) \frac{\partial x'}{\partial \xi_n}(\tau, \xi) - H'(\tau, \xi) \frac{\partial t'}{\partial \xi_n}(\tau, \xi), \tag{6.15}$$

where

$$H'(\tau, \xi) = H(\tau, \lambda'(\tau, \xi), x'(\tau, \xi), u'(\tau, \xi)).$$

Proof. Differentiating $C' = C \circ \Phi^{-1}$ we obtain

$$\frac{\partial C'}{\partial \xi_n}(\tau, \xi) = \frac{\partial C}{\partial t}(t, p) \frac{\partial t}{\partial \xi_n}(\tau, \xi) + \frac{\partial C}{\partial p}(t, p) \frac{\partial p}{\partial \xi_n}(\tau, \xi). \tag{6.16}$$

Using Lemma 3.1 and the analogous formula for $\frac{\partial x'}{\partial \xi_n}(\tau, \xi)$, (6.15) follows. □

Lemma 6.2 allows us to evaluate the derivatives of C' on this singular set. By (6.9) we have

$$\frac{\partial C'}{\partial \xi_n}(\tau, \xi) = 0 \quad \text{for } (\tau, \xi) \in S'. \tag{6.17}$$

Lemma 6.3. *For $(\tau, \xi) \in S'_0$ we have*

$$\frac{\partial^2 C'}{\partial \xi_n^2}(\tau, \xi) = 0, \tag{6.18}$$

$$6 \frac{\partial^2 C'}{\partial \tau \partial \xi_n}(\tau, \xi) + \frac{\partial^3 C'}{\partial \xi_n^3}(\tau, \xi) = 0, \tag{6.19}$$

$$18 \frac{\partial^3 C'}{\partial \tau \partial \xi_n^2}(\tau, \xi) + \frac{\partial^4 C'}{\partial \xi_n^4}(\tau, \xi) = 0. \tag{6.20}$$

Proof. These relations follow by differentiating (6.16) and using Lemma 6.2 to evaluate the derivatives on S'_0. □

Having these basic relations available, we now evaluate the differences in the cost. First we blow up coordinates to desingularize the singular set near the cusp point via

$$\begin{aligned} \Theta : (-1, 1) \times (0, \delta) \ &\to \ (0, 3\delta^2) \times (-\delta, \delta), \\ (\alpha, \beta) &\longmapsto (\tau, \xi_n) = (3\beta^2, \alpha\beta). \end{aligned} \tag{6.21}$$

In these coordinates h is given by

$$\widetilde{h}(\alpha, \beta) = (\alpha\beta)^3 - 3\beta^2 \cdot \alpha\beta = \beta^3(\alpha^3 - 3\alpha)$$

and therefore the preimages of a point $(t,x) = \sigma'(\tau,\xi) \in R_0$ are given by the solutions to the cubic polynomial

$$0 = \mu^3 - 3\beta^2\mu - \beta^3(\alpha^3 - 3\alpha).$$

Using another blow-up $\mu = \beta h$, we can eliminate β and obtain the reduced equation

$$0 = h^3 - 3h - (\alpha^3 - 3\alpha) = (h-\alpha)(h^2 + \alpha h + \alpha^2 - 3). \tag{6.22}$$

The solutions can therefore be analyzed in time-slices $\tau = const$ and are given by the trivial root $h_0(\alpha) = \alpha$ corresponding to the preimage in D_0' and the two zeroes

$$h_\pm(\alpha) = -\frac{\alpha}{2} \pm \sqrt{3}\sqrt{1 - \left(\frac{\alpha}{2}\right)^2} \tag{6.23}$$

corresponding to roots in $D_\mp'$ (note the reversal of signs). The properties of the solutions h_κ, $\kappa \in \{-1,0,1\}$ given in the following lemma follow directly from the definition or are immediately verified.

Lemma 6.4.

$$h_\pm^2(\alpha) + \alpha h_\pm(\alpha) + \alpha^2 \equiv 3, \tag{6.24}$$

$$h_+^2(\alpha) + h_+(\alpha)h_-(\alpha) + h_-^2(\alpha) \equiv 3, \tag{6.25}$$

$$h_+(\alpha) + h_-(\alpha) + \alpha = 0. \tag{6.26}$$

We now compare the values of the functions V_κ, $\kappa \in \{-1,0,1\}$.

Proposition 6.2. *Let $(t,x) = \sigma_0'(\tau,\xi)$ and $(\tau,\xi_n) = (3\beta^2, \alpha\beta)$. Then*

$$\begin{aligned} V_0(t,x) - V_\pm(t,x) &= C'(3\beta^2,\alpha\beta) - C'(3\beta^2, h_\pm(\alpha)\beta) \\ &= \frac{1}{72}\frac{\partial^4 C'}{\partial \xi_n^4}(0,0)\left(\alpha - h_\pm(\alpha)\right)^3, \beta^4\left(\alpha + h_\pm(\alpha) + r_\pm(\alpha,\beta)\right), \end{aligned} \tag{6.27}$$

where $r_\pm(\alpha,\beta)$ is of order $o(1)$ uniformly for all $\alpha \in [-1,1]$, i.e., $\lim_{\beta\to 0+} r_\pm(\alpha,\beta) = 0$ uniformly in α over $[-1,1]$.

Proposition 6.3. *Let $(t,x) = \sigma_0'(\tau,\xi)$ and $(\tau,\xi_n) = (3\beta^2, \alpha\beta)$. Then*

$$\begin{aligned} V_+(t,x) - V_-(t,x) &= C'(3\beta^2, h_+(\alpha)\beta) - C'(3\beta^2, h_-(\alpha)\beta) \\ &= \frac{1}{72}\frac{\partial^4 C'}{\partial \xi_n^4}(0,0)\left(h_+(\alpha) - h_-(\alpha)\right)^3 \\ &\qquad \beta^4\left(h_+(\alpha) + h_-(\alpha) + r_0(\alpha,\beta)\right) \end{aligned} \tag{6.28}$$

where $r_0(\alpha,\beta)$ is of order $o(1)$ uniformly for all $\alpha \in [-1,1]$, i.e., $\lim_{\beta\to 0+} r_0(\alpha,\beta) = 0$ uniformly in α over $[-1,1]$.

The proofs of these propositions are similar and we prove only Proposition 6.3. Note the symmetry in the formulas which are identical if the roots are interchanged.

Proof. Writing C' as integral over its ξ_n derivative, we obtain

$$\begin{aligned} \Delta_0(\alpha,\beta) = V_+(t,x) - V_-(t,x) &= C'(3\beta^2, h_+(\alpha)\beta) - C'(3\beta^2, h_-(\alpha)\beta) \\ &= \frac{1}{2}\left(\int_{-1}^{1} \frac{\partial C'}{\partial \xi_n}(3\beta^2, \beta\gamma_0(\alpha,s))ds\right)(h_+(\alpha) - h_-(\alpha))\beta, \end{aligned}$$

where

$$\gamma_0(\alpha,s)=\frac{1}{2}(1+s)h_+(\alpha)+\frac{1}{2}(1-s)h_-(\alpha)=-\frac{\alpha}{2}+s\sqrt{3}\sqrt{1-\left(\frac{\alpha}{2}\right)^2}.$$

We now expand the integrand around $\beta=0$. Using Lemma 6.3 it follows that

$$\begin{aligned}\Delta_0(\alpha,\beta)=\frac{1}{2}\int_{-1}^{1}\Bigg\{&\left(3\frac{\partial^2 C'}{\partial\tau\partial\xi_n}(0,0)+\frac{1}{2}\frac{\partial^3 C'}{\partial\xi_n^3}(0,0)\gamma_0(\alpha,s)^2\right)\beta^2+\Bigg(3\frac{\partial^3 C'}{\partial\tau\partial\xi_n^2}(0,0)\gamma_0(\alpha,s)\\&+\frac{1}{6}\frac{\partial^4 C'}{\partial\xi_n^4}(0,0)\gamma_0(\alpha,s)^3\Bigg)\beta^3+o(\beta^3)ds\Bigg\}(h_+(\alpha)-h_-(\alpha))\beta\\=\frac{1}{2}\int_{-1}^{1}\Bigg\{&\frac{1}{2}\frac{\partial^3 C'}{\partial\xi_n^3}(0,0)\beta^2\left[\gamma_0(\alpha,s)^2-1\right]\\&+\frac{1}{6}\frac{\partial^4 C'}{\partial\xi_n^4}(0,0)\beta^3\left[-\gamma_0(\alpha,s)+\gamma_0(\alpha,s)^3\right]+o(\beta^3)ds\Bigg\}(h_+(\alpha)-h_-(\alpha))\beta.\end{aligned}$$

Now

$$\begin{aligned}\int_{-1}^{1}(\gamma_0(\alpha,s)^2-1)ds&=\int_{-1}^{1}\left[\frac{1}{4}\left(h_+(\alpha)+h_-(\alpha)+s(h_+(\alpha)-h_-(\alpha))\right)^2-1\right]ds\\&=-2+\frac{1}{2}\left(h_+(\alpha)+h_-(\alpha)\right)^2+\frac{1}{6}\left(h_+(\alpha)-h_-(\alpha)\right)^2\\&=-2+\frac{2}{3}\left(h_+(\alpha)^2+h_+(\alpha)h_-(\alpha)+h_-(\alpha)^2\right)=0\end{aligned}$$

and

$$\begin{aligned}\int_{-1}^{1}\gamma_0(\alpha,s)(\gamma_0(\alpha,s)^2-1)ds&=\frac{1}{2}\left(h_+(\alpha)-h_-(\alpha)\right)\int_{-1}^{1}s(\gamma_0(\alpha,s)^2-1)ds\\&=\frac{1}{2}\left(h_+(\alpha)-h_-(\alpha)\right)^2\left(h_+(\alpha)+h_-(\alpha)\right)\int_{-1}^{1}\frac{1}{2}s^2ds\\&=\frac{1}{6}\left(h_+(\alpha)-h_-(\alpha)\right)^2\left(h_+(\alpha)+h_-(\alpha)\right).\end{aligned}$$

Since $h_+(\alpha)-h_-(\alpha)=2\sqrt{3}\sqrt{1-\left(\frac{\alpha}{2}\right)^2}\geq 3$ for $\alpha\in[-1,1]$, (6.28) follows. This proves the proposition. □

For $\alpha\in[-1,1]$ we have $\alpha+h_+(\alpha)\in[1,2]$ and $\alpha+h_-(\alpha)\in[-2,-1]$. Since $h_-(\alpha)<\alpha<h_+(\alpha)$ on the open interval $(-1,1)$, it follows that $\Delta_\pm(\alpha,\beta)$ has constant sign on $(-1,1)\times(0,\delta)$ for δ sufficiently small. Since the sign of $(\alpha+h_\pm(\alpha))$ is given by ±1, this sign is given by the *nonzero* sign of $-\frac{\partial^4 C'}{\partial\xi_n^4}(0,0)$. To see this, note that for $(\tau,\xi)\in S_0'$

$$\frac{\partial^4 C'}{\partial\xi_n^4}(\tau,\xi)=3\frac{\partial\lambda'}{\partial\xi_n}(\tau,\xi)\frac{\partial^3 x'}{\partial\xi_n^3}(\tau,\xi)-3\frac{\partial H'}{\partial\xi_n}(\tau,\xi)\frac{\partial^3 t}{\partial\xi_n^3}(\tau,\xi)\tag{6.29}$$

and

$$\frac{\partial H'}{\partial\xi_n}(\tau,\xi)=\frac{\partial\lambda'}{\partial\xi_n}(\tau,\xi)\frac{\partial H'}{\partial\lambda'}(\tau,\xi)+\frac{\partial H'}{\partial x'}(\tau,\xi)\frac{\partial x'}{\partial\xi_n}(\tau,\xi)+\frac{\partial H'}{\partial u'}(\tau,\xi)\frac{\partial u'}{\partial\xi_n}(\tau,\xi).\tag{6.30}$$

But for $(\tau,\xi)\in S'$

$$\frac{\partial H'}{\partial u'}(\tau,\xi)\frac{\partial u'}{\partial \xi_n}(\tau,\xi)=\frac{\partial H}{\partial u}(t,p)\left(\frac{\partial u}{\partial t}(t,p)\frac{\partial t}{\partial \xi_n}(\tau,\xi)+\frac{\partial u}{\partial p}(t,p)\frac{\partial p}{\partial \xi_n}(\tau,\xi)\right)=0$$

since $\frac{\partial t}{\partial \xi_n}(\tau,\xi)=0$ and in general $\frac{\partial H}{\partial u}(t,p)\frac{\partial u}{\partial p}(t,p)\equiv 0$ by (3.16). In addition, since $\frac{\partial x'}{\partial \xi_n}(\tau,\xi)=0$ on S' we have that

$$\frac{\partial H'}{\partial \xi_n}(\tau,\xi)=\frac{\partial \lambda'}{\partial \xi_n}(\tau,\xi)f(\tau,x'(\tau,\xi),u'(\tau,\xi)). \tag{6.31}$$

Combining this with

$$\frac{\partial^3 x'}{\partial \xi_n^3}(\tau,\xi)=\frac{\partial x}{\partial t}(t,p)\frac{\partial^3 t}{\partial \xi_n^3}(\tau,\xi)+\frac{\partial^3 x}{\partial p^3}(t,p)\left(\frac{\partial p}{\partial \xi_n}(\tau,\xi),\frac{\partial p}{\partial \xi_n}(\tau,\xi),\frac{\partial p}{\partial \xi_n}(\tau,\xi)\right) \tag{6.32}$$

$$+3\frac{\partial^2 x}{\partial p^2}(t,p)\left(\frac{\partial p}{\partial \xi_n}(\tau,\xi),\frac{\partial^2 p}{\partial \xi_n^2}(\tau,\xi)\right)+\frac{\partial x}{\partial p}(t,p)\frac{\partial^3 p}{\partial \xi_n^3}(\tau,\xi), \tag{6.33}$$

it follows that for $(\tau,\xi)\in S_0'$

$$\frac{\partial^4 C'}{\partial \xi_n^4}(\tau,\xi)=3v^T(t,p)\frac{\partial \lambda}{\partial p}(t,p)\left[\frac{\partial^3 x}{\partial p^3}(t,p)\cdot(v(t,p),v(t,p),v(t,p))\right.$$

$$\left.+3\frac{\partial^2 x}{\partial p^2}(t,p)\cdot\left(\frac{\partial v}{\partial p}(t,p)v(t,p),v(t,p)\right)\right], \tag{6.34}$$

which is nonzero by the transversality condition (6.3).

Proposition 6.2 therefore implies that, away from the submanifolds $S_\pm$—where the graphs of V_0 and $V_\pm$ are tangent—the graphs of $V_\pm$ lie either below or above the graph of V_0 depending on the sign in the transversality condition (6.3). Proposition 6.3 allows us to compare V_+ and V_-. It follows from this proposition and the identity (6.26) that V_+ and V_- intersect if and only if

$$\delta(\alpha,\beta)=\alpha-r_0(\alpha,\beta)=0. \tag{6.35}$$

Since $\lim_{\beta\searrow 0^+} r_0(\alpha,\beta)=0$, it follows that the only solution in the limit $\beta\searrow 0$ is $\alpha=0$. Furthermore, $\lim_{\beta\searrow 0^+}\delta(1,\beta)=1$ and $\lim_{\beta\searrow 0^+}\delta(-1,\beta)=-1$ so that we have for some sufficiently small $\beta_0>0$

$$\delta(1,\beta)>0 \quad\text{and}\quad \delta(-1,\beta)<0,\quad 0<\beta<\beta_0. \tag{6.36}$$

Hence, by the mean-value theorem there exists a zero in $\alpha\in(-1,1)$ for all $\beta\in(0,\beta_0]$. But

$$\frac{\partial \delta}{\partial \alpha}(\alpha,\beta)=1-\frac{\partial r_0}{\partial \alpha}(\alpha,\beta) \tag{6.37}$$

and the limiting properties of $r_0(\alpha,\beta)$ extend to $\frac{\partial r_0}{\partial \alpha}(\alpha,\beta)$, since $r_0(\alpha,\beta)$ is C^4 in α on $[-1,1]$. Thus, $\frac{\partial \delta}{\partial \alpha}(\cdot,\beta)$ is positive and hence, $\delta(\cdot,\beta)$ is strictly increasing over $[-1,1]$. The unique zero $\alpha=A(\beta)$ defines the cut-locus in the normal coordinates. By the implicit function theorem the map $\beta\mapsto A(\beta)$ is C^1 on $(0,\beta_0)$. The curve $\alpha=A(\beta)$ separates $[-1,1]\times(0,\beta_0)$ into the open subsets $\{\alpha<A(\beta)\}$ and $\{\alpha>A(\beta)\}$ and on each one of these $\Delta_0(\alpha,\beta)$ has a constant sign. The diffeomorphic image of the n-dimensional embedded submanifold

$$(\beta, \xi_1, \dots, \xi_{n-1}) \mapsto (3\beta^2, \xi_1, \dots, \xi_{n-1}, A(\beta)\beta) \tag{6.38}$$

under the map $\sigma' = \sigma \circ \Phi^{-1}$ defines the cut-locus Γ in R or (t,x)-space. This proves Theorem 6.1.

This structure has several interesting and important implications on the structure of an optimal synthesis near the image (t_0, x_0) of a simple-cusp point (t_0, p_0). Trajectories $x(\cdot, p)$ different from $x(\cdot, p_0)$ lose optimality in a neighborhood V of (t_0, x_0) at the cut-locus prior to reaching the conjugate point at the fold-locus. Even though these trajectories are strong local extrema until they reach the fold-locus, the portion after the cut-locus will not be part of any optimal synthesis. These portions are no longer optimal since there are better trajectories (namely the ones coming from $D_\pm$ if the original trajectories are from $D_\mp$) which, however, are not close to the original trajectories in the sense of a calculus of variations. This feature and the structure of optimal trajectories are exactly the same, as it has been developed in [22] and is described in detail in [32], for the codimension 0 case of time-optimal control to an equilibrium point in $\mathbb{R}^3$.

Also note that it follows from Proposition 6.2 that the graph of V_0 lies above the graphs of V_+ and V_- if and only if

$$\frac{\partial^4 C'}{\partial \xi_n^4}(0,0) < 0, \tag{6.39}$$

and in this case, the minimizing branches generate a viscosity solution to the Hamilton–Jacobi–Bellman equation. If $\frac{\partial^4 C'}{\partial \xi_n^4}(0,0) > 0$, then this solution will be discontinuous. However, since we consider only one parametrized flow of extremals and not the optimal control problem per se, it is not a priori possible to connect the second case to the value function of the optimal control problem.

Example (see [10]). Consider the one-dimensional control problem to minimize

$$J(u) = \frac{1}{2}\int_0^T u^2 dt + \frac{1}{2}\left(x(T)^4 - x(T)^2\right) \tag{6.40}$$

subject to $\dot{x} = u$, $u \in \mathbb{R}$. As before, $u = -\lambda$, and now with $p = x(T)$, a smoothly parametrized family of extremals is given by

$$x(t,p) = p - (T-t)(p - 2p^3), \quad u(t,p) = p - 2p^3 = -\lambda(t,p). \tag{6.41}$$

Hence,

$$S = \{(t,p) \in (-\infty, T] \times \mathbb{R} : 6(T-t)p^2 - (T-t) + 1 = 0\}. \tag{6.42}$$

Here x_{pp} vanishes for $p = 0$ while x_{tp} and x_{ppp} are nonzero at $p = 0$. Thus the point $(T-1, 0)$ is a simple cusp point while all other singular points are fold points.

7. Conclusion.

This structure of the parametrized value near a simple-cusp point is interesting since it confirms that it is the more degenerate singularity, the simple-cusp, which dominates the local behavior at the point over the less degenerate fold-points nearby. Indeed the fold points become irrelevant and are not part of an optimal (regular) synthesis. This, of course, immediately raises the question about the roles of even more degenerate singularities. The results of [22, 32] verify, however, that in dimension three the structure as it was described here is the typical one (under codimension 0 assumptions on the Lie-brackets) also for time-optimal control to an equilibrium point for a nonlinear system. This is consistent with the fact that the fold and simple cusp singularities are the only generic singularities for maps between two-dimensional manifolds (with time generating the third dimension in the parametrization). For higher-dimensional systems, however, this suggests that the more degenerate singularities will play the decisive role. Naturally, the more degenerate the singularities are, the more difficult they are to analyze. Still, an analysis of the Morin-singularities [20, Chap. VII, Thm. 4.1] does not appear to be impossible and the geometry of some of the more degenerate ones like the swallowtail- or butterfly-catastrophe seems to be well understood [29]. Furthermore, generically the most degenerate singularities occur only in isolated points and thus, having the construction of a regular synthesis in mind, it may only be necessary to construct a few local syntheses near reference points. However, in order to analyze higher-dimensional systems, one probably needs to come up with the general structure relating singularities in the parametrized flow of extremals to the corresponding local solutions of the Hamilton–Jacobi–Bellman equation. Looking at the results of this paper, such a relation may indeed exist. Both for the fold and the simple-cusp singularity, the graph of the parametrized value exhibits the structure of the singular set of the next degenerate singularity, i.e., of a cusp for the fold and of a swallow-tail for the simple-cusp. Whether this is a mere coincidence or part of a general pattern which generally holds for transversal S_{1_k} singularities remains to be seen.

Acknowledgment. We would like to thank two anonymous referees for their valuable suggestions.

REFERENCES

[1] A. Agrachev and R. Gamkrelidze, *Symplectic methods for optimization and control*, in Geometry of Feedback and Optimal Control, B. Jakubczyk and W. Respondek, eds., Monogr. Textbooks Pure Appl. Math. 207, Dekker, New York, 1998, pp. 19–77

[2] V. I. Arnold, S. M. Gusein-Zade, and A. N. Varchenko, *Singularities of differentiable maps*, vol. I, 1985.

[3] L. Berkovitz, *Optimal Control Theory*, Springer-Verlag, New York, 1974.

[4] G. Bliss, *Calculus of Variations*, Open Court Publishing Co., Chicago, 1925.

[5] V. G. Boltyanskii, *Sufficient conditions for optimality and the justification of the dynamic programming method*, SIAM J. Control, 4 (1966), pp. 326–361.

[6] W. M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.

[7] P. Brunovsky, *Existence of regular synthesis for general control problems*, J. Differential Equations, 38 (1980), pp. 317–343.

[8] A. E. Bryson, Jr. and Y. C. Ho, *Applied Optimal Control*, revised printing, Hemisphere Publishing Company, New York, 1975.

[9] C. I. Byrnes and H. Frankowska, *Unicité des solutions optimales et absence de chocs pour les équations d'Hamilton-Jacobi-Bellman et de Riccati*, C.R. Acad. Sci. Paris, t. 315, Série I, 1992, pp. 427–431.

[10] C. I. Byrnes and A. Jhemi, *Shock waves for Riccati Partial Differential Equations arising in nonlinear optimal control*, in Systems, Models and Feedback: Theory and Applications, A. Isidori and T. J. Tarn, eds., Birkhäuser, Cambridge, MA, 1992, pp. 211–225.

[11] P. Cannarsa and H. M. Soner, *On the singularities of the viscosity solutions to Hamilton-Jacobi-Bellman equations*, Indiana Univ. Math J., 36 (1987), pp. 501–524.

[12] P. Cannarsa and H. Frankowska, *Some characterizations of optimal trajectories in control theory*, SIAM J. Control Optim., 29 (1991), pp. 1322–1347.

[13] C. Carathéodory, *Variationsrechnung und Partielle Differential Gleichungen erster Ord-*

nung, Teubner-Verlag, Leipzig, 1936.

[14] N. Caroff and H. Frankowska, *Conjugate points and shocks in nonlinear optimal control*, Trans. Amer. Math. Soc., 348 (1996), pp. 3133–3153.

[15] L. Cesari, *Optimization - Theory and Applications*, Springer-Verlag, New York, 1983.

[16] C. M. Dafermos, *Generalized characteristics and the structure of solutions of hyperbolic conservation laws*, Indiana Univ. Math. J., 26 (1977), pp. 1097–1119.

[17] W. H. Fleming, *The Cauchy problem for a nonlinear first order differential equation*, J. Differential Equations, 5 (1969), pp. 515–530.

[18] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[19] W. Fleming and M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[20] M. Golubitsky and V. Guillemin, *Stable Mappings and Their Singularities*, Springer-Verlag, New York, 1973.

[21] M. Kiefer, *On Singularities in Solutions to the Hamilton-Jacobi-Bellman Equation and Their Implications for the Optimal Control Problem*, D.Sc. Thesis, Washington University, St. Louis, MO, 1997.

[22] A. J. Krener and H. Schättler, *The structure of small-time reachable sets in low dimension*, SIAM J. Control Optim., 27 (1989), pp. 120–147.

[23] I. A. Kupka, *Geometric theory of extremals in optimal control problems* I: *The fold and Maxwell cases*, Trans. Amer. Math. Soc., 299 (1987), pp. 225–243.

[24] J. Noble, *Parametrized families of broken extremals and sufficient conditions for relative minima*, D.Sc. Thesis, Washington University, St. Louis, MO, 1998.

[25] A. Nowakowski, *Field theories in the modern Calculus of Variations*, Trans. Amer. Math. Soc., 309 (1988), pp. 725–752.

[26] B. Piccoli, *Classification of generic singularities for the planar time-optimal synthesis*, SIAM J. Control Optim., 36 (1996), pp. 1914–1946.

[27] B. Piccoli and H. J. Sussmann, *Regular synthesis and sufficiency conditions for optimality*, to appear.

[28] L. S. Pontryagin, V. G. Boltyanskii, V. G. Gamkrelidze, and R. V. Mishchenko, *Mathematical Theory of Optimal Processes*, Wiley-Interscience, New York, 1962.

[29] T. Poston and I. N. Stewart, *Taylor Expansions and Catastrophes, Res. Notes Math.* 7, Pitman Publishing, London, 1977.

[30] H. Schättler, *Hinreichende Bedingungen für ein starkes relatives Minimum bei Kontrollproblemen*, Diplomarbeit der Fakultät für Mathematik der Universität Würzburg, 1982.

[31] H. Schättler, *A local feedback synthesis of time-optimal stabilizing controls in dimension three*, Math. Control Signals Systems, 4 (1991), pp. 293–313.

[32] H. Schättler, *Extremal trajectories, small-time reachable sets and local feedback synthesis: a synopsis of the three-dimensional case*, in Nonlinear Synthesis, Christopher I. Byrnes and Alexander Kurzhansky, eds., Birkhäuser, Boston, 1991, pp. 258–269,

bibitemForum H. Schättler and M. Jankovic, *A synthesis of time-optimal controls in the presence of saturated singular arcs*, Forum Math., 5 (1993), pp. 203–241.

[33] H. J. Sussmann, *Synthesis, presynthesis, sufficient conditions for optimality and subanalytic sets*, in Nonlinear Controllability and Optimal Control, H. Sussmann, ed., Marcel Dekker, New York, 1990, pp. 1–19.

[34] H. J. Sussmann, *The structure of time-optimal trajectories for single-input systems in the plane: The C^∞ nonsingular case*, SIAM J. Control Optim., 25 (1987), pp. 433–465.

[35] H. Sussmann, *The structure of time-optimal trajectories for single-input systems in the plane: The general real analytic case*, SIAM J. Control Optim., 25 (1987), pp. 868–904.

[36] H. Sussmann, *Regular synthesis for time-optimal control of single-input real analytic systems in the plane*, SIAM J. Control Optim., 25 (1987), pp. 1145–1162.

[37] R. Thom, *Les singularités des applications différentiables*, Ann. Inst. Fourier, 6 (1955-56), pp. 43–87.

[38] H. Whitney, *Elementary structure of real algebraic varieties*, Ann. of Math., 66 (1957), pp. 545–556.

[39] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, W.B. Saunders, Philadelphia, PA, 1969.

SIAM J. Control Optim.
Vol. 37, No. 5, pp. 1372–1393

ON PARAMETER ESTIMATION USING LEVEL SETS*

JORDAN M. BERG† AND KENNETH HOLMSTRÖM‡

Abstract. Consider the problem of selecting the member of a parametrized family of curves that best matches a given curve. This is a key step in determining proper values for adjustable parameters in low-order plasma etching and deposition models. Level set methods offer several attractive features for treating such problems. This paper presents a parameter estimation scheme that exploits the level set formulation. The method is completely geometric; there is no need to introduce an arbitrary coordinate system for the curves. Analytic results necessary for the application of gradient descent algorithms are derived, and some preliminary numerical results are presented.

Key words. curve evolution, parameter estimation, level set methods, process modeling, semiconductor manufacturing

AMS subject classifications. 49N50, 35R35, 35R30, 93E10, 93E24

PII. S0363012998336340

1. Introduction.

This work is motivated by the need for accurate low-order phenomenological models of thin film etching and deposition processes. These processes are central to the manufacture of microelectronic devices. Phenomenological models are necessary because of the complexity of the surface chemistry and the plasma-surface interactions. Typically these models lump together numerous unknown rate constants into relatively few parameters [8, 14, 20, 21]. Reliable use of these models for simulation or control then depends on the ability of the user to choose the values of these parameters correctly. The parameter values may be selected based on surface evolution data from scanning electron micrographs. To our knowledge, only one study has investigated methods to optimize this process [8]. That work, while successful, used a nongeometric cost function that requires the user to select points in one-to-one correspondence on the actual and estimated surfaces. This selection introduces an arbitrary component of unknown significance into the procedure and places an undesirable burden on the user. In the following sections we introduce a completely coordinate-free cost function that eliminates this arbitrary element.

In a recent series of papers Sethian and Adalsteinsson apply level set methods to the simulation of feature development in a variety of semiconductor manufacturing applications [1, 18, 19]. Level set methods provide a flexible framework for surface evolution problems. String methods are the main alternatives to level set methods. Here the surface is defined by points, which are advanced according to the surface velocity. Two problems may occur if corners appear in the surface. First, if the surface points are allowed to "bunch up" at a corner, the Courant–Friedrichs–Lewy condition may be violated. Second, points on the surface may move past each other, necessitating "delooping." The level set framework is inherently immune to these

*Received by the editors March 25, 1998; accepted for publication (in revised form) September 30, 1998; published electronically July 9, 1999. This research was partially supported by a grant from the Center for Automation and Robotics at Texas Tech University. A preliminary and highly abridged version of this paper was presented at the 1998 American Control Conference, Philadelphia, PA, and published in *Proc. American Control Conf.*, Vol. 6, 1998, pp. 3758–3762.

http://www.siam.org/journals/sicon/37-5/33634.html

†Department of Mechanical Engineering, Texas Tech University, Lubbock, TX 79409-1021 (jberg@ttu.edu).

‡Center of Mathematical Modeling, Department of Mathematics and Physics, Mälardalen University, P.O. Box 883, S-721 23 Västerås, Sweden (kenneth.holmstrom@mdh.se).

problems. String methods may also become unwieldy when topological transitions occur, as during merging or splitting of surfaces. Again, these transitions are handled automatically using level sets. For a detailed discussion and a survey of the literature see [18]. Both corner formation and topological transition occur in etching and deposition, making this a promising area for the application of level set methods.

Although we are motivated by plasma etching and deposition applications, as described above, the problem we address here is more narrow in scope. We describe a geometric cost function based on level set descriptions of both the curve to be matched and the parametrized family. We then construct derivatives of the cost function in terms of the parameters. This allows us to apply the gradient descent class of minimization methods. A simple example is presented, and the tasks remaining before the results can be applied to the full problem are discussed. Several important properties of the cost function are discussed.

Level set methods have been used in identification. Approaches to real-time estimates of evolving features in plasma etching based on level sets have been put forth in [2, 3, 4, 5]. Santosa devises a level set approach to reconstruct the shape of an unknown object from a discrete set of measurements [16]. This is somewhat different from our situation. In this paper we treat the case where the shape of the object is itself the measurement.

2. Level sets. The approach presented here is general, but in this paper we restrict our attention to curves in the plane. In the level set formulation, an oriented curve $\mathcal{C}$ is represented by the zero level set (ZLS) of a level set function (LSF) $\Phi(x)$, that is, $\mathcal{C} = \{x \in R^2 : \Phi(x) = 0\}$. Clearly the choice of Φ is not unique. To remove this nonuniqueness one may think of $\mathcal{C}$ as defining an equivalence class of LSFs on the plane, where two such functions Φ and Ψ are defined to be equivalent if they have the same signature, that is, if Φ and Ψ have the same sign (or are simultaneously zero) at every point in R^2. As usual when dealing with equivalence classes, it is convenient to choose a canonical element. A good choice here is the *signed distance function.* For any curve in the plane, and given some choice of norm, the magnitude of the signed distance function at a point is the shortest distance (as defined by the norm) to the curve. The signed distance function is negative if the point is inside the curve and positive if the point is outside the curve. The choice of which component is inside and which is outside is essentially arbitrary. In situations with physical meaning the proper choice is generally obvious. Efficient algorithms exist for generating the signed distance function given an arbitrary LSF, particularly when allowed freedom in the choice of norms.

The real importance of the level set approach comes when considering *evolving* curves. Here $\mathcal{C} : [0, 1] \times [0, t_f] \to R^2$ is a parametrized curve evolving in time according to the equation

$$\frac{\partial \mathcal{C}}{\partial t} = \tilde{\beta}(s, t, \ldots)\nu, \tag{1}$$

where ν is the outward pointing unit normal to $\mathcal{C}$. The *speed function* $\tilde{\beta}$ describes the outward normal velocity of $\mathcal{C}$ and may depend on independent variables, on local properties of the curve, or on global considerations [13]. In (1), $\tilde{\beta}$ is defined only on the curve. Now the LSF too is time-dependent. The evolution of the LSF according to (1) is governed by the following PDE:

$$\Phi_t + \beta(x, t, \ldots)\|\nabla\Phi\| = 0. \tag{2}$$

This PDE is derived as follows: The curve $\mathcal{C}(s,t)$ is represented by the ZLS of a function $\Phi : R^2 \times [0,\tau) \to R$. Assume that Φ is negative in the interior and positive in the exterior of the zero level set. We consider the zero level set, defined by

$$\left\{X(t) \in R^2 \,:\, \Phi(X,t) = 0\right\}. \tag{3}$$

We have to find an evolution equation of Φ, such that the evolving curve $\mathcal{C}(t)$ is given by the evolving zero level $X(t)$, i.e., $\mathcal{C}(t) \equiv X(t)$. By differentiating $\Phi(X(t),t) = 0$ we obtain

$$\nabla\Phi(X,t) \cdot X_t + \Phi_t(X,t) = 0. \tag{4}$$

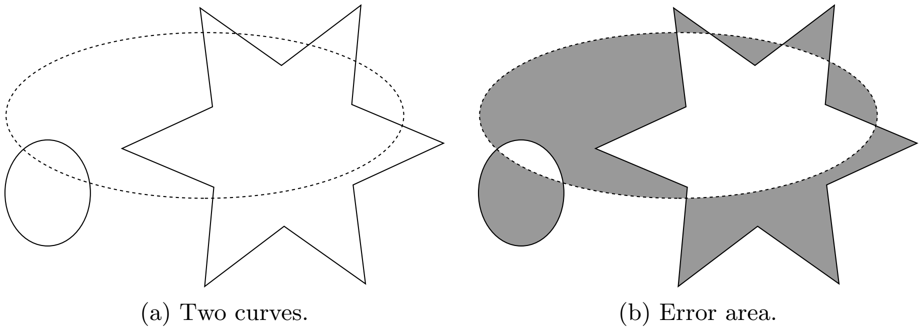
For any level set, the following relation holds:

$$\frac{\nabla\Phi}{\parallel \nabla\Phi \parallel} = \nu. \tag{5}$$

Substituting (5) into (1) to eliminate ν, then placing the resulting expression for $\mathcal{C}_t$ into (4), in place of X_t, gives (2). For more detail on the meaning of (2) when the LSF fails to be differentiable everywhere in space, and for numerical approaches to solving such cases, see [18] and the references therein.

All LSFs in the equivalence class must satisfy an equation like (2). Here β is a function defined everywhere in the plane. Different choices for β are possible, but at any instant they must all coincide on the ZLS itself, and there be equal to $\tilde{\beta}$. We will call any β satisfying this condition *admissible.* To remove the nonuniqueness in the definition of (2) we again turn to equivalence. Any admissible β will map a member of the equivalence class of LSFs at time t_0 into the proper equivalence class at time t. The canonical signed distance function can be recovered from any other member of the class as desired. This process is often referred to as *renormalization.* Note that the evolution (2) will not, in general, preserve the signed distance function.

3. A metric for level set functions. The process of parameter identification will require that we find the parameter values that give, in some sense, the closest match to an observed evolution. To make this rigorous, we must define a metric for LSFs that formalizes the idea of "distance." We turn now to the construction of a suitable function. Our main objective is to avoid the need to parametrize the curves. Such parametrizations are intrinsically arbitrary, and place a burden on the experience and expertise of the end user. This is the motivation for defining a geometric (by which we mean *coordinate-free*) cost function. The idea is presented in Figure 3.1. Figure 3.1(a) shows two closed curves, one consisting of a single connected component, the other consisting of two connected components. In all cases "inside" is taken to be the bounded component of the plane defined by the curves. Since the two curves are not identical, there should be a positive distance between them. We define this distance to be the area of the region between the two curves. This area is shaded light gray in Figure 3.1(b). The example shown in Figure 3.1 is abstract. A geometry that might occur in thin film deposition into a trench or via is shown in Figure 3.2. Here the points contained inside both curves are shaded dark, the points outside both curves are unshaded, and the points inside one curve but not the other—that set of points whose measure defines the distance—are shaded a light gray. That is, given a pair of simple closed curves, $\mathcal{C}_1$ and $\mathcal{C}_2$, let the distance from $\mathcal{C}_1$ to $\mathcal{C}_2$, denoted $\rho(\mathcal{C}_1, \mathcal{C}_2)$, be the total area of points enclosed by either one curve or the other, but



(a) Two curves. (b) Error area.

Fig. 3.1. *Two curves. One* (*dashed*) *is a large ellipse. The other* (*solid*) *consists of two simply connected components, a star and a small ellipse. The distance between the curves is found by summing the areas of the shaded regions.*

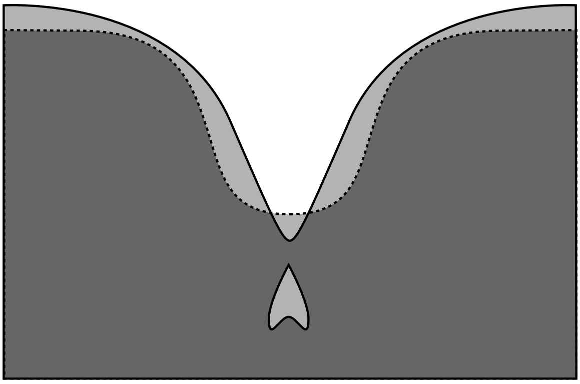


Fig. 3.2. *The estimate* (*dashed*) *differs from the measurement* (*solid*). *The associated cost is then the area between the two curves* (*light gray*).

not both. In three dimensions or higher area is replaced by volume, etc. The definition is extended to any curve defined by a LSF on a compact domain by including appropriate portions of the boundary.

Formally, let A and B be sets. Define the function $\rho(A, B)$ on sets to be the Lebesgue outer measure of the symmetric difference: $\rho(A, B) = \mu^*(S(A, B))$, where $S(A, B) := (A \setminus B) \cup (B \setminus A)$. The symbol $\setminus$ denotes the usual set difference. Unfortunately this is not a metric on sets, because sets that differ only by a set of measure zero will have a distance zero from each other, but are not equal. This problem can be resolved by defining two sets to be equivalent if ρ is zero. Then ρ is a metric on the resulting quotient space [15]. This operation is justified physically in our application since closed curves containing no area can be neglected, at least in terms of the gross structure. Whether these curves have some physical significance at a smaller length scale, or in terms of electrical or mechanical properties of the material, is an interesting question, but beyond the scope of this paper. Having defined $\rho(A, B)$ to be a metric on sets in the above sense, we extend it to closed curves as follows: If I_1 is the interior of curve $\mathcal{C}_1$, and I_2 is the interior of curve $\mathcal{C}_2$, then $\rho(\mathcal{C}_1, \mathcal{C}_2) = \rho(I_1, I_2)$. Finally we overload the notation still further, and extend the metric to LSFs, as fol-

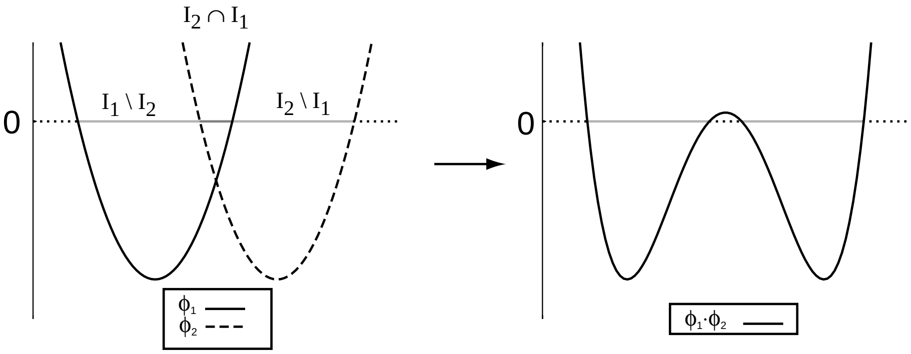


FIG. 3.3. *Product method for constructing error area from level sets.*

lows: If $\mathcal{C}_1$ is the ZLS of Φ and $\mathcal{C}_2$ is the ZLS of Ψ, then $\rho(\Psi, \Phi) = \rho(\mathcal{C}_1, \mathcal{C}_2)$. For the last extension, it is again necessary to consider a quotient space of the set of all LSFs. Here the equivalence relation defined earlier is weakened slightly to $\Phi(x) \sim \Psi(x)$ if $\text{sign}(\Phi) = \text{sign}(\Psi)$ *almost everywhere.* In the context of this paper these are technical details, and do not affect the computations. We refer to the metrics thus defined on plane curves and LSFs as the *area metric*, and refer to its value as the *error area.*

The *Hausdorff metric* is often used to compare curves in the plane. The "asymmetric Hausdorff metric" between curves $\mathcal{C}_1$ and $\mathcal{C}_2$ is defined, given some norm, by placing spheres of radius ϵ on each point of $\mathcal{C}_1$. The infimal value of ϵ for which the union of these spheres contains every point of $\mathcal{C}_2$ is the asymmetric Hausdorff metric from $\mathcal{C}_1$ to $\mathcal{C}_2$. The (symmetric) Hausdorff metric is the larger of the two asymmetric Hausdorff metrics. This paper is not intended as a comparison of the two metrics—only the area metric will be considered below. Much of the development can be accomplished using the Hausdorff metric in place of the area metric. In the context of the etching and deposition problems that motivate this work, the Hausdorff metric has advantages and disadvantages. For example, consider a case where the estimated curve agrees with the measured curve, except for the presence of a tiny bubble far from the free surface. The error according to the Hausdorff metric will be large, even though the bubble would most likely have little physical significance. The area metric would give a small error in this case. On the other hand, given two voids, the area metric does not distinguish between the situation where they are almost touching, and the situation where they are far apart. This potentially troublesome behavior does not occur with the Hausdorff metric.

3.1. The product method. When the estimate and the measurement curve are both characterized by LSFs, the area metric is easy to calculate. If the two LSFs are multiplied pointwise, the result is a new LSF, which we call the *product LSF.* The ZLS of the product LSF defines a curve, and the area of the interior of this curve is exactly the desired value. To see this, let $\mathcal{C}_1$ be the ZLS of Φ and $\mathcal{C}_2$ be the ZLS of Ψ. Recall that the distance between $\mathcal{C}_1$ and $\mathcal{C}_2$ is the measure of the symmetric difference between their interiors. That is, in terms of the LSFs, the measure of the set of points for which one LSF takes a positive value, while the other takes a negative value. If we form $\Gamma(x) = \Phi\Psi$, we see that $\rho(\mathcal{C}_1, \mathcal{C}_2)$ is just the area of the interior of the ZLS corresponding to Γ. Figure 3.3 is a graphical depiction of this construction in one dimension. To calculate the area we generate $\mathcal{C}_3$—the contour corresponding to the ZLS of $\Gamma(x)$ (note that the contour is now composed of *oriented curves* and is no longer just a set)—and apply Green's theorem:

$$\rho = \frac{1}{2}\int_{\mathcal{C}_3} \langle \mathcal{C}_3, \nu \rangle \, ds, \tag{6}$$

where $\langle \cdot, \cdot \rangle$ is the vector inner product. Here, as elsewhere in this paper, s stands for the arc length parameter [9]. Note that it is necessary to assign an LSF to the measurement.

In the procedure above, generating the contour corresponding to the ZLS of Γ is a key step. A geometric algorithm to accomplish this has been developed by Siddiqi, Kimia, and Wang, and is presented in [17].

3.2. The symmetric difference method. In the preceding section the product LSF was used to define the error area. In this section we present an alternative that has better numerical properties. Consider the interior of the ZLS as a set, and note the following relationships between operations on LSFs and set operations on the interiors.

1. Given a LSF Φ, and denoting the interior of the ZLS of Φ as $\mathrm{int}(\Phi)$, then $\mathrm{int}(-\Phi) = \mathrm{int}(\Phi)^c$. That is, the interior of the negative of the ZLS is the *complement* of the interior of the ZLS.
2. Given two LSFs, Φ and Ψ, then $\mathrm{int}(\max(\Phi,\Psi)) = \mathrm{int}(\Phi) \cap \mathrm{int}(\Psi)$.
3. Given two LSFs, Φ and Ψ, then $\mathrm{int}(\min(\Phi,\Psi)) = \mathrm{int}(\Phi) \cup \mathrm{int}(\Psi)$.

The symmetric difference between sets A and B is defined as $S(A,B) = (A \cap B^c) \cup (B \cap A^c)$. It can be shown that $S(A,B) = (A \cup B) \cap (A \cap B)^c$. Identifying A with $\mathrm{int}(\Phi)$ and B with $\mathrm{int}(\Psi)$, we define

$$\Theta \; = \; \max(\underline{\theta}, -\overline{\theta}), \tag{7}$$

where,

$$\underline{\theta} \; = \; \min(\Phi, \Psi), \tag{8}$$

$$\overline{\theta} \; = \; \max(\Phi, \Psi). \tag{9}$$

And it is seen that the error area is the Lebesgue measure of the interior of Θ. We refer to Θ as the symmetric difference LSF.

This formulation is numerically preferable to that of the previous section because if the gradients of the LSFs are near unity everywhere, then the gradient of the symmetric difference LSF will be also. In contrast, the gradients of the product LSF are not preserved. However, even for the symmetric difference method, away from the ZLS the signed distance function is not preserved. To see this, consider the case of the two LSFs, $\Phi(x,y) = x$ and $\Psi(x,y) = y$, whose interiors are the left half-plane and lower half-plane, respectively. These LSFs correspond to the signed distance function in the 1-norm or 2-norm. The union of their interiors is the interior of the level set function $\Theta = \min(\Phi,\Psi)$. The gradient of Θ has unity magnitude everywhere on the ZLS, except at the corner at the origin. To see that the signed distance function is not preserved, consider the point $(-1,-1)$. Although its distance from the ZLS is 2 (in the 1-norm) or $\sqrt{2}$ (in the 2-norm), the value of Θ at $(-1,-1)$ is -1.

4. Parametrized level sets. Ultimately we wish to parametrize the speed function β. In this paper however, we consider the simpler case of a parametrized level set. For the present we suppress also the time dependence of the curves. Our objective then is as follows: We are given a *measured* curve, $\mathcal{M}$, which is the ZLS of $\Phi(x)$. We also have a parametrized family of level set functions, $\Psi(x;\lambda)$. Let $\mathcal{L}(\lambda)$ be the ZLS of $\Psi(x;\lambda)$. We wish to find the value of the parameter vector λ such that $\mathcal{L}(\lambda)$ is closest to $\mathcal{M}$.

One way to proceed is to treat the metric as a cost function and seek to minimize

$$J(\lambda) = \rho(\mathcal{M}, \mathcal{L}(\lambda)) \tag{10}$$

or, in terms of Green's theorem, to minimize

$$J(\lambda) = \frac{1}{2}\int_{\mathcal{E}} \langle \mathcal{E}, \nu \rangle \, ds, \tag{11}$$

where $\mathcal{E}$ is the ZLS of

$$\Gamma(x;\lambda) := \Phi(x)\Psi(x;\lambda). \tag{12}$$

We wish to apply gradient descent methods to accomplish this minimization. To do this, we need to calculate the gradient ρ_i, where we denote partial differentiation with respect to λ_i by a subscript i.

For the smooth segments of the curve we have

$$\rho_i(\lambda) = \int_{\mathcal{E}} \langle \mathcal{E}_i, \nu \rangle \, ds. \tag{13}$$

Equation (13) is derived as follows: Let θ parametrize the curve and take values from 0 to 2π regardless of the value of λ. Of course, this parameter is, in general, no longer the arc length. Now, writing $\mathcal{E}$ explicitly as (x, y), the tangent vector explicitly as $(x_\theta, y_\theta)/\sqrt{x_\theta^2 + y_\theta^2}$, the outward normal ν as $(-y_\theta, x_\theta)/\sqrt{x_\theta^2 + y_\theta^2}$, and recalling that $ds = \sqrt{x_\theta^2 + y_\theta^2}\, d\theta$,

$$\rho = \frac{1}{2}\int_0^{L(\lambda)} \langle \mathcal{E}, \nu \rangle \, ds \tag{14}$$

$$= \frac{1}{2}\int_0^{2\pi} \langle \mathcal{E}, (-y_\theta, x_\theta) \rangle \, d\theta. \tag{15}$$

Differentiating inside the integral, and using the fact that the limits of integration no longer depend on λ,

$$\rho_\lambda = \frac{1}{2}\int_{\mathcal{E}} \langle \mathcal{E}_\lambda, \nu \rangle \, ds + \frac{1}{2}\int_0^{2\pi} \langle \mathcal{E}, (-y_{\theta\lambda}, x_{\theta\lambda}) \rangle \, d\theta. \tag{16}$$

Differentiation by λ and by θ commute. Integration of the second term by parts gives

$$\frac{1}{2}\int_0^{2\pi} \langle \mathcal{E}, (-y_{\theta\lambda}, x_{\theta\lambda}) \rangle \, d\theta = \langle \mathcal{E}, (-y_\lambda, x_\lambda) \rangle |_0^{2\pi} - \frac{1}{2}\int_0^{2\pi} \langle \mathcal{E}_\theta, (-y_\lambda, x_\lambda) \rangle \, d\theta. \tag{17}$$

The leading term in this expression is zero because the curve is closed. To evaluate the remaining term, write the vectors explicitly.

$$-\frac{1}{2}\int_0^{2\pi} \langle \mathcal{E}_\theta, (-y_\lambda, x_\lambda) \rangle \, d\theta \;=\; -\frac{1}{2}\int_0^{2\pi} \langle (x_\theta, y_\theta), (-y_\lambda, x_\lambda) \rangle \, d\theta \tag{18}$$

$$= -\frac{1}{2}\int_0^{2\pi} (-x_\theta y_\lambda + x_\lambda y_\theta)\, d\theta \tag{19}$$

$$= \frac{1}{2}\int_0^{2\pi} \langle (x_\lambda, y_\lambda), (-y_\theta, x_\theta)\rangle\, d\theta \tag{20}$$

$$= \frac{1}{2}\int_{\mathcal{E}} \langle \mathcal{E}_\lambda, \nu\rangle\, ds. \tag{21}$$

Substituting this result into (16) gives (13).

Now consider the change in $\mathcal{E}$ corresponding to a change in the ith parameter λ_i. We write $\Gamma(X(s;\lambda);\lambda) = 0$. So, $\Gamma_i + \langle \nabla\Gamma, \mathcal{E}_i\rangle = 0$. In fact, differentiation with respect to a parameter gives the same form we found when we derived the evolution equation (2), namely $\Gamma_i + \langle \mathcal{E}_i, \nu\rangle \|\nabla\Gamma\| = 0$. This time, however, we arrange the terms as follows:

$$\langle \mathcal{E}_i, \nu\rangle = -\frac{\Gamma_i}{\|\nabla\Gamma\|}. \tag{22}$$

Now, substituting (22) into (13), we obtain

$$\rho_i = -\int_{\mathcal{E}} \Gamma_i/\|\nabla\Gamma\|\, ds. \tag{23}$$

The derivatives of Γ are replaced by $\Gamma_i = \Phi\Psi_i$, and $\nabla\Gamma = \Phi\nabla\Psi + \Psi\nabla\Phi$. Since $\Phi = 0$ on $\mathcal{M}$ and $\Psi = 0$ on $\mathcal{L}$,

$$\rho_i = -\int_{\mathcal{M}\cup\mathcal{L}} \Phi\Psi_i/\|(\Phi\nabla\Psi + \Psi\nabla\Phi)\|\, ds \tag{24}$$

$$= -\int_{\mathcal{L}} \Phi\Psi_i/\|\Phi\nabla\Psi\|\, ds \tag{25}$$

$$= -\int_{\mathcal{L}} \operatorname{sign}(\Phi)\Psi_i/\|\nabla\Psi\|\, ds. \tag{26}$$

5. Corner effects. The formulas (13) and (26) of the previous section apply only to the smooth portion of the curves, and the contours will fail to be smooth at a finite number of corners. If the original curves are smooth, which we assume here for convenience, then the error contours defined by the ZLS of the product LSF will have corners only where the original curves intersect. It is now shown that, to first order, the contribution at each of these corners is zero.

Let $\mathcal{M}$ be the measured curve, $\mathcal{L}$ be the original estimated curve, corresponding to some nominal value for the parameter vector, and $\mathcal{L}'$ be the estimated curve at some new parameter vector, obtained by varying only the element λ_i by $\Delta\lambda_i$. The change in the curve along the outward normal ν, denoted δ_i, will be

$$\delta_i = \langle (\partial\mathcal{L}/\partial\lambda_i)\Delta\lambda_i, \nu\rangle \tag{27}$$

$$= \langle \mathcal{L}_i, \nu\rangle \Delta\lambda_i. \tag{28}$$

We assume that $\mathcal{L}_i$ is continuous. Now consider some neighborhood of each corner point, small enough that $\langle \mathcal{L}_i, \nu\rangle$ may be treated as approximately constant inside that neighborhood and constructed so as to intersect every curve in the family $\mathcal{L} + \langle \mathcal{L}_i, \nu\rangle t,\ t \in [0, \Delta\lambda_i]$, at a right angle. We also choose the neighborhood sufficiently small that the segments of $\mathcal{M}$ and $\mathcal{L}$ contained within it may be approximated

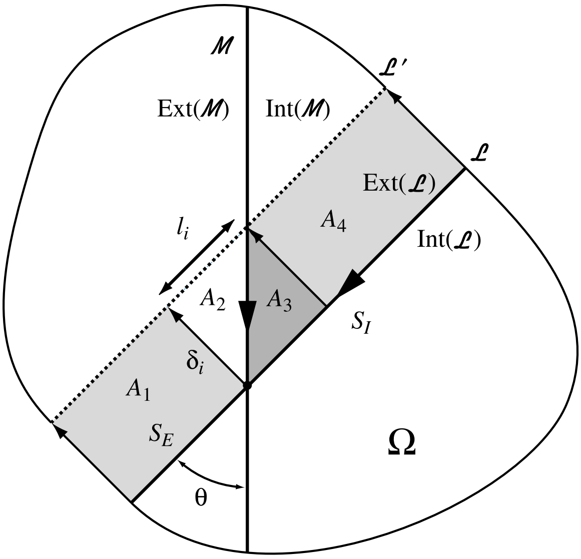


FIG. 5.1. *Close-up of corner effects.*

arbitrarily closely by straight lines. Figure 5.1 shows one such neighborhood, which we denote as Ω. Outside of these neighborhoods (13) holds; we now calculate the difference in one such neighborhood between ρ_i as predicted by (13) and the actual ρ_i. The sum of all such terms will give the total value required to correct the smooth approximation to ρ_i. Denote the length of the curve $\mathcal{L}$ that lies in both Ω and the interior of $\mathcal{M}$ by S_I, and the length of $\mathcal{L}$ in Ω but outside of $\mathcal{M}$ by S_E. The neighborhood is chosen small enough that $\langle \mathcal{L}_i, \nu \rangle$ is approximated arbitrarily well by a constant value. Hence, inside Ω, (13) becomes

$$\rho_i^s = (S_E - S_I)\langle \mathcal{L}_i, \nu \rangle, \tag{29}$$

where we use the fact that the outward normal to $\mathcal{L}$ coincides with the outward normal to the ZLS of the product LSF on the exterior of $\mathcal{M}$ and is opposite to it on the interior of $\mathcal{M}$. The superscript s indicates that this is the smooth approximation. Then

$$\Delta\rho^s = \rho_i^s \Delta\lambda_i = S_E\delta_i - S_I\delta_i. \tag{30}$$

Or, in terms of the regions shown in Figure 5.1,

$$\Delta\rho^s = A_1 - (A_2 + A_3 + A_4). \tag{31}$$

However, in terms of the regions shown in Figure 5.1, the actual change in the area metric is

$$\Delta\rho^a = A_1 + A_2 - (A_3 + A_4). \tag{32}$$

So the corrected expression for $\Delta\rho^a$ is

$$\Delta\rho^a = \Delta\rho^s + 2A_2 = \Delta\rho^s + \delta_i l_i. \tag{33}$$

Dividing through by $\Delta\lambda_i$ gives

$$\frac{\Delta\rho^a}{\Delta\lambda_i} = \frac{\Delta\rho^s}{\Delta\lambda_i} + \langle \mathcal{L}_i, \nu\rangle l_i. \tag{34}$$

As $\Delta\lambda_j$ goes to zero this becomes

$$\frac{\partial\rho^a}{\partial\lambda_i} = \frac{\partial\rho^s}{\partial\lambda_i} + \langle \mathcal{L}_i, \nu\rangle l_i. \tag{35}$$

So $\langle \mathcal{L}_i, \nu\rangle l_i$ is the correction term that must be applied to (13) or (26) to account for the corner. Writing l_i in terms of the corner angle, θ, we see that $l_i = \delta_i / \tan\theta = \langle \mathcal{L}_i, \nu\rangle \Delta\lambda_i / \tan\theta$. Thus, as $\Delta\lambda_i$ goes to zero, the correction term $\langle \mathcal{L}_i, \nu\rangle^2 \Delta\lambda_i / \tan\theta$ vanishes, and so the corners do not affect the calculation of the first derivative. On the other hand, it is clear that for small angles of intersection, the region for which a first-order approximation to the area metric is accurate will be very small. Therefore we anticipate that such geometries will cause significant numerical difficulties in minimization.

For the case in which the curve itself has corners, the gradients calculated by (26) will require correction. We will not consider such curves in the present paper. For a treatment of corners in a related context see [3].

5.1. Gradients by the symmetric difference method. To derive the gradient using the symmetric difference method we replace the product LSF Γ in (12) with the symmetric difference LSF Θ from (7). This substitution does not affect the calculation until (23). There we find

$$\Theta_i = \begin{cases} \Psi_i, & \Psi = 0, \Phi > 0, \\ -\Psi_i, & \Psi = 0, \Phi < 0, \\ 0 & \text{otherwise}, \end{cases} \tag{36}$$

where the case that both Φ and Ψ are zero along a measurable portion of the curve is neglected. This corresponds to a degenerate LSF, and the numerical scheme will typically not detect such a segment, since no sign change occurs. To see that (36) is correct, note that almost everywhere on the ZLS of the symmetric difference LSF either Φ or Ψ, but not both, is zero. When Φ is zero then the ZLS is locally determined only by Φ. Since Φ has no dependence on parameters, Φ_i, and so Θ_i, is zero. When $\Psi = 0$ the ZLS is locally determined only by Ψ. Then

$$\underline{\theta} = \min(\Phi, \Psi) = \begin{cases} \Phi, & \Phi < 0, \\ \Psi, & \Phi > 0. \end{cases} \tag{37}$$

$$\overline{\theta} = \max(\Phi, \Psi) = \begin{cases} \Psi, & \Phi < 0, \\ \Phi, & \Phi > 0. \end{cases} \tag{38}$$

And so,

$$\Theta = \max(\underline{\theta}, -\overline{\theta}) = \begin{cases} \max(\Phi, -\Psi), & \Phi < 0, \\ \max(\Psi, -\Phi), & \Phi > 0. \end{cases} \tag{39}$$

$$= \begin{cases} -\Psi, & \Phi < 0, \\ \Psi, & \Phi > 0. \end{cases} \tag{40}$$

Taking the gradient gives

$$\nabla\Theta = \begin{cases} -\nabla\Psi, & \Phi < 0, \\ \nabla\Psi, & \Phi > 0. \end{cases} \tag{41}$$

So the final expression for $\partial\rho/\partial\lambda_i$ becomes

$$\rho_i = -\int_{\mathcal{L}} \operatorname{sign}(\Phi)\Psi_i/\|\nabla\Psi\|\, ds. \tag{42}$$

That is, the expression for the gradient is the same for both the product LSF and the symmetric difference LSF. Since the area metric and its gradient are the same in both methods, it is natural to ask why one differs from the other. The answer is that the numerical implementation of the contour tracing algorithm will give different results. In the case where both LSFs are signed-distance functions, the symmetric difference method results in a LSF whose intersection with any grid line is piecewise linear. Therefore the interpolation routines used to locate the zeros will give better results for this case. However the gradient calculation is an integral over the ZLS of the estimate only. Therefore the method chosen does not affect the computation of the gradient.

6. Examples.

Here we consider two simple examples. In the first, the measured curve $\mathcal{M}$ is a circle of radius R, centered on (x_0, y_0). Choosing the canonical LSF to represent the curve, we write $\Phi(x,y) = \sqrt{(x-x_0)^2 + (y-y_0)^2} - R$. As a parametrized LSF we also choose a circle, parametrized by the position of its center, and its radius:

$$\Psi(x,y;x_c,y_c,\Pi) = \sqrt{(x-x_c)^2 + (y-y_c)^2} - \Pi. \tag{43}$$

Note that $\|\nabla\Psi\| = \|\nabla\Phi\| = 1$. The derivatives of the estimated LSF with respect to the parameters are calculated exactly, as follows:

$$\Psi_{x_c} = -\frac{(x-x_c)}{\sqrt{(x-x_c)^2 + (y-y_c)^2}}, \tag{44}$$

$$\Psi_{y_c} = -\frac{(y-y_c)}{\sqrt{(x-x_c)^2 + (y-y_c)^2}}, \tag{45}$$

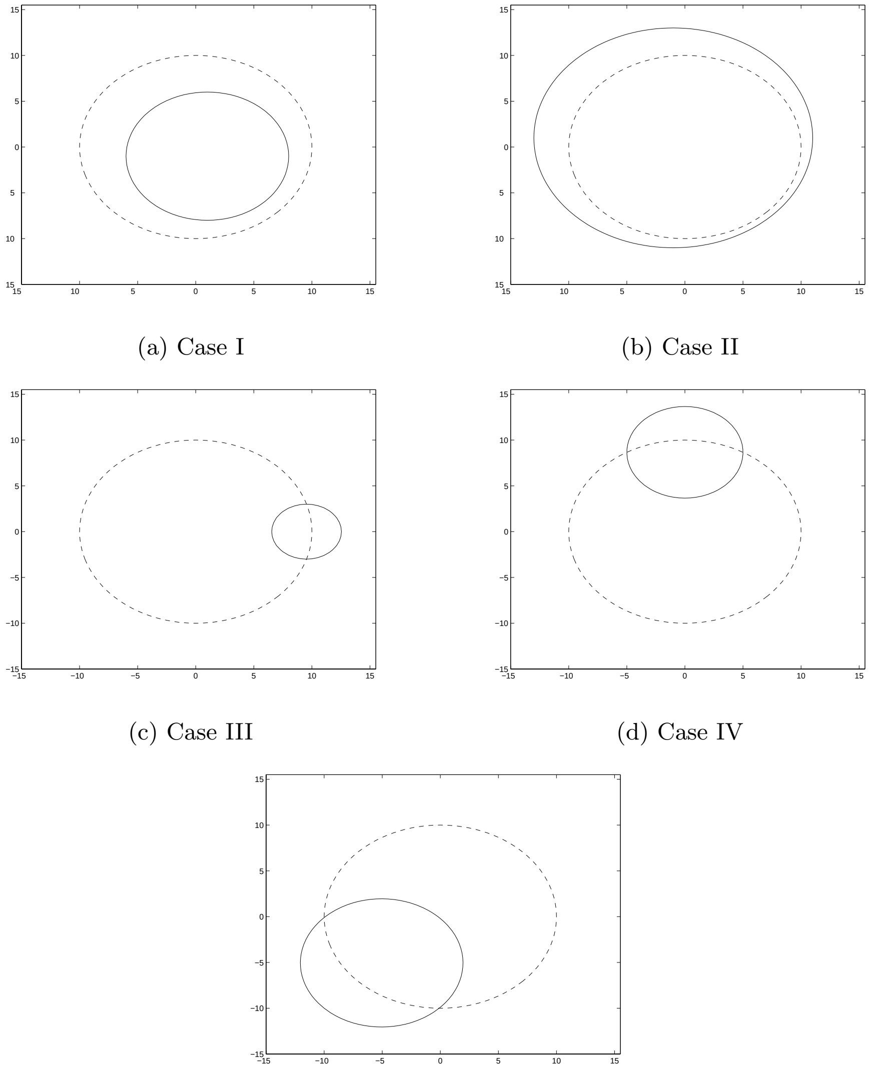
$$\Psi_{\Pi} = -1. \tag{46}$$

In the optimal solution the curves match exactly, and the cost function is zero.

For the second example the parametrized level set is unchanged, but the measured curve is taken as an ellipse, that is, the ZLS of $\Phi(x,y) = \sqrt{((x-x_0)/a)^2 + ((y-y_0)/b)^2} - 1$. This leaves the calculations unchanged, but the optimal estimate no longer gives a perfect match.

6.1. Area metric and derivative computation.

The computation of the area metric and its gradient to parameter variation was checked using five static circular geometries; see Figure 6.1. In the first, the estimated curve is a circle completely contained within the measured circle. In this case the area metric is the difference in the areas, the gradient terms ρ_{x_c} and ρ_{y_c} are zero, and ρ_{Π} is the negative of the circumference. In the second case, the positions are reversed, with the measurement contained in the estimate. The only difference between this and Case I is the sign of



(a) Case I

(b) Case II

(c) Case III

(d) Case IV

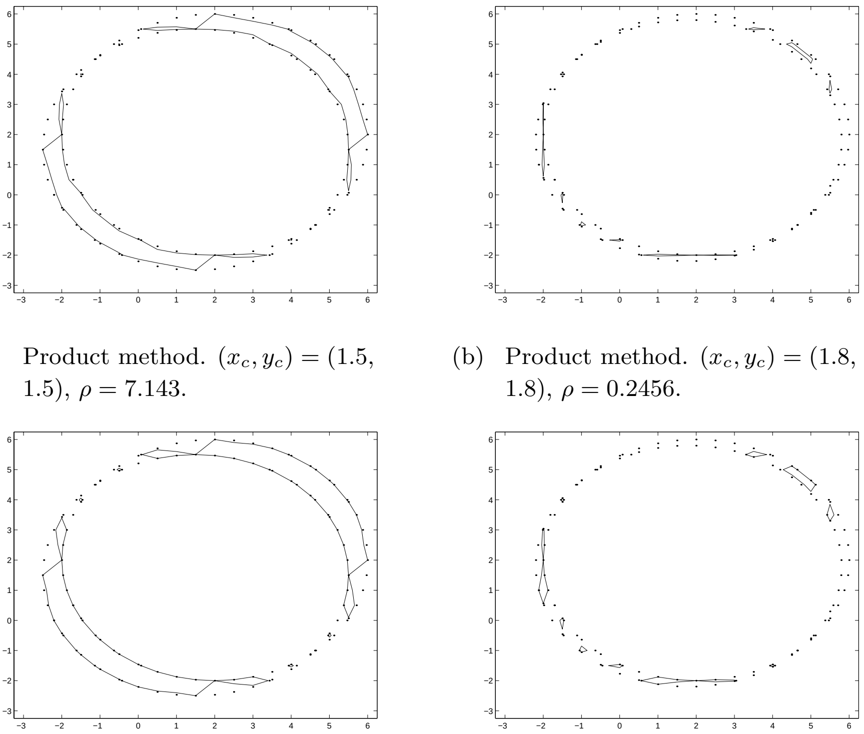
(e) Case V

FIG. 6.1. *Cases for testing area metric and gradient computations.*

the gradient term corresponding to the radius. In Cases III–V the estimated circle lies on the measured circle in such a way that half the boundary of the estimate lies inside the measured circle and half outside. This should result in $\rho_\Pi = 0$. In Case III the center of the estimated circle is displaced from the center of the measurement only in the x direction, in Case IV, only in the y direction, and in Case V, in both x and y. The gridsize was 0.5 in both the x and y directions for all calculations. Table 6.1 compares the computed results to the theoretical values. Because the gradient calculations depend only on the ZLS of the estimated LSF, the gradients are the same for the two calculation methods. However, the symmetric difference method appears to be the more accurate way to calculate the cost function itself.

TABLE 6.1
Computed cost function and gradient vs. actual values ($h = 0.5$).

		ρ	ρ_{x_c}	ρ_{y_c}	ρ_Π
	Theory	160.221	0	0	–43.982
Case I	Product	159.083	0.000	0.000	–43.969
	Sym. Diff.	160.216	0.000	0.000	–43.969
	Theory	138.230	0	0	75.398
Case II	Product	135.274	0.000	0.000	75.390
	Sym. Diff.	138.183	0.000	0.000	75.390
	Theory	310.457	12	0	0
Case III	Product	310.119	11.991	0.000	0.160
	Sym. Diff.	310.314	11.991	0.000	0.160
	Theory	296.042	0	20	0
Case IV	Product	295.496	1.000	19.967	–0.359
	Sym. Diff.	295.869	1.000	19.967	–0.359
	Theory	259.060	–19.799	–19.799	0
Case V	Product	257.962	–19.002	–20.532	–1.091
	Sym. Diff.	258.725	–19.002	–20.532	–1.091



(a) Product method. $(x_c, y_c) = (1.5, 1.5)$, $\rho = 7.143$.

(b) Product method. $(x_c, y_c) = (1.8, 1.8)$, $\rho = 0.2456$.

(c) Symmetric difference method. $(x_c, y_c) = (1.5, 1.5)$, $\rho = 8.588$.

(d) Symmetric difference method. $(x_c, y_c) = (1.8, 1.8)$, $\rho = 0.7756$.

FIG. 6.2. *Contour tracing using* (a, b) *the product method and* (c, d) *the symmetric difference method. True and estimated curves are dotted. The true curve is a circle of radius* 4 *centered at* (2, 2). *The estimated curve is a circle of radius* 4. *The centers are as noted. The resulting cost functions are also given.* $h = 0.5$.

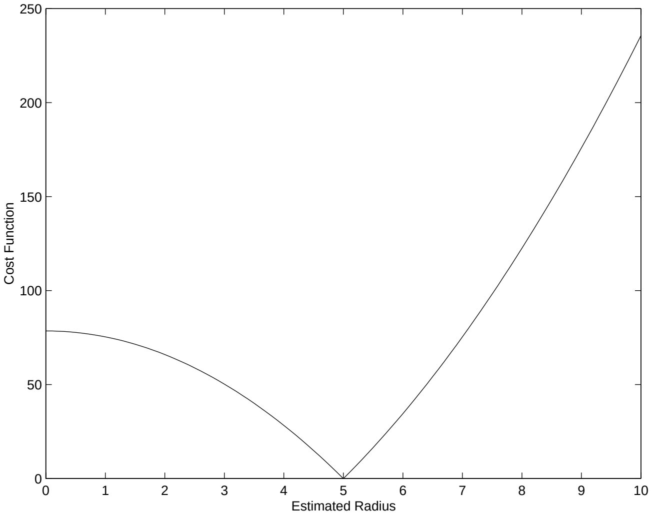


FIG. 6.3. *The cost function $\rho = \pi|R^2 - \Pi^2|$ for $R = 5$.*

6.2. Contour tracing. The error calculation becomes more difficult as the estimate approaches the true value. The reason is that the contour defining the error area becomes difficult to trace accurately. In this section, we examine this behavior more closely and compare the results from the product method with the results from the symmetric difference method. The test case is a circle centered at (2,2) with a radius of 4. Figures 6.2(a) and 6.2(b) show the true curve compared to a series of progressively closer estimates (both true and estimated curves are dotted) with the computed contours as calculated by the product method. Figures 6.2(c) and 6.2(d) show the same test using the symmetric difference method. Again, the gridsize in both the x and y directions is 0.5. Although both methods struggle with the contour as the curves become close, the symmetric difference method does a better job—as is evident from the plots and the computed values of ρ. The contour tracing algorithm implemented here is crude. Linear interpolation is used to locate zero crossings on the computational grid, and those points are connected to form contours. No shock placement logic is currently used. Improving the error calculation further requires either a smaller gridsize or a higher order interpolation method for contour tracing.

6.3. Numerical properties. The first example, in which both the measured and estimated curves are circles, gives several useful insights into the numerical properties of this problem. Consider the case in which the circles are concentric, and the only free parameter is the estimated radius, Π. Then the cost function is $\rho = \pi|R^2 - \Pi^2|$, where R is the true radius. Figure 6.3 shows this cost function. The function is nonsmooth and is not well approximated by a quadratic at the optimal point. The gradient is $\rho_\Pi = (2\pi)\,\mathrm{sign}(\Pi - R)\,\Pi$. The first derivative is undefined at the optimal point, and the magnitude of the gradient does not go to zero at the optimum.

Because of the presence of the absolute value in the area metric, we refer to direct minimization of this cost function as minimizing the 1-norm in the following discussion. The differentiability problem can be partially addressed by forming the cost function

$$J = \tfrac{1}{2}\rho(\mathcal{M}, \mathcal{L}(\lambda))^2. \tag{47}$$

We refer to this cost function as the 2-norm. The 2-norm and its derivatives, $J_i = \rho\rho_i$, are evaluated using the formulas derived above. Using this cost function, the case shown above becomes differentiable everywhere with zero gradient at the optimal point; in general we expect the 2-norm to be smoother than the 1-norm, but cannot guarantee differentiability. Both optimization strategies are illustrated in the example.

For either formulation, when the interiors of the two curves are disjoint, the estimate will converge to a local minimum. Namely, the estimated curve will simply decrease in radius until it vanishes (unless the parameter values are constrained). This observation holds true in general, and care must be taken to avoid either starting the minimization with disjoint interiors, or allowing such a situation during the minimization process.

In the case of a circular estimated curve, it is easy to prevent the interior of the initial guess from being disjoint with the interior of the measured curve. To do so it is sufficient to center the estimate at the origin, and then pick the initial radius sufficiently large. This may seem to be special to the example, but it is not. Consider the c-level set, $\{x \in R^2 : \Psi(x,t) = c\}$. If we use the signed distance function for Ψ, when c is sufficiently large, the measured curve must be contained in the interior of the estimate. So we replace $\Psi(x;\lambda)$ by $\tilde{\Psi}(x;\tilde{\lambda})$, where $\tilde{\Psi} = \Psi - c$, and $\tilde{\lambda} = (\lambda, c)$. Of course, this does not guarantee that subsequent iterations will not cause the interiors to become disjoint.

6.4. Numerical minimization results. The 2-norm is expected to be smoother than the 1-norm. Reflecting this difference, two different methods were chosen for the minimization. The 2-norm was minimized using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method [10]. This method makes a quadratic approximation to the cost function. The 1-norm, for which a quadratic approximation is expected to be poor near the optimum, was minimized using the method of steepest descent. In both cases the line search used was the one described by Fletcher [10, p. 34], with standard choice of parameters $\rho = 0.01$, $\tau_1 = 9$, and cubic interpolation. The line search accuracy parameter σ was set to 0.9 for the BFGS algorithm (inexact line search), and 0.1 for the steepest descent algorithm (exact line search). The specific implementation of these algorithms was the TOMLAB optimization environment package in Matlab. For more information see [11, 12].

To get some feel for the effects of discretization, the minimizations were carried out first with the true curve centered on a grid point (the origin) and concentric with the initial estimated circle, and next with the center of the true curve not on a grid point, and with the true curve not concentric with the initial estimated circle. In the cases where the true curve is circular the actual optimal estimate is obvious; in the cases where the true curve is elliptical it is necessary to calculate the optimal estimate. Assuming that the optimal estimate is concentric with the true ellipse, this calculation is quite easy. Under this assumption only one parameter is free—the radius of the estimated circle. Then $\langle \mathcal{E}_\Pi, \nu \rangle$ is unity everywhere on the curve, so (26) reduces to $\rho_\Pi = S_I - S_E$, where S_I and S_E are the total length of the arcs of the estimated circle inside and outside the ellipse, respectively. There are two such arcs inside the ellipse, which by symmetry are of equal length. Likewise there are two arcs of equal length outside the ellipse. Setting $\rho_\Pi = 0$ at the optimal point, we see that all four arcs

Table 6.2
Numerical minimization results.

Case	Method	$x_c{}^\star$	$y_c{}^\star$	$\Pi^\star$	$\rho^\star$	N_I	N_f
Centered Circle	True	0.0	0.0	8.0	0.0		
	1-norm	$-7.257 \cdot 10^{-16}$	$-8.844 \cdot 10^{-16}$	8.000	0.0	1	8
	2-norm	$-7.257 \cdot 10^{-16}$	$-8.844 \cdot 10^{-16}$	8.000	0.0	1	8
Skewed Circle	True	4.1	2.4	8.0	0.0		
	1-norm	3.959	2.465	8.065	0.6498	6	67
	2-norm	3.775	2.373	7.913	3.642	14	67
Centered Ellipse	True	0.0	0.0	7.071	80.44		
	1-norm	$-8.343 \cdot 10^{-16}$	$-1.017 \cdot 10^{-16}$	6.953	78.21	2	25
	2-norm	$-8.343 \cdot 10^{-16}$	$-1.017 \cdot 10^{-16}$	6.953	78.21	2	25
Skewed Ellipse	True	2.6	-4.1	7.071	80.44		
	1-norm	2.749	-4.108	7.033	78.15	10	105
	2-norm	2.448	-4.128	7.110	78.40	11	33

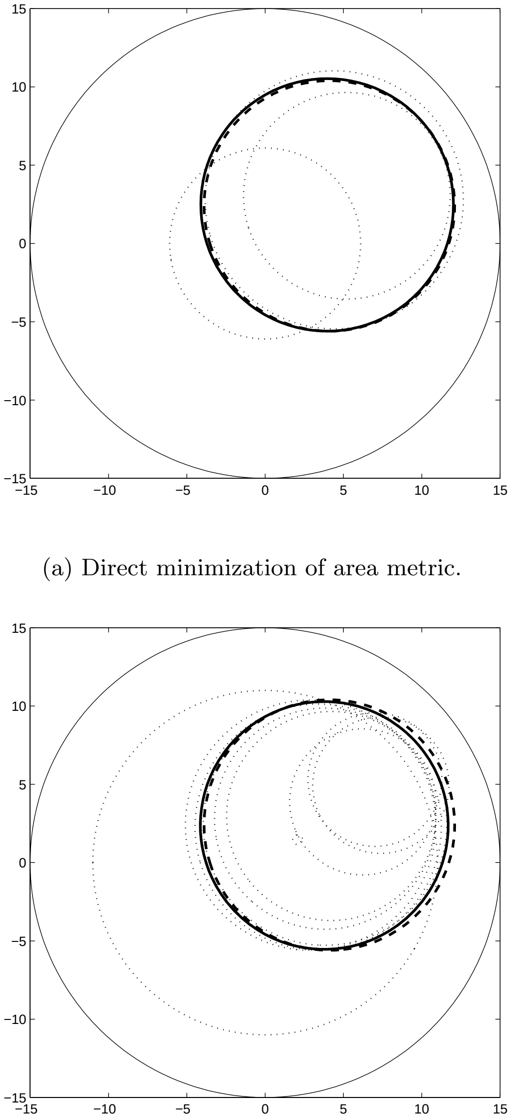
must have the same arc length. Therefore the angle defined by each arc must be $\pi/2$. Writing the equation for the ellipse in polar form $r = r(\theta)$, in a coordinate system with origin on the center point of the ellipse, we see that the optimal radius $\Pi^\star$ of the estimated circle is just given by $\Pi^\star = r(\pi/4)$. For the ellipse $(x/a)^2 + (y/b)^2 = 1$ used in the examples, where $a = 5\sqrt{5}$ and $b = a/2$, that gives $\Pi^\star = 5\sqrt{2} \approx 7.071$. Once this value has been obtained, the corresponding error area may be calculated by direct integration. In the cases considered below the optimal value is approximately 80.44.

All cases used the symmetric difference method to calculate the area metric, and the gridsize was 0.5 in the x and y directions. That is, the computational grid was 61 by 61 points square. The value of Π was constrained to be greater than the gridsize h. The first line search presented a problem, as the steps taken were much too large, and the first estimate often had an interior disjoint with the true curve. This situation leads to convergence to a nonglobal minimum. The problem was prevented by using the normalized gradient, rather than the gradient itself, for the first line search only.

Three tests for convergence were applied. Convergence condition 1 is an absolute change of less than 10^{-10} in the norm of the difference of sequential estimates of the parameter vector. Convergence condition 2 is a relative reduction in the cost function of less than 10^{-10} in ten consecutive iterations. Convergence condition 3 is that the calculated cost is less than or equal to a specified lower limit (zero in this case). Convergence condition 4, used only for the BFGS method, is that the norm of the gradient is less than 10^{-1}.

In all cases except the first the algorithms stopped because of convergence condition 1. However, this test does not always indicate small search steps. When the accuracy of the function and gradient computation is not sufficient or when we have an ill-conditioned problem the convergence may be due to a zero or close-to-zero step length. The directed derivative is negative but very small and the line search cannot reduce the objective function. In such cases the optimization algorithm terminates too early, and a separate analysis is needed to determine how far from the local optimum we are. We get these premature stops for our last three test examples, but as we know the true answer it is clear that the result is acceptable. One reason for premature termination may be the lost accuracy in the contour tracing algorithm. As has been previously stated, our implementation is crude and may be improved.

The results are summarized in Table 6.2. There the optimal parameter estimates are compared with the true values. The number of iterations required for conver-



(a) Direct minimization of area metric.

(b) Minimization of $J = \frac{1}{2}\rho^2$.

FIG. 6.4. *True curve is circle* (*dashed*). *Initial estimate* (*light solid*) *is circle of radius* 15 *centered on* $(0, 0)$. (a) *Steepest descent algorithm converges to final estimate* (*heavy solid*) *in* 6 *iterations* (*dotted*). (b) $J = \frac{1}{2}\rho^2$. *BFGS algorithm gives final estimate* (*heavy solid*) *in* 14 *iterations* (*dotted*).

gence, N_I, is listed, as are the total number of function evaluations, N_f (which for our test cases is also the total number of gradient evaluations), and the computed area metric of the final solution. Note that the computed values can be less than the true minimum because of errors associated with discretization and contour tracing.

In the case of the centered circle, both the steepest descent minimization of the

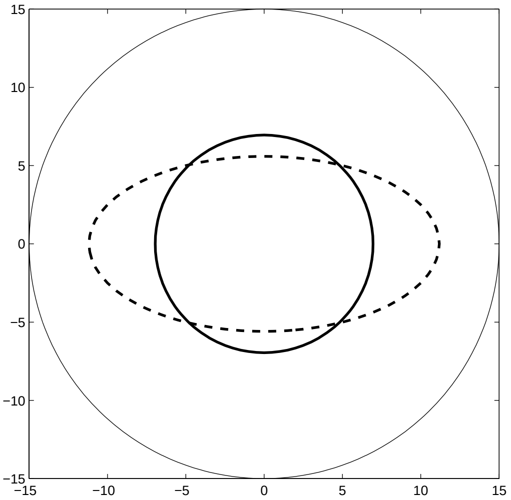
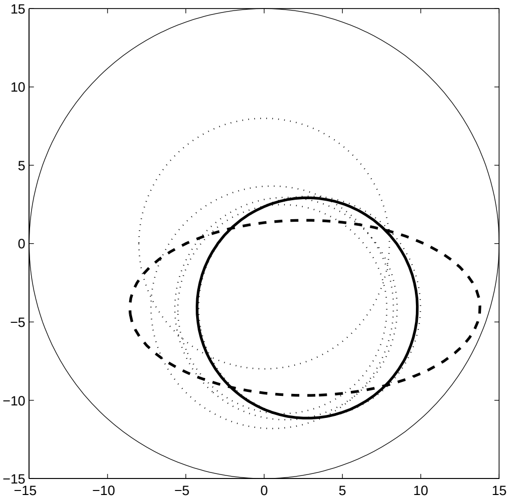


FIG. 6.5. *True curve is ellipse* (*dashed*)*. Initial estimate* (*light solid*) *is circle of radius* 15 *centered on* $(0, 0)$*. Both minimization schemes converge to final estimate* (*heavy solid*) *in* 2 *iterations. Results are identical for the two methods.*

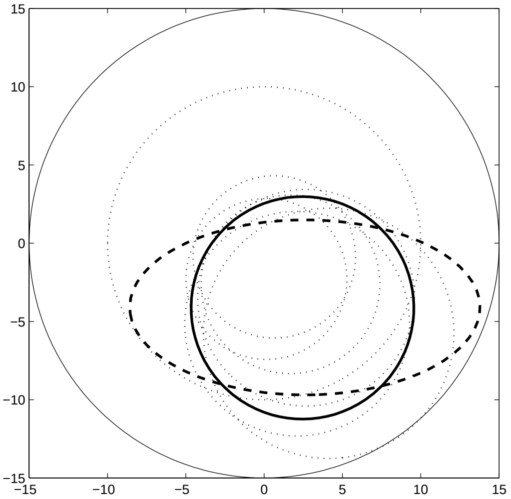
1-norm and the BFGS minimization of the 2-norm converge immediately to the optimal solution with very little error on the first line search. The results of the two methods are identical. This is expected, since the gradients are normalized on the first line search, and these methods are using the same search step in the first iteration. The true curves, starting curves, final estimate, and intermediate curves for the remaining cases are shown in Figures 6.4–6.6. The steepest descent minimization of the 1-norm does fairly well for the skewed circle. The relatively poor performance of the BFGS minimization of the 2-norm is partly due to the inaccurate line search for the first normalized gradient step. Away from the optimum the gradient of the 2-norm is very large. This leads to a very large second step that hits the lower constraint on Π, overshoots the minimum, and changes the center of the approximating circle too much. This suggests a two-phase algorithm, with one or more normalized gradient steps with accurate line search, before switching to the BFGS method. Possibly the 1-norm should be used as the objective function for the first phase and the 2-norm for the second.

Another possible contribution to errors when using the BFGS method is due to the extremely small angle (approaching zero as the method converges) between the estimated and true curves. The BFGS method approximates the Hessian based on gradient information, and, as shown in section 5, the error in the first-order estimates may be very large when this angle is small. This conjecture is supported by the results of the skewed ellipse case. Here the angles at the corner between the estimated and true curves are much less acute, and the BFGS method produces a result of comparable accuracy to steepest descent, while converging much more rapidly.

It is not the purpose of this paper to compare the two minimization formulations. Rather, these computations are intended to show that either of the cost functions



(a) Direct minimization of area metric.



(b) Minimization of $J = \frac{1}{2}\rho^2$.

FIG. 6.6. *True curve is ellipse* (*dashed*). *Initial estimate* (*light solid*) *is circle of radius* 15 *centered on* $(0, 0)$. (a) *Direct minimization of the area metric* ρ. *Steepest descent algorithm converges* (*heavy solid*) *in* 10 *iterations* (*dotted*). (b) $J = \frac{1}{2}\rho^2$. *BFGS algorithm converges* (*heavy solid*) *in* 11 *iterations* (*dotted*).

developed above are feasible choices for the deposition and etching application. We believe that the results support this conclusion. On the other hand, the results of this section do suggest that the particular numerical formulation used will be an important consideration when applying this technique to practical problems. Many modifications to the numerical formulation are possible—including second-order correction terms at the corners, for example—which may significantly improve performance.

7. Parametrized speed functions. This work is motivated by applications to curve evolution problems. In such cases, the initial curve is assumed known, as is the measured shape at one or more subsequent times. The cost function corresponding to the 1-norm is then

$$J(\lambda) = \sum_k \rho(\mathcal{M}_k, \mathcal{L}(t_k; \lambda)), \tag{48}$$

and the cost function corresponding to the 2-norm is

$$J(\lambda) = \frac{1}{2} \sum_k \rho(\mathcal{M}_k, \mathcal{L}(t_k; \lambda))^2, \tag{49}$$

where $\mathcal{M}_k$ is the kth measured curve, t_k is the time corresponding to the kth measurement, and $\mathcal{L}(t_k; \lambda)$ is the curve obtained by propagating the estimate, with parameter vector λ, to $t = t_k$. The calculation of ρ and ρ_i, and so J and J_i, proceeds exactly as has already been described, with one important difference. The Ψ_i term that appears in (26) is no longer given explicitly by the parametrization. Rather, the λ dependence of $\Psi(x; \lambda)$ arises from the variation of β with λ in the evolution equation

$$\Psi_t + \beta(\,\cdot\,; \lambda)\|\nabla\Psi\| = 0. \tag{50}$$

Thus we will need to determine Ψ_i from β_i instead of writing it directly.

One way to accomplish this is to differentiate both sides of (50) with respect to the ith parameter. The result is a PDE for the gradient, coupled to the original evolution equation. This type of *sensitivity equation* approach is described further in [6, 7]. For the level set evolution equation, assuming only space and time dependence of the speed function, the result of the differentiation is as follows:

$$\Psi_{ti} + \beta_i\|\nabla\Psi\| + \frac{\beta}{\|\nabla\Psi\|}\,(\Psi_x\Psi_{xi} + \Psi_y\Psi_{yi}) = 0. \tag{51}$$

Exchanging the order of differentiation gives

$$\Psi_{it} + \beta_i\|\nabla\Psi\| + \frac{\beta}{\|\nabla\Psi\|}\,(\Psi_x\Psi_{xi} + \Psi_y\Psi_{yi}) = 0. \tag{52}$$

Finally, denote the appropriate gradient term by $S^{(i)} := \Psi_i$ and write

$$S_t^{(i)} + \beta_i\|\nabla\Psi\| + \frac{\beta}{\|\nabla\Psi\|}\,(\Psi_x S_x^{(i)} + \Psi_y S_y^{(i)}) = 0 \tag{53}$$

or

$$S_t^{(i)} + \beta\langle \nu, \nabla S^{(i)}\rangle = -\beta_i\|\nabla\Psi\|, \tag{54}$$

where ν is the outward pointing unit normal to the estimated level set. Once (50) is solved, $\nabla\Psi$ can be computed. Since β is known, and β_i can be computed, (54) is a linear first-order PDE. Initially the various $S^{(i)}$ are everywhere zero, since the starting curve for the model is exactly the true starting curve and does not depend on the speed function β. Boundary conditions are not needed as long as the estimated curve evolves outward everywhere on the boundary. If this is not the case, periodic boundary conditions may be applied [18]. Important behavior, such as curvature or orientation dependence of the speed function, will require additional terms in (54). Once (54) is solved, the computed value for Ψ_i (that is, $S^{(i)}$) is substituted into (26). Because only the values on the estimated curve are required, it would be highly desirable to implement a local solution method, such as the narrowband techniques of Sethian [18]. We note again that in all other ways, computation in the time-dependent evolutionary case will follow the procedures developed in this paper.

Acknowledgments. The authors would like to thank Allen Tannenbaum of the University of Minnesota and Anthony Yezzi of the Massachusetts Institute of Technology for their help on the theoretical and numerical analysis of curve evolution. Thanks also to Paul Bray of Texas Instruments and Ric Borges and Brian Li of Avant! Corporation for very helpful discussions on process simulation and modeling.

REFERENCES

[1] D. Adalsteinsson and J. A. Sethian, *A level set approach to a unified model for etching, deposition, and lithography* I: *algorithms and two-dimensional simulations*, J. Comput. Phys., 120 (1995), pp. 128–144.

[2] J. M. Berg, A. Yezzi, and A. R. Tannenbaum, *Phase transitions, curve evolution, and the control of semiconductor manufacturing processes*, Proceedings of the 35th IEEE CDC, Kobe, Japan, 1996, pp. 3376–3381.

[3] J. M. Berg, A. Yezzi, and A. R. Tannenbaum, *Phase Transitions, Curve Evolution, and the Control of Semiconductor Manufacturing Processes*, Preprint 1454, Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, MN, 1997.

[4] J. M. Berg, A. Yezzi, and A. R. Tannenbaum, *Toward Real-Time Estimation of Surface Evolution in Plasma Etching: Isotropy, Anisotropy, and Self-Calibration*, Proceedings of the 36th IEEE CDC, San Diego, CA, 1997, pp. 860–865.

[5] J. M. Berg, A. Yezzi, and A. R. Tannenbaum, *Curve evolution models for real-time identification with application to plasma etching*, IEEE Trans. Automat. Control, 44 (1999), pp. 99–101.

[6] J. Borggaard and J. Burns, *A sensitivity equation approach to shape optimization in fluid*, in Flow Control, M. D. Gunzburger, ed., Springer-Verlag, New York, 1995.

[7] J. Borggaard and J. Burns, *A PDE sensitivity equation method for optimal aerodynamic design*, J. Comput. Phys., 136 (1997), pp. 366–384.

[8] T. S. Cale, M. B. Chaara, and A. Hasper, *Estimating local deposition conditions and kinetic parameters using film profiles*, Proc. Mat. Res. Soc. Symp., 260 (1992), pp. 393–398.

[9] M. P. do Carmo, *Differential Geometry of Curves and Surfaces*, Prentice–Hall, Englewood Cliffs, NJ, 1976.

[10] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., John Wiley, New York, 1987.

[11] K. Holmström and M. Björkman, *The TOMLAB NLPLIB toolbox for nonlinear programming*, Adv. Model. Optim., 1 (1999), pp. 70–86.

[12] K. Holmström, *The TOMLAB optimization environment in Matlab*, Adv. Model. Optim., 1 (1999), pp. 47–69.

[13] B. B. Kimia, A. Tannenbaum, and S. W. Zucker, *On the evolution of curves via a function of curvature*, I: *The classical case*, J. Math. Anal. Appl., 163 (1992), pp. 438–458.

[14] J. P. McVittie, J. C. Rey, A. J. Bariya, M. M. IslamRaja, L. Y. Cheng, S. Ravi, and K. C. Saraswat, *SPEEDIE: A Profile Simulator for Etching and Deposition*, Proceedings of the SPIE Symposium: Advanced Techniques for Integrated Circuits Process., 1392 (1990), pp. 126–138.

[15] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed., McGraw–Hill, New York, 1976.

[16] F. Santosa, *A level-set approach for inverse problems involving obstacles*, ESAIM: Control Optim. Calc. Var., 1 (1996), pp. 17–33.

[17] K. Siddiqi, B. Kimia, C.-W. Wang, *Geometric shock-capturing ENO schemes for subpixel interpolation, computation, and curve evolution*, Graphical Models and Image Processing, 59 (1997), pp. 278–301.

[18] J. A. Sethian, *Level Set Methods: Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, Cambridge University Press, Cambridge, UK, 1996.

[19] J. A. Sethian and D. Adalsteinsson, *An Overview of Level Set Methods for Etching, Deposition, and Lithography Development*, IEEE Trans. Semiconduct. Manufacturing, 10 (1997), pp. 167–184.

[20] J. A. Sethian and D. Adalsteinsson, *Terrain: Topography Simulation for IC Technology*, Technology Modeling Associates, Inc., Sunnyvale, CA, 1997.

[21] E. Zawaideh and N. S. Kim, *A plasma etching model based on a generalized transport approach*, J. Appl. Phys., 62 (1987), pp. 2498–2507.

SIAM J. Control Optim.
Vol. 37, No. 5, pp. 1394–1418

LACK OF TIME-DELAY ROBUSTNESS FOR STABILIZATION OF A STRUCTURAL ACOUSTICS MODEL[*]

GEORGE AVALOS[†], IRENA LASIECKA[‡], AND RICHARD REBARBER[§]

Abstract. In this paper we consider a natural robustness question for a model for structural acoustics. This model, which has been of great interest in recent years, is represented by a wave equation in $\mathbb{R}^2$ coupled to a Kelvin–Voigt beam; the coupling is natural physically, and is represented mathematically by highly unbounded operators. We assume that the observation consists of point evaluation of the beam position, the beam velocity, and the wave velocity. We are interested in the effect of arbitrarily small delays in the feedback loop on a controller that uses these observations. We show that it is not possible to construct a dynamic stabilizer of a very general form—including static feedback—such that the stabilization is robust with respect to delays in the feedback loop. In order to do this we need to carefully analyze the input-to-output map. Finally, we relate these results to already existing numerical results obtained for a Galerkin approximation of the system.

Key words. time delays, robust stabilization, coupled partial differential equations, transfer functions, dynamic stabilization, structural acoustics

AMS subject classifications. 93C20, 93D09, 93D15, 93D25, 35M10

PII. S0363012997331135

1. Introduction.

In this paper we consider a natural robustness question for a model for structural acoustics. This model has been of great interest in recent years (see Banks et al. [6, 7, 8, 9, 10, 11, 12] and the references therein, Avalos [1], and Avalos and Lasiecka [2, 3]). In this introduction we roughly describe the problem and our results, and we leave the precise problem description and technical details for later sections.

Most of the mathematical analysis of this structural acoustics model has been done with a two-dimensional approximation; see [6, 7, 12] for details about the modeling for this system. We represent an acoustic cavity by a rectangular region in $\mathbb{R}^2$ and we consider one side of the boundary to be a flexible beam and the other three sides to be hard walls. Sound waves inside the cavity are described by a wave equation. The displacement of the flexible beam, which we call the active boundary, is described by a beam equation with Kelvin–Voigt damping. The coupling of these two equations is natural physically and is represented mathematically by highly unbounded operators. In the references above the control is formed by using piezoceramic patches on the active boundary, which produce moment forces when a voltage is applied. The control goal we consider here is stabilization of the system, in particular the attenuation of the acoustic pressure in the cavity. This attenuation might be described by exponential stability of an appropriate state space model, or by input-output stability of an appropriate input-output model.

[*] Received by the editors December 5, 1997; accepted for publication (in revised form) December 2, 1998; published electronically July 22, 1999.
http://www.siam.org/journals/sicon/37-5/33113.html

[†] Department of Mathematics, Texas Tech University, Lubbock, TX 79409 (avalos@me.ttu.edu). The work of this author was partially supported by NSF grant DMS-9710981.

[‡] Department of Applied Mathematics, Thornton Hall, University of Virginia, Charlottesville, VA 22903 (il2v@amsun.apma.virginia.edu). The work of this author was partially supported by NSF grant DMS-9504822 and ARO grant MA DAAAHO4-96-0059.

[§] Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, NE 68588-0323 (rrebarbe@math.unl.edu). The work of this author was partially supported by NSF grant DMS-9623392.

In this paper we assume that the full state of the system is not available for use by the controller and that we have access to the following observations of the system: point evaluation of the beam displacement at some point or points on the beam, point evaluation of the beam velocity at some point or points on the beam, and point evaluation of the acoustic pressure at some points in the cavity. We consider controls determined by a dynamic compensator that uses these observations as its input.

We are interested in the effect of arbitrarily small delays in the feedback loop on the stability of the closed-loop system. Roughly speaking, we say that feedback stabilization of a system is *robust with respect to delays* if delays introduced into the feedback loop do not destroy the stability provided that the delays are sufficiently small. In this paper we answer the following question: *Is it possible to construct a dynamic stabilizer (or a static feedback) for this structural acoustics systems such that the stabilization is robust with respect to small delays in the feedback loop?*

For a large class of dynamic stabilizers and static feedbacks we show that the answer to this question is negative, since we find that there exists a sequence of delays $\{\varepsilon_j\}$ such that $\varepsilon_j \to 0$ and the closed-loop system with delay ε_j has an unstable transfer function pole. See section 5 for precise statements of the relevant results. In Banks, Demetriou, and Smith [8] the effect of delays on an H^∞ dynamic compensator for this system is studied numerically. In particular, in [8] delays are inserted in three places in the feedback loop, and the robustness with respect to these delays is studied for a Galerkin approximation of the closed-loop system. In section 6 we adapt our results to the setup in [8] and give a heuristic explanation for the results in [8] based on a frequency domain analysis.

The approach we use for these lack-of-robustness results is systems theoretic in the sense that we first show that our system is in a particular class of systems and then appeal to theorems about this class. The class in question is the class of *regular* systems, which is a very general class whose basic properties were developed by Weiss in [25] and [26], and for which there is now a substantial amount of control machinery; see for instance [17, 18, 19, 21, 22, 23, 27, 29]. In order to show that our system is regular when the observation includes point observations of the beam displacement and velocity, but does not include point observations of the acoustic pressure, we rely heavily on results by Avalos and Lasiecka [2] and Avalos [1]. When the observation includes point evaluation of acoustic pressure, proving regularity involves a careful analysis of the map from control to observation. This regularity is of interest independent of robustness questions, since many other control questions, e.g., adaptive control [18], stability radii [19], or dynamic stabilization [27], can be studied in the regular systems framework.

In the case when the observation does not include point evaluation of acoustic pressure, we have a natural (A, B, C, D) state space realization for the system, and our results can be stated in terms of exponential stability. In the case where the observation includes point evaluation of acoustic pressure, the state space that accommodates both the control and observation does not coincide with the basic energy space—it requires $(1/2)$ more derivatives in the wave variable—so we present our results in input-output form.

The paper is organized as follows. In section 2 we present the controlled, observed structural acoustics system. In section 3 we show that when the observation is point evaluation of acoustic pressure, then the input-to-output map is well posed, that is, for any $T > 0$ this map is in $\mathcal{L}(L^2[0, T; U], L^2[0, T; Y])$, where U is the control space and Y is the observation space. In section 4 we prove regularity of the system. In section 5 we discuss some results on robustness and lack of robustness with respect to delays, which can be immediately applied to our system. In section 6 we relate and compare our results to the numerical studies in [8]. In all the sections except section 3, we allow the acoustic cavity to be either a rectangle or any region in $\mathbb{R}^2$ with a smooth boundary. In section 3 we require the cavity to be a rectangle. In [5] the results in section 3 are obtained, using very different techniques, when Ω is a region in $\mathbb{R}^2$ or $\mathbb{R}^3$ with smooth boundary.

2. The controlled, observed structural acoustics system. In this section we present and analyze the partial differential equation model. This model is based on the one in [9, 12], but we will use a scaled, slightly abstracted version. Let Ω be either a rectangular region in $\mathbb{R}^2$ or a region in $\mathbb{R}^2$ with Lipschitz boundary Γ. Let Γ_0 be a smooth (C^2) segment of Γ with endpoints a and b. Let $z = z(t,x)$ for $t \in [0,\infty)$ and $x \in \Omega$, let $v = v(t,\xi)$ for $t \in [0,\infty)$ and $\xi \in \Gamma_0$, and let $\partial/\partial\nu$ denote the outward normal derivative. Let $U = \mathbb{R}^r$ and $B \in \mathcal{L}(U, H^{-\alpha}(\Gamma_0))$, where α will be specified throughout to be

$$\begin{aligned} &\alpha = \frac{7}{4} \text{ when } \Omega \text{ is rectangular} \\ &\text{and } \alpha = \frac{5}{3} \text{ when } \Omega \text{ has a smooth boundary.} \end{aligned} \tag{2.1}$$

We refer to the following as the *structural acoustics model*:

$$\begin{aligned} z_{tt} &= \Delta z \text{ on } [0,\infty) \times \Omega, \\ \frac{\partial z}{\partial \nu} &= v_t \text{ on } [0,\infty) \times \Gamma_0, \\ \frac{\partial z}{\partial \nu} &= 0 \text{ on } [0,\infty) \times \Gamma \setminus \Gamma_0, \\ v_{tt} &= -\Delta^2 v - \Delta^2 v_t - z_t + Bu \text{ on } [0,\infty) \times \Gamma_0, \\ v(a,t) = v_t(b,t) &= \frac{\partial v(a,t)}{\partial x} = \frac{\partial v(b,t)}{\partial x} = 0 \quad \forall\, t \in [0,\infty). \end{aligned} \tag{2.2}$$

The model discussed in [6, 7, 8, 9, 10, 11, 12], suitably scaled, is a special case of (2.2): in this case Ω is a rectangular region, Γ_0 is the bottom side of the rectangle, and B is of the form

$$B = \sum_{i=1}^{r} \alpha_i \delta'(\eta_i),$$

where $\delta'(\eta_i)$ are derivatives of delta functions evaluated at $\eta_i \in \Gamma_0$ and $\alpha_i \in \mathbb{R}$. The physical interpretation for this particular control operator is that its control action is realized by the strategic placement of piezoelectric ceramic patches on the (flexible) boundary Γ_0; a voltage is subsequently applied through these patches and the resulting bending moments can be interpreted as derivatives of delta functions.

The observation $y(t)$ considered in [6] and [8] is a vector with components of the form $v(t,\xi_0)$ for $\xi_0 \in \Gamma_0$, $v_t(t,\xi_0)$ for $\xi_0 \in \Gamma_0$, and $z_t(t,x_0)$ for $x_0 \in \overline{\Omega}$; these observations are chosen because they are likely to be physically observable. The acoustic pressure is proportional to z_t, so we often refer to an observation of z_t as an observation of acoustic pressure.

In [1], [2] the control system (2.2) was put into the following state-space form. Define the operator $A : L^2(\Omega) \supset \mathcal{D}(A) \to L^2(\Omega)$ by $A = -\Delta$ with

$$\mathcal{D}(A) = \left\{ z \in L^2(\Omega)/\mathbb{R} \cap H^2(\Omega) \middle| \frac{\partial z}{\partial \nu} = 0 \text{ on } \Gamma \right\},$$

where $L^2(\Omega)/\mathbb{R} = \{f \in L^2(\Omega) \mid \int_\Omega f = 0\}$; viz, $L^2(\Omega)/\mathbb{R}$ is the orthogonal complement of the space of constant functions in $L^2(\Omega)$. A is symmetric positive definite on $L^2(\Omega)$, so fractional powers of A are well defined. In particular, from [14] we have

$$\mathcal{D}(A^{\beta/2}) = L^2(\Omega)/\mathbb{R} \cap H^\beta(\Omega) \quad \text{for } \beta \in \left[0, \frac{3}{2}\right). \tag{2.3}$$

Define $\mathring{\mathbf{A}} : L^2(\Gamma_0) \supset \mathcal{D}(\mathring{\mathbf{A}}) \to L^2(\Gamma_0)$ by $\mathring{\mathbf{A}} = \Delta^2$ with

$$\mathcal{D}(\mathring{\mathbf{A}}) = H^4(\Gamma_0) \cap H_0^2(\Gamma_0).$$

$\mathring{\mathbf{A}}$ is symmetric positive definite on $L^2(\Gamma_0)$, so its fractional powers are well defined. In particular, by [14] we have

$$\mathcal{D}\left(\mathring{\mathbf{A}}^{\beta/4}\right) = H_0^\beta(\Gamma_0) \quad \text{for } \beta \in \left[0, \frac{5}{2}\right). \tag{2.4}$$

Since $\mathring{\mathbf{A}}$ is symmetric positive definite on a Hilbert space and the dual space of H_0^β is $H^{-\beta}$, we see that

$$\mathcal{D}\left(\mathring{\mathbf{A}}^{\beta/4}\right)' = H^{-\beta}(\Gamma_0) \quad \text{for } \beta \in \left[0, \frac{5}{2}\right). \tag{2.5}$$

Define

$$H_1 = \mathcal{D}(A^{\frac{1}{2}}) \times L^2(\Omega) = \left(L^2(\Omega)/\mathbb{R} \cap H^1(\Omega)\right) \times L^2(\Omega) \tag{2.6}$$

and

$$H_0 = \mathcal{D}\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right) \times L^2(\Gamma_0) = H_0^2(\Gamma_0) \times L^2(\Gamma_0).$$

Let $A_1 : H_1 \supset \mathcal{D}(A_1) \to H_1$ and $A_0 : H_0 \supset \mathcal{D}(A_0) \to H_0$ be defined by

$$\begin{aligned} A_1 &:= \begin{bmatrix} 0 & I \\ -A & 0 \end{bmatrix} \quad \text{with} \\ D(A_1) &= \{[z_1, z_2]^T \in D(A) \times D(A^{\frac{1}{2}})\} \end{aligned}$$

and

$$\begin{aligned} A_0 &:= \begin{bmatrix} 0 & I \\ -\mathring{\mathbf{A}} & -\mathring{\mathbf{A}} \end{bmatrix} \quad \text{with} \\ D(A_0) &= \left\{ [v_1, v_2]^T \in \left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)\right]^2 \ni v_1 + v_2 \in D\left(\mathring{\mathbf{A}}\right) \right\}. \end{aligned}$$

We define the Neumann map N on $H^s(\Gamma)$ for $s > -1/2$ by setting $Ng := z$ for $g \in H^s(\Gamma)$, where z is the unique solution in $L^2(\Omega)/\mathbb{R}$ of the equation

$$\langle Az, v\rangle_{[D(A^{\frac{1}{2}})]' \times D(A^{\frac{1}{2}})} = \int_\Omega \nabla z \cdot \nabla v = \langle g, v\rangle_{H^s(\Gamma)\times[H^s(\Gamma)]'}$$

for all $v \in D(A^{\frac{1}{2}})$. From [14], we have that

$$N \in \mathcal{L}\left(L^2(\Gamma_0), \mathcal{D}(A^{\frac{3}{4}-\epsilon})\right) \text{ for arbitrary } \epsilon > 0,$$

and this boundedness further implies that

$$AN \in \mathcal{L}\left(L^2(\Gamma_0), [\mathcal{D}(A^{\frac{1}{4}+\epsilon})]'\right) \text{ for arbitrary } \epsilon > 0. \tag{2.7}$$

We also define $\gamma : H^1(\Omega) \to H^{\frac{1}{2}}(\Gamma_0)$ by $\gamma(z) = z|_{\Gamma_0}$ and

$$C = \begin{bmatrix} 0 & 0 \\ 0 & \gamma^* \end{bmatrix},$$

so $C \in \mathcal{L}(H_0, \{0\} \times [D(A^{\frac{1}{2}})]')$. Let $\mathcal{X} = H_1 \times H_0$; we refer to the usual product norm on $\mathcal{X}$ by $\|\cdot\|$, and other norms will be indicated by an appropriate subscript. Now define $\mathcal{A} : \mathcal{D}(\mathcal{A}) \supset \mathcal{X} \to \mathcal{X}$ by

$$\mathcal{A} := \begin{bmatrix} A_1 & C \\ -C^* & A_0 \end{bmatrix}$$

with

$$\begin{aligned} D(\mathcal{A}) \quad = \quad & \{[z_1, z_2, v_1, v_2]^T \in [D(A^{\frac{1}{2}})]^2 \times \left[D\left(\mathring{\mathbf{A}}^{\frac{1}{2}}\right)\right]^2 \text{ such that} \\ & -z_1 + Nv_2 \in D(A) \text{ and such that } v_1 + v_2 \in D\left(\mathring{\mathbf{A}}\right) \Big\}. \end{aligned}$$

Let

$$X(t) = [z(t), z_t(t), v(t), v_t(t)]^T. \tag{2.8}$$

If $u(t) \equiv 0$, then (2.2) is formally equivalent to

$$\dot{X}(t) = \mathcal{A}X(t). \tag{2.9}$$

It is shown in Theorem 1.1 in [1] that $\mathcal{A}$ is the generator of a strongly continuous semigroup $S(t)$ on $\mathcal{X}$; see also Banks and Smith [11].

Let $R(s, \mathcal{A}) = (sI - \mathcal{A})^{-1}$. For $\alpha \in \mathbb{R}$, let $\mathbb{C}_\alpha := \{s \in \mathbb{C} \mid \mathbf{Re}(s) > \alpha\}$. The system under consideration in this paper differs from the system in [3] in that the wave equation here has Neumann conditions on $\Gamma \setminus \Gamma_0$, while the wave equation in [3] has Dirichlet conditions. The arguments in [3] can be easily modified to show that $S(t)$ is not exponentially stable, so Proposition 2 in Prüss [20] implies that

$$\sup_{s \in \mathbb{C}_0} \|R(s, \mathcal{A})\|_{\mathcal{L}(\mathcal{X})} = \infty. \tag{2.10}$$

The arguments in [3] can also be modified to show that $S(t)$ is strongly stable and that

$$\sigma(\mathcal{A}) \cap i\mathbb{R} = \emptyset, \tag{2.11}$$

where $\sigma(\mathcal{A})$ is the spectrum of $\mathcal{A}$. Let $s \in i\mathbb{R}$. If $\{s_n\} \subset \overline{\mathbb{C}}_0$, $s = \lim_{n\to\infty} s_n$, and $\{\|R(s_n, \mathcal{A})\|\}$ is unbounded, then $s \in \sigma(\mathcal{A})$, which is in contradiction to (2.11). Hence the only way (2.10) can be true is if

$$\limsup_{|s|\to\infty, s\in\mathbb{C}_0} \|R(s, \mathcal{A})\|_{\mathcal{L}(\mathcal{X})} = \infty. \tag{2.12}$$

Condition (2.12) is important for our lack-of-robustness results.

Let

$$\mathcal{B} = [0, 0, 0, B]^T \in \mathcal{L}\left(U, \{0\}^3 \times H^{-\alpha}(\Gamma_0)\right).$$

Then (2.2) is formally equivalent to

$$\dot{X}(t) = \mathcal{A}X(t) + \mathcal{B}u(t), \tag{2.13}$$

which holds pointwise in time in

$$X_{-1} := [\mathcal{D}(\mathcal{A}^*)]'. \tag{2.14}$$

$S(t)$ extends to a semigroup on $[\mathcal{D}(\mathcal{A}^*)]'$ and (2.2) is also formally equivalent to

$$z_{tt} = -Az + ANv_t, \tag{2.15}$$

$$v_{tt} + \mathring{\mathbf{A}}v_t + \mathring{\mathbf{A}}v = -N^*Az_t + Bu, \tag{2.16}$$

which holds pointwise in time in $[\mathcal{D}(A)]'$ and $[\mathcal{D}(\mathring{\mathbf{A}})]'$, respectively.

We now discuss the observations of this system. Let C_η be a point evaluation at η; we will use the same notation when $\eta \in \Gamma_0$ or when $\eta \in \overline{\Omega}$. Let $\xi_0 \in \Gamma_0$ and $x_0 \in \overline{\Omega}$. We define the following operators on $\mathcal{X}$ by

$$\mathcal{C}_1 := [0, 0, C_{\xi_0}, 0], \qquad \mathcal{C}_2 := [0, 0, 0, C_{\xi_0}], \qquad \mathcal{C}_3 := [0, C_{x_0}, 0, 0]. \tag{2.17}$$

We are ultimately interested in observations that contain several terms of this type, but it is most convenient to analyze these separately; in particular, our analysis of $\mathcal{C}_3$ necessarily will be much different (and more difficult) than our analysis of $\mathcal{C}_1$ or $\mathcal{C}_2$.

We wish to show that (2.2) with any of these observations is a *regular* system. For a detailed discussion of regular systems, see Weiss [25, 26]. Definition 2.1 given below for a *well-posed*, controlled, observed system is not as detailed as that given in [25], but it is equivalent when the system is given by a boundary-controlled partial differential equation of the type we are currently considering.

Definition 2.1. *Let $\mathcal{X}, Y$, and U be Hilbert spaces. A system with state $X(t) \in \mathcal{X}$, input $u(t) \in U$, and observation $y(t) \in Y$ is well posed if for some (and hence all) $T > 0$,*

(1) *$X(t) = S(t)X(0)$ for a strongly continuous semigroup $S(t)$ when $u(t) \equiv 0$;*
(2) *The map $u(\cdot) \to X(T)$ is bounded from $L^2[0, T; U]$ into $\mathcal{X}$;*
(3) *The map $X(0) \to y(\cdot)$ is bounded from $\mathcal{X}$ into $L^2[0, T; Y]$ when $u(t) \equiv 0$;*

(4) *The map* $\mathcal{T} : u(\cdot) \to y(\cdot)$ *with* $X(0) = 0$ *is bounded from* $L^2[0,T;U]$ *into* $L^2[0,T;Y]$.

For many purposes we need only be concerned with the input-output map $\mathcal{T}$.

DEFINITION 2.2. *If an input-output map* $\mathcal{T}$ *satisfies condition* (4)*, we say that it is a well-posed input-output system.*

If the controlled system is formally represented by (2.13) and condition (2) of Definition 2.1 is satisfied, then we say that $\mathcal{B}$ is an *admissible control (input) operator* for $S(t)$. If the uncontrolled system with observation is formally represented by (2.9) and

$$y(t) = \mathcal{C}X(t) \tag{2.18}$$

and condition (3) of Definition 2.1 is satisfied, then we say that $\mathcal{C}$ is an *admissible observation (output) operator* for $S(t)$. From Proposition 3.2 in [25], condition (4) in Definition 2.1 implies that if

$$L^2_\alpha[0,\infty,U] := \left\{ f \in L^2_{\mathrm{loc}}[0,\infty;U] \mid \int_0^\infty \|f(t)\|^2_U e^{-2\alpha t}\,dt < \infty \right\},$$

then there exists $\beta \in \mathbb{R}$ such that

$$\mathcal{T} \in \mathcal{L}(L^2_\beta[0,\infty;U), L^2_\beta[0,\infty;Y)). \tag{2.19}$$

While the well-posedness of the system with the observation operator $\mathcal{C}_1$ follows from basic Sobolev embeddings and the interior smoothness guaranteed by the state-space topology, this is not the case for the observation operators $\mathcal{C}_2$ and $\mathcal{C}_3$. Indeed, the topology generated by the state space $\mathcal{X}$ allows for point evaluation of the v component, but it does not allow us to define a pointwise evaluation of v_t—roughly speaking, $1/2+\varepsilon$ derivatives are missing. In order to handle this difficulty in the case of $\mathcal{C}_2$, we shall use "additional" smoothness results developed for structural acoustic problems in [2], which say that v_t has greater smoothness than the state space guarantees; in particular, $v_t \in L^2(0,T;H^2(\Gamma_0))$. This allows us to use Sobolev's imbeddings to obtain the well-posedness with the $\mathcal{C}_2$ observation.

In the case of the observation operator $\mathcal{C}_3$, the situation is much more delicate. The state space guarantees L^2 smoothness of z_t in the cavity. However, in order to use a Sobolev's imbedding one would need to have $z_t \in H^{3/2+\varepsilon}(\Omega)$, so $3/2+\varepsilon$ derivatives are missing. It can be shown (even for the one-dimensional example) that $L^2(0,T;H^2(\Gamma_0))$ tangential smoothness of the boundary input v_t does not produce sufficient interior smoothness for the variable z_t to make Sobolev embeddings useful for interior-point evaluation. In fact, one could show (with additional nontrivial work) that the maximal internal smoothness of z_t is $H^{1/2}(\Omega)$, but even this does not suffice to take pointwise evaluation in two dimensions. To cope with this difficulty, we resort to completely different arguments, which are based on microlocal analysis for the general smooth domains, and on very delicate calculations involving harmonic analysis when Ω is a rectangle. The rectangular case is dealt with in section 3, while the smooth domain case is treated in [5].

PROPOSITION 2.3. *The system* (2.2) *with observation* $y(t) = \mathcal{C}_1X(t)$ *or observation* $y(t) = \mathcal{C}_2X(t)$ *is well posed.*

Proof. Condition (1) in Definition 2.1 has already been established in [2]. Condition (2) follows immediately from Lemma 2.1 in [2]. To verify condition (3), first note that $\mathcal{C}_1$ is bounded on $\mathcal{X}$, hence is an admissible observation operator for any semigroup on $\mathcal{X}$. Now let the projections P_0 and P_1 be defined by

$$P_0[z_1, z_2, v_1, v_2]^T = [v_1, v_2]^T,$$
$$P_1[z_1, z_2, v_1, v_2]^T = [z_1, z_2]^T.$$

Theorem 1.1 in [2] shows that for any $X_0 \in \mathcal{X}$ and $T > 0$,

$$P_0 S(\cdot) X_0 \in L^2\left(0, T; \mathcal{D}(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times \mathcal{D}(\mathring{\mathbf{A}}^{\frac{1}{2}})\right).$$

Let $X_0 \in \mathcal{X}$ and $X(t) = S(t)X_0$ be of the form (2.8). Using the Sobolev embedding theorem and the fact that $\mathcal{D}(\mathring{\mathbf{A}}^{1/2}) = H_0^2(\Gamma_0)$, we see that $\mathcal{C}_2 S(\cdot) X_0 = v_t(\cdot, \xi_0) \in L^2(0,T)$. By the principle of uniform boundedness, $\mathcal{C}_2$ is an admissible observation operator for $S(t)$.

To verify condition (4), note that Proposition 2.3 in [2] shows that for any $T > 0$ and $u \in L^2[0, T; U]$,

$$P_0 \int_0^{\cdot} S(\cdot - \tau) \mathcal{B} u(\tau)\, d\tau \in L^2\left(0, T; \mathcal{D}(\mathring{\mathbf{A}}^{\frac{1}{2}}) \times \mathcal{D}(\mathring{\mathbf{A}}^{\frac{1}{2}})\right).$$

Using the same reasoning as we used for the admissibility of $\mathcal{C}_1$ and $\mathcal{C}_2$, we see that (2.2) with this observation is well posed. □

If Z is a Banach space, define $H_\alpha^\infty(Z)$ to be all analytic Z-valued functions $\mathbf{H}(s)$ for which there exists $M > 0$ such that $\|\mathbf{H}(s)\| < M$ for all $s \in \mathbb{C}_\alpha$. If $\alpha = 0$ and if Z is clear from the context, we denote $H_\alpha^\infty(Z)$ by H^∞. Let $y(t)$ be the observation of a system resulting from a given u and zero initial conditions. If we denote the Laplace transform of y by $\hat{y}$ and if

$$\hat{y}(s) = \mathbf{H}(s)\hat{u}(s),$$

then we say that $\mathbf{H}(s)$ is the *transfer function* of the system. From Proposition 3.2 in [25], (2.19) is equivalent to $\mathbf{H} \in H_\beta^\infty(\mathcal{L}(U, Y))$, leading to the following definition.

DEFINITION 2.4. *A transfer function $\mathbf{H}(s)$ is well posed if $\mathbf{H} \in H_\beta^\infty(\mathcal{L}(U, Y))$ for some $\beta \in \mathbb{R}$.*

DEFINITION 2.5. *A transfer function $\mathbf{H}$ (or, equivalently, an input-output map $\mathcal{T}$) is* input-output stable *if*

$$\mathbf{H} \in H^\infty(\mathcal{L}(U, Y)) \text{ (or, equivalently, if } \mathcal{T} \in \mathcal{L}(L^2[0, \infty; U), L^2[0, \infty; Y))).$$

DEFINITION 2.6. *A system is* regular *if it is well posed and if the following condition is satisfied by its transfer function $\mathbf{H}(s)$:*

(5) $\lim_{s \to \infty, s \in \mathbb{R}} \mathbf{H}(s) =: D \in \mathcal{L}(U, Y)$ *exists.*

In this case we say that D is the feedthrough of the transfer function and of the system.

DEFINITION 2.7. *If an input-output system is well posed and the transfer function satisfies condition* (5)*, then we say that the input-output system is regular and that its transfer function is regular.*

It is shown in [25] that if a controlled, observed system with semigroup generator $\mathcal{A}$, input operator $\mathcal{B}$, and observation operator $\mathcal{C}$ is regular, it has the following state space representation, with state space $\mathcal{X}$:

$$\begin{aligned} \dot{X}(t) &= \mathcal{A} X(t) + \mathcal{B} u(t), \\ y(t) &= \mathcal{C}_\Lambda X(t) + D u(t), \end{aligned} \tag{2.20}$$

where $\mathcal{C}_\Lambda$ is the *Lebesgue extension* of $\mathcal{C}$; see [25].

Remark 2.8. If a controlled, observed system is well posed but not regular, it has a state space representation, but the abstract form of the observation (when the control is present) is much more complicated (and less analogous to finite dimensional systems) than that in (2.20) (see Salamon [24]).

Since $\mathcal{B}$ is an admissible input operator and $\mathcal{C}_1$ is a bounded observation operator, we get the following simple result, which follows immediately from [25].

PROPOSITION 2.9. *The system* (2.2) *with observation* $\mathcal{C}_1 X(t)$ *is regular.*

3. Observation of acoustic pressure. In the case where the observation operator is $\mathcal{C}_3$ (given in (2.17)) the well-posedness of the system is a much more delicate issue than the case treated by Proposition 2.3. This observation is *not* an admissible observation for $S(t)$ when the state space is $\mathcal{X}$. However, we will see in Corollary 3.4 that when Ω is a rectangular region, the input-output system is well posed. In Theorem 4.8 we see that the transfer function for (2.2) with this observation is regular, i.e., it also satisfies condition (5) in Definition 2.6.

Let $T > 0$ and let Ω be a rectangular domain in $\mathbb{R}^2$ with boundary Γ. Let $[z(t), z_t(t)]^T$ be the solution of the wave equation

$$\begin{aligned} &z_{tt}(t,x,y) = \Delta z(t,x,y) \quad \text{on } (0,T)\times\Omega, \\ &\frac{\partial z}{\partial\nu}(t,\zeta) = \mu(t,\zeta) \quad \text{on } (0,T)\times\Gamma, \qquad (3.1) \\ &z(0,x,y) = z_t(0,x,y) = 0 \quad \text{on } \Omega. \end{aligned}$$

With boundary data $\mu \in L^2(0,T;H^{5/4}(\Gamma))$, we already know from Avalos [4] that

$$[z, z_t] \in C\left([0,T]; H^1(\Omega)\times L^2(\Omega)\right).$$

We assume that $\mu(t,\cdot) = 0$ for $t > T$. We now prove the following "trace" result, which allows for pointwise evaluation of the velocity z_t at a point in $\overline{\Omega}$.

THEOREM 3.1. *For every fixed* $(x_0,y_0)\in\overline{\Omega}$, *the mapping* $\mu \to z_t(\cdot,x_0,y_0)$ *is in*

$$\mathcal{L}(L^2(0,T;H^{5/4}(\Gamma)), L^2(0,T)).$$

In particular, there exists $M > 0$, *independent of* (x_0,y_0), *such that for every* $\mu \in L^2(0,T;H^{5/4}(\Gamma))$,

$$\|z_t(\cdot,x_0,y_0)\|_{L^2(0,T)} \le M\,\|\mu\|_{L^2\left(0,T;H^{\frac{5}{4}}(\Gamma)\right)}. \qquad (3.2)$$

Remark 3.2. Note that this result *does not* follow from a direct application of the classical Sobolev embedding theorem.

Proof. Without loss of generality, we can set $\Omega := \{(x,y)\in\mathbb{R}^2 \mid 0 < x,y < \pi\}$, and since the mapping of concern here is linear, we can set $u \equiv 0$ except on the side $\{y = 0, 0 \le x \le \pi\}$. Fix $(x_0,y_0)\in\overline{\Omega}$. Let $\{\lambda_{mn},\Phi_{mn}\}_{m,n=1}^{\infty}$ denote, respectively, the eigenvalues and orthonormalized eigenfunctions of the operator A defined in section 2. These are given explicitly by

$$\begin{aligned} &\lambda_{mn} = n^2 + m^2 \quad \text{for } m,n = 1,2,\dots, \\ &\Phi_{mn}(x,y) = \tfrac{2}{\pi}\cos nx\cos my \quad \text{for } m,n = 1,2,\dots. \end{aligned}$$

We can explicitly write out the solution $[z, z_t]$ of (3.1)—this is done in [4] and [16]—as

$$z(t,x,y) = \sum_{m,n=1}^{\infty} \left\{ \frac{1}{\sqrt{n^2+m^2}} \int_0^t \sin\sqrt{n^2+m^2}(t-\tau)\mu_n(\tau)d\tau \right\} \Phi_{mn}(x,y);$$

$$\text{(3.3)} \quad z_t(t,x,y) = \sum_{m,n=1}^{\infty} \left\{ \int_0^t \cos\sqrt{n^2+m^2}(t-\tau)\mu_n(\tau)d\tau \right\} \Phi_{mn}(x,y),$$

where

$$\mu_n(t) := \frac{2}{\pi} \int_0^{\pi} \mu(t,\xi)\cos(n\xi)\,d\xi.$$

We now need the following proposition.

PROPOSITION 3.3. *For arbitrary $y_0 \in [0,\pi]$, the map $\mu \to z_t(\cdot,\cdot,y=y_0)$ is in*

$$\mathcal{L}(L^2(0,T;H^{\frac{5}{4}}(0,\pi)), L^2(0,T;H^{\frac{3}{4}}(0,\pi))).$$

In particular, there exists $M > 0$, independent of $y_0 \in [0,\pi]$, such that for all $\mu \in L^2(0,T;H^{\frac{5}{4}}(0,\pi))$,

$$\|z_t(\cdot,\cdot,y=y_0)\|_{L^2(0,T;H^{\frac{3}{4}}(0,\pi))} \le M\,\|\mu\|_{L^2(0,T;H^{\frac{5}{4}}(0,\pi))}.$$

Proof of Proposition 3.3. Define the operator $A_\Pi : L^2(0,\pi) \to L^2(0,\pi)$ by

$$A_\Pi = -\frac{d^2}{dx^2} \quad \text{with } \mathcal{D}(A_\Pi) = \left\{ g \in H^2(0,\pi) \middle| \frac{\partial g(\pi)}{\partial x} = \frac{\partial g(0)}{\partial x} = 0 \right\}.$$

A_Π is self-adjoint, positive semidefinite, so its positive fractional powers are well defined. Its respective eigenvalues and orthonormalized eigenvectors $\{\lambda_n, \Phi_n\}_{n=0}^{\infty}$ are given explicitly by

$$\text{(3.4)} \qquad \lambda_n = n^2 \quad \text{and } \Phi_n(x) = \frac{2}{\pi}\cos nx.$$

Then, for g in $L^2(0,\pi)$ and $\eta \ge 0$,

$$A_\Pi^{\frac{\eta}{2}} g = A_\Pi^{\frac{\eta}{2}} \sum_{n=1}^{\infty} (g,\Phi_n)\,\Phi_n = \sum_{n=1}^{\infty} (g,\Phi_n)\,n^\eta \Phi_n$$

with the L^2-convergence of the series on the right-hand side occurring if and only if $g \in D(A_\Pi^{\frac{\eta}{2}})$. Using this equality and the explicit representation of z_t in (3.3), we obtain

(3.5)
$$A_\Pi^{\frac{\eta}{2}} z_t(t,\cdot,y=y_0) = \sum_{m,n=1}^{\infty} n^\eta \left\{ \int_0^t \cos\sqrt{n^2+m^2}(t-\tau)\mu_n(\tau)d\tau \right\} \Phi_n(\cdot)\cos my_0.$$

Furthermore, from [14], for all $g \in H^\eta(0,\pi)$, $0 \le \eta < 3/2$, we can take

$$\text{(3.6)} \qquad \|g\|^2_{H^\eta(0,\pi)} = \left\| A_\Pi^{\frac{\eta}{2}} g \right\|^2_{L^2(0,\pi)}.$$

Using this and the orthonormality of $\{\Phi_n\}$, we have that for every $g \in H^\eta(0,\pi)$, $0 \le \eta \le 3/2$,

$$\|g\|^2_{H^\eta(0,\pi)} = \sum_{n=1}^{\infty} |(g,\Phi_n)|^2\, n^{2\eta}. \tag{3.7}$$

Extending the boundary input μ by zero outside the interval $[0,T]$, we take the Laplace transform in time of both sides of (3.5) with transform variable

$$\lambda = \gamma + i\omega$$

to thereby obtain

$$\widehat{A_\Pi^{\frac{\eta}{2}} z_t}(\lambda,\cdot,y=y_0) = \sum_{m,n=1}^{\infty} \frac{\lambda n^\eta}{\lambda^2+\lambda_{mn}} \widehat{\mu_n}(\lambda)\Phi_n(\cdot)\cos my_0. \tag{3.8}$$

Therefore, we can use the generalized Parseval's relation (see [13, p. 212]), (3.6), (3.8), and (3.7) to obtain for fixed $\eta \in [0,3/2]$ and any $\gamma \ge 0$

$$2\pi \int_0^\infty e^{-2\gamma t} \|z_t(t,\cdot,y=y_0)\|^2_{H^\eta(0,\pi)}\,dt = 2\pi \int_0^\infty e^{-2\gamma t} \left\| A_\Pi^{\frac{\eta}{2}} z_t(t,\cdot,y=y_0)\right\|^2_{L^2(0,\pi)} dt$$

$$= \int_{-\infty}^{\infty} \left\| \widehat{A_\Pi^{\frac{\eta}{2}} z_t}(\gamma+i\omega,\cdot,y=y_0)\right\|^2_{L^2(0,\pi)} d\omega$$

$$= \int_{-\infty}^{\infty} \sum_{n=1}^{\infty} \left| \sum_{m=1}^{\infty} \frac{\lambda}{\lambda^2+\lambda_{mn}} \cos my_0 \right|^2 |\widehat{\mu_n}(\lambda)|^2\, n^{2\eta} d\omega. \tag{3.9}$$

We now fix $\gamma > 0$ and specify that $\eta \in [0,5/4]$. Suppose we can find some $\theta \in [0, 5/4-\eta]$ and some positive constant $C_0 > 0$ such that

$$\frac{1}{n^\theta} \left| \sum_{m=1}^{\infty} \frac{\lambda}{\lambda^2+\lambda_{mn}} \cos my_0 \right| < C_0, \tag{3.10}$$

where C_0 does not depend on $\omega \in \mathbb{R}$ or $n = 1, 2, \ldots$. Then (3.9) is equal to

$$\int_{-\infty}^{\infty} \sum_{n=1}^{\infty} \left| \sum_{m=1}^{\infty} \frac{\lambda n^{-\theta}}{\lambda^2+\lambda_{mn}} \cos my_0 \right|^2 |\widehat{\mu_n}(\lambda)|^2\, n^{2(\eta+\theta)} d\omega$$

$$\le C_0 \int_{-\infty}^{\infty} \sum_{n=1}^{\infty} |\widehat{\mu_n}(\lambda)|^2\, n^{2(\eta+\theta)} d\omega = C_0 \int_{-\infty}^{\infty} \left\| \widehat{A_\Pi^{\frac{\eta+\theta}{2}} \mu}(\gamma+i\omega) \right\|^2_{L^2(0,\pi)} d\omega$$

$$= 2\pi C_0 \int_0^T e^{-2\gamma t} \|\mu(t)\|^2_{H^{\eta+\theta}(0,\pi)}\,dt \le 2\pi C_0 \int_0^T \|\mu(t)\|^2_{H^{\frac{5}{4}}(0,\pi)}\,dt. \tag{3.11}$$

Hence, if we can verify (3.10) for some $\theta \in [0, 5/4-\eta]$, then (3.9) and (3.11) imply that there exists C_1 such that

$$\int_0^T \|z_t(t,\cdot,y=y_0)\|^2_{H^\eta(0,\pi)}\,dt \le C_1 \int_0^T \|\mu(t)\|^2_{H^{5/4}(0,\pi)}. \tag{3.12}$$

To attain the sought-after estimate (3.10), it suffices to ascertain the convergence of each of the following sums for some value of $\theta \in [0, 5/4 - \eta]$, with each convergence being again independent of $\omega \in \mathbb{R}$ and $n \in \mathbb{N}$:

(i) $\frac{1}{n^\theta} \left| \sum_{m=1}^{\infty} \frac{2\gamma\omega^2}{(\gamma^2 - \omega^2 + m^2 + n^2)^2 + 4\gamma^2\omega^2} \right|$,

(ii) $\frac{1}{n^\theta} \left| \sum_{m=1}^{\infty} \frac{\omega(\gamma^2 - \omega^2 + m^2 + n^2)}{(\gamma^2 - \omega^2 + m^2 + n^2)^2 + 4\gamma^2\omega^2} \right|$.

In fact, the absolute convergence of these sums, independent of ω and n, has already been shown in [4] for $\theta = 1/2$ (see in particular equations (3.19), (3.33), (3.55), and (3.76) in [4]). With the estimate (3.10) being established for $\theta = 1/2$, we see that (3.12) is true for any η such that $\theta = 1/2 \in [0, 5/4 - \eta]$ and so in particular for $\eta = 3/4$. □

Conclusion of the proof of Theorem 3.1. Now upon the use of the Sobolev embedding theorem in one dimension and Proposition 3.3, we have that for any $(x_0, y_0) \in [0, \pi] \times [0, \pi]$, there exists $M, \tilde{M} > 0$ such that

$$\int_0^T |z_t(t, x = x_0, y = y_0)|^2 \, dt \le \tilde{M} \int_0^T \|z_t(t, \cdot, y = y_0)\|^2_{H^{\frac{3}{4}}(0,\pi)} \, dt$$
$$\le M \|\mu\|^2_{L^2(0,T;H^{\frac{5}{4}}(0,\pi))} .$$

This finishes the proof of Theorem 3.1. □

COROLLARY 3.4. *The system* (2.2) *with observation* $y(t) = \mathcal{C}_3 X(t)$ *is input-output well posed.*

Proof. The first three equations in (2.2) are of the same form as the wave equation (3.1) when $\mu = 0$ on $\Gamma \setminus \Gamma_0$ and $\mu = v_t$ on Γ_0. Theorem 1.1 in [2] implies that there exists $M_1 > 0$ such that

$$\|v_t\|_{L^2(0,T;H^2(\Gamma_0))} \le M_1 \|u\|_{L^2(0,T;U)} . \tag{3.13}$$

Combining this with (3.2) and the fact that $\|\cdot\|_{H^{5/4}(\Gamma_0)} \le \|\cdot\|_{H^2(\Gamma_0)}$, we see that there exists $M_2 > 0$ such that

$$\|z_t(\cdot, x_0, y_0)\|_{L^2(0,T)} \le M_2 \|u\|_{L^2(0,T;U)} ,$$

so the input-output map from $u \in L^2(0,T;U)$ into $\mathcal{C}_3 X = z_t(\cdot, x_0, y_0) \in L^2(0,T)$ is well posed. □

4. Regularity results. In this section we show that when the observation is $\mathcal{C}_2 X(t)$ the system is regular, and when the observation is $\mathcal{C}_3 X(t)$ the system is input-output regular. Since we have shown that the system is well posed in the former case and input-output well posed in the latter case, it suffices to show in both of these cases that the transfer function satisfies condition (5) in Definition 2.6. We first need a few technical lemmas.

Let $\alpha > 0$—we will eventually specialize to α as in (2.1), but all the following lemmas are true for more general α. For $s > 0$, let $\mathbf{T}(s)$ be defined on $\mathcal{D}(\mathring{\mathbf{A}}^{\alpha/4})'$ by

$$\mathbf{T}(s) := \left[\frac{s^2}{s+1} + \mathring{\mathbf{A}} \right]^{-1} . \tag{4.1}$$

Lemma 4.1. *For*

$$0 \leq \theta < 4 - \alpha,$$

there exists $M > 0$ such that for all $s > 0$,

$$\|\mathbf{T}(s)\|_{\mathcal{L}\mathcal{D}\left(\mathring{\mathbf{A}}^{\alpha/4}\right)',\mathcal{D}\left(\mathring{\mathbf{A}}^{\theta/4}\right)} \leq \frac{M}{\left(1+\frac{s^2}{s+1}\right)^{1-\frac{1}{4}(\theta+\alpha)}}. \tag{4.2}$$

Proof. For θ in the prescribed range and $w \in \mathcal{D}(\mathring{\mathbf{A}}^{\alpha/4})'$, we have

$$\|\mathbf{T}(s)w\|_{\mathcal{D}\left(\mathring{\mathbf{A}}^{\theta/4}\right)} = \left\|\left[\frac{s^2}{s+1}+\mathring{\mathbf{A}}\right]^{-1} w\right\|_{\mathcal{D}\left(\mathring{\mathbf{A}}^{\theta/4}\right)}$$

$$= \left\|\mathring{\mathbf{A}}^{\frac{\theta}{4}}\left[\frac{s^2}{s+1}+\mathring{\mathbf{A}}\right]^{-1} w\right\|_{L^2(\Gamma_0)} = \left\|\mathring{\mathbf{A}}^{\frac{\theta}{4}}\left[\frac{s^2}{s+1}+\mathring{\mathbf{A}}\right]^{-1} \mathring{\mathbf{A}}^{\frac{\alpha}{4}}\mathring{\mathbf{A}}^{-\frac{\alpha}{4}} w\right\|_{L^2(\Gamma_0)}.$$

Using Krein [15, Eq. (5.15), p. 115] and (2.4), we see that if $s > 0$, there exists $M > 0$ such that

$$\|\mathbf{T}(s)w\|_{\mathcal{D}\left(\mathring{\mathbf{A}}^{\theta/4}\right)} \leq \frac{M}{\left[1+\frac{s^2}{s+1}\right]^{1-\frac{1}{4}(\theta+\alpha)}}\left\|\mathring{\mathbf{A}}^{-\frac{\alpha}{4}} w\right\|_{L^2(\Gamma_0)}$$

$$= \frac{M}{\left[1+\frac{s^2}{s+1}\right]^{1-\frac{1}{4}(\theta+\alpha)}}\|w\|_{\mathcal{D}\left(\mathring{\mathbf{A}}^{\alpha/4}\right)'}. \qquad \square$$

Remark 4.2. The same type of estimate could be obtained using the fact that $\mathring{\mathbf{A}}$ generates an analytic semigroup, but we did not use analyticity here.

Lemma 4.3. *For large enough $s > 0$ the operator*

$$I + \frac{s^2}{s+1}\mathbf{T}(s)N^*A(s^2+A)^{-1}AN$$

is boundedly invertible on $\mathcal{L}(\mathcal{D}(\mathring{\mathbf{A}}^{\theta/4}))$ for $\theta \in [0, 4-\alpha)$. In particular, there exists $M > 0$ such that for large enough $s > 0$,

$$\left\|\left(I + \frac{s^2}{s+1}\mathbf{T}(s)N^*A(s^2+A)^{-1}AN\right)^{-1}\right\|_{\mathcal{L}\left(\mathcal{D}\left(\mathring{\mathbf{A}}^{\theta/4}\right)\right)} \leq M. \tag{4.3}$$

Proof. We prove this result by considering the Neumann series for

$$\left(I + \frac{s^2}{s+1}\mathbf{T}(s)N^*A(s^2+A)^{-1}AN\right)^{-1};$$

in particular, we show that

$$\lim_{s\to\infty}\left\|\frac{s^2}{s+1}\mathbf{T}(s)N^*A(s^2+A)^{-1}AN\right\|_{\mathcal{L}\left(\mathcal{D}\left(\mathring{\mathbf{A}}^{\theta/4}\right)\right)} = 0. \tag{4.4}$$

Let $w \in \mathcal{D}(\mathring{\mathbf{A}}^{\theta/4})$. From (2.7), $N^*A \in \mathcal{L}(\mathcal{D}(A^{\frac{1}{4}+\epsilon}), L^2(\Gamma_0))$, where $\epsilon > 0$ is arbitrarily small, so there exists $M > 0$ such that

$$\begin{aligned}
\left\|N^*A(s^2+A)^{-1}ANw\right\|_{\mathcal{D}\left(\mathring{\mathbf{A}}^{\alpha/4}\right)'} &\le \left\|N^*A(s^2+A)^{-1}ANw\right\|_{L^2(\Gamma_0)} \\
&\le M\left\|(s^2+A)^{-1}ANw\right\|_{D\left(A^{\frac{1}{4}+\epsilon}\right)} \\
&= M\left\|A^{\frac{1}{4}+\epsilon}(s^2+A)^{-1}ANw\right\|_{L^2(\Omega)} \\
&= M\left\|A^{\frac{1}{2}+2\epsilon}(s^2+A)^{-1}A^{\frac{3}{4}-\epsilon}Nw\right\|_{L^2(\Omega)}. \qquad (4.5)
\end{aligned}$$

Using equation (5.15) on page 115 in [15] and (2.7), we see that

$$(4.5) \le \frac{M}{(1+s^2)^{\frac{1}{2}-2\epsilon}}\left\|A^{\frac{3}{4}-\epsilon}Nw\right\|_{L^2(\Omega)} \le \frac{M}{(1+s^2)^{\frac{1}{2}-2\epsilon}}\|w\|_{L^2(\Gamma_0)}$$

and therefore

$$\left\|N^*A(s^2+A)^{-1}ANw\right\|_{\mathcal{D}\left(\mathring{\mathbf{A}}^{\alpha/4}\right)'} \le \frac{M}{(1+s^2)^{\frac{1}{2}-2\epsilon}}\|w\|_{L^2(\Gamma_0)}. \tag{4.6}$$

Since $\theta > 0$, using (4.6) with (4.2) yields for large $s > 0$

$$\frac{s^2}{s+1}\left\|\mathbf{T}(s)N^*A(s^2+A)^{-1}ANw\right\|_{\mathcal{L}\left(\mathcal{D}\left(\mathring{\mathbf{A}}^{\theta/4}\right)\right)} = \mathcal{O}\left(\frac{1}{s^{1-\frac{1}{4}(\theta+\alpha)-4\epsilon}}\right).$$

Since $\theta+\alpha < 4$ by hypothesis, for $\epsilon > 0$ small enough the equation above implies that (4.4) is true. Hence

$$\begin{aligned}
&\left\|\left(I+\frac{s^2}{s+1}\mathbf{T}(s)N^*A(s^2+A)^{-1}AN\right)^{-1}\right\|_{\mathcal{L}\left(\mathcal{D}\left(\mathring{\mathbf{A}}^{\theta/4}\right)\right)} \\
&\qquad \le \frac{1}{1-\frac{s^2}{s+1}\left\|\mathbf{T}(s)N^*A(s^2+A)^{-1}ANw\right\|_{\mathcal{L}\left(\mathcal{D}\left(\mathring{\mathbf{A}}^{\theta/4}\right)\right)}}
\end{aligned}$$

for $s > 0$ large enough, which finishes the proof of Lemma 4.3. □

For $s > 0$, let

$$\mathbf{H}^1(s) := \left\{I+\frac{s^2}{s+1}\mathbf{T}(s)N^*A(s^2+A)^{-1}AN\right\}^{-1}\frac{s}{s+1}\mathbf{T}(s)B. \tag{4.7}$$

Now let α be as in (2.1), and recall that $B \in \mathcal{L}(U, H^{-\alpha}(\Gamma_0))$. Since $\mathcal{D}(\mathring{\mathbf{A}}^{\eta/4}) = H^{\eta}{}_0(\Gamma_0)$ for $\eta \in [0, 5/2)$, we see that $B \in \mathcal{L}(U, \mathcal{D}(\mathring{\mathbf{A}}^{\alpha/4})')$. Combining Lemmas 4.1 and 4.3 we immediately obtain the following result for the operator $\mathbf{H}^1(s)$.

Lemma 4.4. *$\mathbf{H}^1(s)$ is an element of $\mathcal{L}(U, H^{\theta}(\Gamma_0))$ for $s > 0$ large enough and $\theta \in [0, 4-\alpha)$ and for every θ there exists $M > 0$ such that*

$$\left\|\mathbf{H}^1(s)\right\|_{\mathcal{L}\left(U,\mathcal{D}\left(\mathring{\mathbf{A}}^{\theta/4}\right)\right)} \le \frac{M}{\left[1+\frac{s^2}{s+1}\right]^{1-\frac{1}{4}(\theta+\alpha)}}.$$

We now use these lemmas to prove the regularity of the system (2.2) with observation $\mathcal{C}_2 X(t)$. We first need to show that $\mathbf{H}^1$ is the transfer function from u to v_t.

Lemma 4.5. *Let the solution of* (2.2) *be of the form* (2.8) *with* $u \in L^2_\gamma(0,\infty;U)$ *for some* $\gamma \in \mathbb{R}$ *and with* $X(0) = 0$. *Then, for sufficiently large* $s > 0$,

$$\widehat{v_t}(s) = \mathbf{H}^1(s)\widehat{u}(s). \tag{4.8}$$

Proof. Applying the Laplace transform (in time t) to (2.15), (2.16) for $s \in \mathbb{C}_\beta$, we have, formally,

$$\begin{cases} s^2\widehat{z} = -A\widehat{z} + sAN\widehat{v} \iff \widehat{z} = s(s^2+A)^{-1}AN\widehat{v}, \\ \\ s^2\widehat{v} + \mathring{\mathbf{A}}\widehat{v} + s\mathring{\mathbf{A}}\widehat{v} = -sN^*A\widehat{z} + B\widehat{u}. \end{cases}$$

Substituting the first equation above into the second then yields

$$\left(s^2 + (s+1)\mathring{\mathbf{A}} + s^2N^*A(s^2+A)^{-1}AN\right)\widehat{v} = B\widehat{u}. \tag{4.9}$$

To deal with this equation, we will use the relation

$$(s^2 + (\tilde{A} + \tilde{B}))^{-1} = [I + (s^2+\tilde{A})^{-1}\tilde{B}]^{-1}(s^2+\tilde{A})^{-1}$$

for any operators $\tilde{A}$ and $\tilde{B}$ and scalars s for which all the inverses in the relation exist. Letting $\tilde{A} = (s+1)\mathring{\mathbf{A}}$ and $\tilde{B} = s^2N^*A(s^2+A)^{-1}AN$, we obtain formally

$$\begin{aligned} &\left(s^2 + (s+1)\mathring{\mathbf{A}} + s^2N^*A(s^2+A)^{-1}AN\right)^{-1} \\ &= \left(I + \left(s^2+(s+1)\mathring{\mathbf{A}}\right)^{-1} s^2N^*A(s^2+A)^{-1}AN\right)^{-1}\left(s^2+(s+1)\mathring{\mathbf{A}}\right)^{-1} \\ &= \left(I + \frac{s^2}{s+1}\mathbf{T}(s)N^*A(s^2+A)^{-1}AN\right)^{-1}\frac{1}{s+1}\mathbf{T}(s). \end{aligned} \tag{4.10}$$

Using Lemma 4.3, we see that (4.10) is indeed valid for sufficiently large real s. Using (4.10) in (4.9), we get

$$\widehat{v}(s) = \left(I + \frac{s^2}{s+1}\mathbf{T}(s)N^*A(s^2+A)^{-1}AN\right)^{-1}\frac{1}{s+1}\mathbf{T}(s)B\widehat{u}(s)$$

or

$$\widehat{v_t}(s) = s\hat{v}(s) = \mathbf{H}^1(s)\widehat{u}(s),$$

which finishes the lemma. □

It follows from Lemma 4.5 that for sufficiently large real s the transfer function from $u(t)$ to $y = \mathcal{C}_2X(t)$ is

$$\mathbf{H}_1(s) := C_{\xi_0}\mathbf{H}^1(s)$$

(where we recall that C_{ξ_0} is the point evaluation at $\xi_0 \in \Gamma_0$). Since the system (2.2) with this observation is well posed, there exists $\beta \in \mathbb{R}$ such that $\mathbf{H}_1 \in H^\infty_\beta(\mathcal{L}(U,Y))$; hence this transfer function can be extended analytically for $s \in \mathbb{C}_\beta$.

THEOREM 4.6. $\mathbf{H}_1(s)$ *is a regular transfer function with feedthrough* 0.

Proof. The theorem is proved if we can show that

$$\lim_{s\to\infty, s\in\mathbb{R}} \left\|C_{\xi_0}\mathbf{H}^1(s)\right\|_{\mathcal{L}(U,\mathbb{R})} = 0. \tag{4.11}$$

For $1/2 < \theta < 4-\alpha$, Lemma 4.4 and the Sobolev embedding theorem imply that there exists $M > 0$ such that

$$\left\|C_{\xi_0}\mathbf{H}^1(s)\right\|_{\mathcal{L}(U,\mathbb{R})} \le M\left\|\mathbf{H}^1(s)\right\|_{\mathcal{L}(U,H^\theta(\Gamma_0))} \le \frac{M}{\left[1+\frac{s^2}{s+1}\right]^{1-\frac{1}{4}(\theta+\alpha)}}.$$

Taking the limit as $s \to \infty$ along the real axis finishes the proof of Theorem 4.6. □

We now turn to the transfer function $\mathbf{H}_2(s)$ for (2.2) with the observation $\mathcal{C}_3X(t)$. In the previous section it is proved that this transfer function is well posed, so there exists $\beta \in \mathbb{R}$ such that $\mathbf{H}_2 \in H^\infty_\beta(\mathcal{L}(U,Y))$. We will show that this transfer function is also regular with feedthrough 0. Let

$$\mathbf{H}^2(s) = s(s^2+A)^{-1}AN\mathbf{H}^1(s).$$

We first show that $\mathbf{H}^2$ is the transfer function from u into the velocity z_t of the wave component.

LEMMA 4.7. *Let the solution of* (2.2) *be of the form* (2.8) *with* $u \in L^2_\beta(0,\infty;U)$ *and* $X(0) = 0$. *Then, for sufficiently large real* s,

$$\widehat{z_t}(s) = \mathbf{H}^2(s)\widehat{u}(s). \tag{4.12}$$

Proof. Applying the Laplace transform in time to (2.15), we obtain

$$s^2\widehat{z}(s) = -A\widehat{z}s + AN\widehat{v_t}(s)$$

or, equivalently, using (4.8),

$$\widehat{z} = (s^2+A)^{-1}AN\widehat{v_t} = (s^2+A)^{-1}AN\mathbf{H}^1(s)\widehat{u}(s);$$

(4.12) follows immediately from this. □

It follows from Lemma 4.7 that for sufficiently large real s the transfer function from $u(t)$ to $y(t) = \mathcal{C}_3X(t)$ is

$$\mathbf{H}_2(s) := C_{x_0}\mathbf{H}^2(s).$$

THEOREM 4.8. $\mathbf{H}_2(s)$ *is a regular transfer function with feedthrough* 0.

Proof. Since $\alpha/8 < 1/4$, we can choose $\varepsilon \in (0, 1/4-\alpha/8)$. Using the Sobolev embedding theorem and (2.3), there exist $M_1, M_2 > 0$ such that for $s > 0$,

$$\begin{aligned}\|\mathbf{H}_2(s)\|_{\mathcal{L}(U,\mathbb{R})} &= \left\|C_{x_0}\mathbf{H}^2(s)\right\|_{\mathcal{L}(U,\mathbb{R})} \le M_1\left\|\mathbf{H}^2(s)\right\|_{\mathcal{L}(U,H^{1+\varepsilon}(\Omega)\cap L^2(\Omega)/\mathbb{R})}\\ &= M_2\left\|\mathbf{H}^2(s)\right\|_{\mathcal{L}(U,\mathcal{D}(A^{(1+\varepsilon)/2}))}.\end{aligned} \tag{4.13}$$

The right side of (4.13) is equal to

$$\begin{aligned}&M_2s\left\|A^{\frac{1}{2}+\frac{\varepsilon}{2}}(s^2+A)^{-1}AN\mathbf{H}^1(s)\right\|_{\mathcal{L}(U,L^2(\Omega))}\\ &\quad= M_2s\left\|A^{\frac{3}{4}+\varepsilon}(s^2+A)^{-1}A^{\frac{3}{4}-\frac{\varepsilon}{2}}N\mathbf{H}^1(s)\right\|_{\mathcal{L}(U,L^2(\Omega))}\\ &\quad\le \frac{sM_2}{(1+s^2)^{\frac{1}{4}-\varepsilon}}\left\|A^{\frac{3}{4}-\frac{\varepsilon}{2}}N\mathbf{H}^1(s)\right\|_{\mathcal{L}(U,L^2(\Omega))} \le \frac{sM_3}{(1+s^2)^{\frac{1}{4}-\varepsilon}}\left\|\mathbf{H}^1(s)\right\|_{\mathcal{L}(U,L^2(\Gamma_0))}\end{aligned} \tag{4.14}$$

for some $M_3 > 0$, where we use equation (5.15) in [15] for the second-to-last inequality and (2.7) for the last inequality. Using Lemma 4.4 with $\theta = 0$, the right side of (4.14) is

$$\leq \frac{sM_3}{(1+s^2)^{\frac{1}{4}-\varepsilon}\left(1+\frac{s^2}{s+1}\right)^{1-\frac{\alpha}{4}}}.$$

This shows that

$$\|\mathbf{H}_2(s)\|_{\mathcal{L}(U,\mathbb{R})} = \mathcal{O}\left(s^{-\frac{1}{2}+2\varepsilon+\frac{\alpha}{4}}\right).$$

By our choice of ε, we see that $-1/2 + 2\varepsilon + \alpha/4 < 0$, so

$$\lim_{s\to\infty, s\in\mathbb{R}} \|\mathbf{H}_2(s)\|_{\mathcal{L}(U,\mathbb{R})} = 0.$$

5. Lack of robustness for dynamic stabilization of regular systems.

In this section we present results from the literature on lack of robustness with respect to delays. These results can be easily applied to the structural acoustics system in this paper, or in fact any regular system that satisfies (2.12) and has $R(s, A)$ analytic on $\mathbb{C}_0$ and continuous on $\overline{\mathbb{C}_0}$. We first need to discuss dynamic stabilization in this setting. In the definitions and theorems in this section we deal with generic regular systems, and not just the structural acoustics system in section 2.

Let X, X_c, U, and Y be Hilbert spaces, and recall that X_{-1} is defined as in (2.14). Let Σ_p be a regular system represented by

$$\dot{x}_p = Ax_p + Bu_p, \tag{5.1}$$

$$y_p = C_\Lambda x_p + Du_p, \tag{5.2}$$

where A generates a semigroup on X, $B : U \to X_{-1}$, $C : \mathcal{D}(A) \to Y$, and $D : U \to Y$, where we recall that C_Λ is the Lebesgue extension of C; see [25]. The subscript p stands for *plant*, the system we wish to stabilize. The transfer function for (5.1), (5.2) is $\mathbf{H}(s) := C_\Lambda R(s, A)B + D$.

Let Σ_c be a regular system represented by

$$\dot{x}_c = A^c x_c + B^c u_c, \tag{5.3}$$

$$y_c = C^c_\Lambda x_c + D^c u_c, \tag{5.4}$$

where A^c generates a semigroup on the Hilbert space X_c, $B^c : Y \to (X_c)_{-1}$, $C^c : \mathcal{D}(A^c) \to U$, and $D^c : Y \to U$. The subscript c stands for *controller*. The transfer function for (5.3), (5.4) is $\mathbf{H}^c(s) := C^c_\Lambda R(s, A^c)B^c + D^c$.

We can formally form a closed loop of Σ_p and Σ_c by letting

$$u_p = y_c + v_p, \tag{5.5}$$

$$u_c = y_p + v_c, \tag{5.6}$$

where we assume that the dimension of u_p is equal to the dimension of y_c and the dimension of u_c is equal to the dimension of y_p. This closed loop is illustrated in Figure 5.1 (when $\varepsilon = 0$).

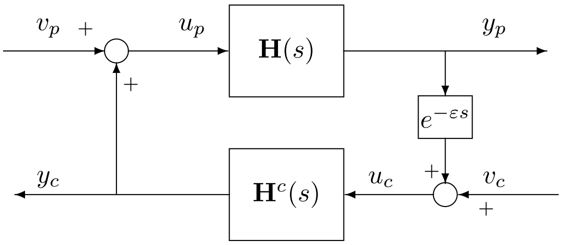


FIG. 5.1. *The closed loop of* Σ_p *and* Σ_c*.*

There are several difficulties associated with making the connections (5.5) and (5.6); these are covered in detail in Weiss and Curtain [27]. Briefly, we assume that $\mathbf{H}^c$ is an *admissible feedback transfer function* for $\mathbf{H}$, which means that $I-\mathbf{H}^c(s)\mathbf{H}(s)$ is invertible for all s in some right half plane, and its inverse is a well posed $\mathcal{L}(U)$-valued transfer function. This condition guarantees that the parallel connection of Σ_p and Σ_c is a well-posed linear system $\Sigma_{p,c}$ with state $[x_p, x_c]^T \in X \times X_c$, input $[v_p, v_c]^T \in U \times Y$, and output $[y_p, y_c]^T \in Y \times U$. If we assume in addition that $I - D^cD$ is invertible in $\mathcal{L}(U)$, then it is shown in [27] that $\Sigma_{p,c}$ is regular as well, with generating operator

$$A_{p,c} := \begin{bmatrix} A + B(I - D^cD)^{-1}D^cC_\Lambda & B(I - D^cD)^{-1}C^c_\Lambda \\ B^c(I - DD^c)^{-1}C_\Lambda & A^c + B^c(I - CC^c)^{-1}DC^c_\Lambda \end{bmatrix} \tag{5.7}$$

on its natural domain.

DEFINITION 5.1. *The regular system* Σ_c *is a regular stabilizing controller for* Σ_p *if* $\mathbf{H}^c$ *is an admissible feedback transfer function for* $\mathbf{H}$*,* $I-D^cD$ *is invertible in* $\mathcal{L}(U)$*, and* $A_{p,c}$ *generates an exponentially stable semigroup on* $X \times X_c$*.*

Remark 5.2. The above definition of a regular stabilizing controller includes static feedback as a special case, since a regular controller can be of the form $y = D^cu$. In this case $A_{p,c}$ generates an exponentially stable semigroup if and only if $A + B(I - D^cD)^{-1}D^cC_\Lambda$ does.

We now consider the effect of a time delay in the plant output. Let $\varepsilon > 0$ and suppose that (5.6) is replaced by

$$u_c(t) = y_p(t - \varepsilon) + v_c(t), \tag{5.8}$$

which is illustrated in Figure 5.1. The transfer function for this system is

$$\mathbf{F}^\varepsilon(\mathbf{H}, \mathbf{H}^c) := \begin{bmatrix} \mathbf{H}(I - e^{-\varepsilon\cdot}\mathbf{H}^c\mathbf{H})^{-1} & \mathbf{H}\mathbf{H}^c(I - e^{-\varepsilon\cdot}\mathbf{H}\mathbf{H}^c)^{-1} \\ e^{-\varepsilon\cdot}\mathbf{H}^c\mathbf{H}(I - e^{-\varepsilon\cdot}\mathbf{H}^c\mathbf{H})^{-1} & \mathbf{H}^c(I - e^{-\varepsilon\cdot}\mathbf{H}\mathbf{H}^c)^{-1} \end{bmatrix},$$

that is,

$$[\hat{v}_p, \hat{v}_c]^T = \mathbf{F}^\varepsilon[\hat{y}_p, \hat{y}_c]^T.$$

DEFINITION 5.3. *Suppose* $\mathbf{F}^0(\mathbf{H}, \mathbf{H}^c)$ *is input-output stable (see Definition* 2.5*). Then we say that the input-output stability of* $\mathbf{F}^0(\mathbf{H}, \mathbf{H}^c)$ *is robust with respect to delays if there exists* $\varepsilon^* > 0$ *such that* $\mathbf{F}^\varepsilon(\mathbf{H}, \mathbf{H}^c)$ *is input-output stable for all* $\varepsilon \in [0, \varepsilon^*)$*.*

Remark 5.4. In the case where the controller is static, it is easy to show that $\mathbf{F}^{\varepsilon}(\mathbf{H}, \mathbf{H}^c)$ is input-output stable if and only if $(I - e^{-\varepsilon \cdot}\mathbf{H}\mathbf{H}^c)^{-1}$ is input-output stable. Hence in this case the input-output stability of $\mathbf{F}^0(\mathbf{H}, \mathbf{H}^c)$ is robust with respect to delays if there exists $\varepsilon^* > 0$ such that $(I - e^{-\varepsilon \cdot}\mathbf{H}D^c)^{-1}$ is input-output stable for all $\varepsilon \in [0, \varepsilon^*)$, and we say that the input-output stability of $(I - \mathbf{H}D)^{-1}$ is robust with respect to delays.

Remark 5.5. If a regular system is exponentially stable (i.e., the semigroup is exponentially stable), then it is input-output stable; see Weiss [25]. Hence, if a regular transfer function is not stable, no regular realization of it is going to be exponentially stable. On the other hand, input-output stability does not necessarily imply exponential stability of the underlying semigroup generator, or even strong stability. Hence, even if we do not identify a state space realization for $\mathbf{F}^{\varepsilon}(\mathbf{H}, \mathbf{H}^c)$, a lack of robustness of the input-output stability of $\mathbf{F}^0(\mathbf{H}, \mathbf{H}^c)$ with respect to delays is quite strong.

We now present results about lack of robustness with respect to delays, and apply these results to our structural acoustics model. These results are simple corollaries of the following frequency domain theorem from [17].

THEOREM 5.6 ([17, Thm. 8.5]). *Suppose U and Y are finite dimensional,* $\mathbf{H}\mathbf{H}^c$ *is regular, and* $\mathbf{F}^0(\mathbf{H}, \mathbf{H}^c)$ *is input-output stable. If*

$$\limsup_{|s| \in \mathbb{C}_0, s \to \infty} \|\mathbf{H}(s)\|_{\mathcal{L}(U,Y)} = \infty, \tag{5.9}$$

then the input-output stability of $\mathbf{F}^0(\mathbf{H}, \mathbf{H}^c)$ *is not robust with respect to delays. In particular, there exist sequences* $\{\varepsilon_n\}$ *and* $\{p_n\}$ *with*

$$\varepsilon_n > 0, \qquad \varepsilon_n \to 0, \qquad p_n \in \mathbb{C}_0, \qquad |\mathrm{Im}(p_n)| \to \infty,$$

such that for any $n \in \mathbb{N}$, p_n *is a pole of* $\mathbf{H}\mathbf{H}^c(I - e^{-\varepsilon_n \cdot}\mathbf{H}\mathbf{H}^c)^{-1}$ *and hence of the overall closed-loop transfer function* $\mathbf{F}^e(\mathbf{H}, \mathbf{H}^c)$.

The following two corollaries will be useful for the system under consideration in this paper.

COROLLARY 5.7. *Suppose U and Y are finite dimensional,* $\Sigma_p = (A, B, C)$ *is a regular system,* $R(s, A)$ *is analytic on* $\mathbb{C}_0$ *and continuous on* $\overline{C_0}$, *and*

$$\limsup_{|s| \to \infty, s \in \mathbb{C}_0} \|R(s, A)\|_{\mathcal{L}(X)} = \infty. \tag{5.10}$$

If there exists a regular stabilizing controller Σ_c *for* Σ_p, *then the input-output stability of* $\mathbf{F}^0(\mathbf{H}, \mathbf{H}^c)$ *is not robust with respect to delays. In particular, the conclusions in Theorem* 5.6 *hold.*

Proof. In Weiss and Rebarber [28] it is shown that if there exists a regular stabilizing controller for Σ_p, then (5.9) holds if and only if (5.10) holds, so we can apply Theorem 5.6. □

As a special case of Corollary 5.7, we consider the static feedback case. We do not need the analyticity of $R(s, A)$ in $\mathbb{C}_0$ here, since we use the results in [21] instead of the results in [28].

COROLLARY 5.8. *Suppose U and Y are finite dimensional,* $\Sigma_p = (A, B, C)$ *is a regular system, and* (5.10) *holds. If there exists* $K \in \mathcal{L}(Y, U)$ *such that the closed-loop generator* $A + BKC_\Lambda$ *generates an exponentially stable semigroup, then the stability of* $(I - \mathbf{H}K)^{-1}$ *is not robust with respect to delays. In particular, the conclusions in Theorem* 5.6 *hold for the poles of* $(I + e^{\varepsilon_n \cdot}\mathbf{H}DK)^{-1}$.

We now turn to the structural acoustics model we have analyzed in previous sections. Since this system satisfies the conditions in Corollary 5.7, we can immediately apply Corollaries 5.7 and 5.8.

Case 1. Suppose the observation is given by

$$\mathcal{C}X(t) = [v(\alpha_1), \ldots , v(\alpha_j), v_t(\beta_1), \ldots , v_t(\beta_k)]^T \tag{5.11}$$

for $\{\alpha_i\}_{i=1}^{j}, \{\beta_i\}_{i=1}^{k} \subset \Gamma_0$. In particular, we assume in this case that the observation does not include any point evaluations of acoustic pressure. Then, by Proposition 2.3 and Theorem 4.6, the system (2.2) with this observation is a regular system, which can be represented by (2.20). Since this system satisfies (2.12), Corollary 5.7 applies to this system, and we obtain the following result.

THEOREM 5.9. *There is no regular dynamic controller using an observation of the form* (5.11) *that stabilizes the structural acoustics model robustly with respect to delays.*

COROLLARY 5.10. *Let $\mathcal{A}$ and $\mathcal{B}$ be as in sections* 2, 3, *and* 4, *and $\mathcal{C}$ be any observation of the form* (5.11). *Then there does not exist $K \in \mathcal{L}(Y, U)$ such that $\mathcal{A} + \mathcal{B}K\mathcal{C}$ generates a C_0-semigroup in a way that is robustly stable with respect to delays.*

Remark 5.11. Note that we have not made any claims about whether one can find an exponentially stabilizing regular dynamic controller in the case where the observation is only taken along the beam and not inside the cavity. We believe that it is unlikely that there is such a stabilizing dynamic controller, but proving this might be difficult: most such lack-of-stabilizability results require either the input operator or observation operator to be bounded, which is not the case here. However, if the control design anticipates small but uncertain delays, then this result shows that no dynamic stabilizer can do the job.

Case 2. Suppose the observation is given by

$$\mathcal{C}X(t) = [v(\alpha_1), \ldots , v(\alpha_j), v_t(\beta_1), \ldots , v_t(\beta_k), z_t(x_i), \ldots , z_t(x_l)]^T \tag{5.12}$$

for $\{\alpha_i\}_{i=1}^{j}, \{\beta_i\}_{i=1}^{k} \subset \Gamma_0$ and $\{x_i\}_{i=1}^{l} \subset \overline{\Omega}$.

As discussed in section 1, since this observation includes point evaluation of acoustic pressure, there is no natural and convenient state space realization; hence we state our results in terms of transfer functions and input-output stabilization. Since $\mathbf{H}(s) = \mathcal{C}_\Lambda R(s, A)\mathcal{B}$ and (5.10) holds, it is possible that $\mathbf{H}$ is unstable, but this is not guaranteed, since there might be many pole-zero cancellations. We will take the point of view that if $\mathbf{H}(s)$ is stable, then it can be robustly stabilized by the zero feedback, and focus our attention on the case where $\mathbf{H}$ is unstable.

THEOREM 5.12. *Suppose the transfer function* $\mathbf{H}$ *for the system* (2.2) *with observation* (5.12) *is not stable. Suppose further that* $\mathbf{H}^c$ *is well posed and is such that* $\mathbf{F}^0(\mathbf{H}, \mathbf{H}^c)$ *is stable (see Definition* 2.5*). Then the stability of* $\mathbf{F}^0(\mathbf{H}, \mathbf{H}^c)$ *is not robust with respect to delays. In particular, the conclusions in Theorem* 5.6 *hold.*

Proof. By Theorems 4.6 and 4.8, $\mathbf{H}$ is regular with feedthrough 0. Since $\mathbf{H}^c$ is well posed, $\mathbf{H}\mathbf{H}^c$ is regular with feedthrough 0. Since $\mathbf{H}$ is analytic in $\overline{\mathbb{C}_0}$, the only way it can be unstable is if (5.9) is satisfied. The conclusion follows from Theorem 5.6. □

Remark 5.13. The above results will still be true if $\mathcal{C}$ is replaced by any observation operator such that the open-loop system is regular. We focus on the particular observations (5.11) and (5.12) because point observation of the beam displacement, beam velocity, and acoustic pressure are well motivated physically.

6. Comparison with numerical results.

The feedback control we considered in the previous section (see Figure 5.1) is not the same as the control discussed in [8]. In [8] a noise term $\eta(t)$ and an exogenous forcing function $f(t)$ are included in the plant, and the feedback compensator includes a tracking term. In that paper numerical studies are done on a (necessarily finite dimensional) Galerkin approximation to the system and controller. Among these studies is a look at the effect of delays introduced in three places in the feedback loop for this finite dimensional model. In this section we give some results about the effect of these delays on the full infinite-dimensional model for this system. We then give some heuristic explanations for the numerical results obtained in [8] for the finite dimensional model, using a frequency domain analysis.

In [8] the fourth equation in (2.2) is replaced by

$$v_{tt}(\xi,t) = -\Delta^2 v(\xi,t) - \Delta^2 v_t(\xi,t) - z_t(\xi,t) + Bu(t) + b(\xi)f(t), \qquad \xi\in\Gamma_0, \quad t\in[0,\infty), \tag{6.1}$$

where b represents the spatial distribution of the forcing term. In [8] $b(\zeta)\equiv 1$, that is, the forcing term acts the same on all points of the active boundary. We further include the noise term $\mathcal{E}\eta(t)$ in the model, where $\eta(t)\in W$, a Hilbert space, and $\mathcal{E}\in\mathcal{L}(W,\mathcal{X})$. Therefore (2.13) is augmented to

$$\dot{X}(t) = \mathcal{A}X(t) + \mathcal{B}u(t) + \mathcal{B}_1 f(t) + \mathcal{E}\eta(t), \tag{6.2}$$

where

$$\mathcal{B}_1 = [0,0,0,b]^T \in \mathcal{X}.$$

Let $\mathcal{C}$ be of the form (5.11) or (5.12), let p be the dimension of the range of $\mathcal{C}$, and let $Y=\mathbb{R}^p$. The observation for (6.2) will be formally given by

$$y(t) = \mathcal{C}_\Lambda X(t) + E\eta(t), \tag{6.3}$$

where $E\in\mathcal{L}(W,Y)$. We do not worry here about the admissibility of $\mathcal{C}$ for $S(t)$, since we will be doing our robustness analysis in the frequency domain, and that analysis is justified by the work in sections 3 and 4. Also, we do not have a feedthrough term $Du(t)$ in the observation (6.3) because, as shown in section 4, the feedthrough is zero.

If the initial condition is $X(0)=X_0$, the system (6.2), (6.3) can be described in input-output terms as

$$\hat{y}(s) = \mathbf{H}(s)\hat{u}(s) + \mathbf{H}_1(s)\hat{f}(s) + \mathbf{H}_2(s)\hat{\eta}(s) + \hat{\nu}(s); \tag{6.4}$$

here ν is the observation of the system, depending on the initial data X_0, but with $u\equiv 0$, $\eta\equiv 0$, and $f\equiv 0$. $\mathbf{H}$ is the transfer function from u to y (analyzed in detail in sections 3 and 4), $\mathbf{H}_1$ is the transfer function from f to y, and $\mathbf{H}_2$ is the transfer function from η to y. If the observation $\mathcal{C}$ does not contain point observations of acoustic pressure, we can write

$$\hat{\nu}(s) = \mathcal{C}_\Lambda R(s,\mathcal{A})X_0, \qquad \mathbf{H}(s) = \mathcal{C}_\Lambda R(s,\mathcal{A})\mathcal{B}, \tag{6.5}$$

$$\mathbf{H}_1(s) = \mathcal{C}_\Lambda R(s,\mathcal{A})\mathcal{B}_1, \qquad \mathbf{H}_2(s) = \mathcal{C}_\Lambda R(s,\mathcal{A})\mathcal{E} + E. \tag{6.6}$$

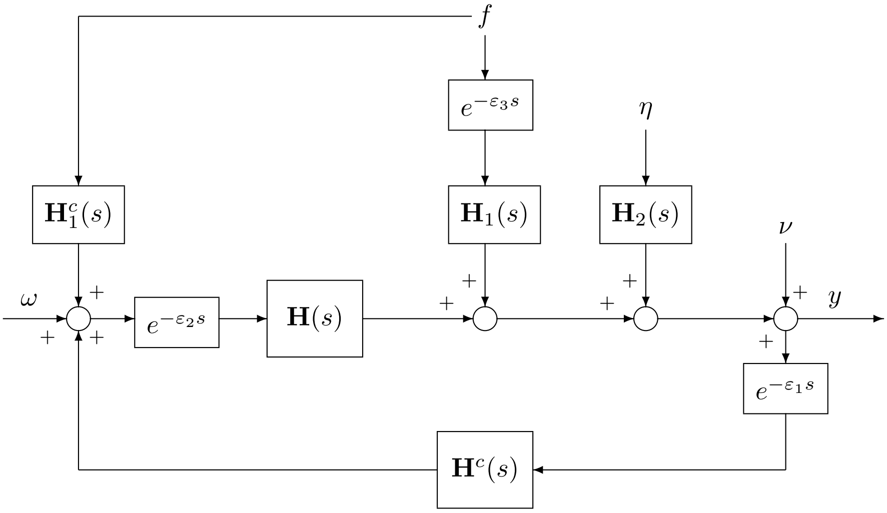


FIG. 6.1. *Controller with tracking and delays.*

If $\mathcal{C}$ does contain point observations of acoustic pressure, results in sections 3 and 4 show that $\mathbf{H}_1$ is regular with feedthrough 0. In this case, since $\mathcal{C}$ is not even bounded on $\mathcal{D}(\mathcal{A})$, the other transfer functions in (6.5), (6.6) are regular only if X_0 and $\mathcal{E}$ are sufficiently smooth. We will assume in this section that this is the case.

We will not be concerned about the exact nature of the feedback here, but we will insist that the controller be of the following form in the frequency domain:

$$\hat{u}(s) = \mathbf{H}^c(s)\hat{y}(s) + \mathbf{H}_1^c(s)\hat{f}(s) + \hat{\omega}(s), \tag{6.7}$$

where $\mathbf{H}^c$ and $\mathbf{H}_1^c$ are well-posed transfer functions, $\mathbf{H}_1^c$ contains any tracking terms, and $\hat{\omega}$ contains initial state information for the compensator. The feedback control in [8] is of this form if the solutions to two algebraic Riccati equations exist and are of sufficient smoothness; it is also of this form if it is a finite dimensional controller computed using the Galerkin approximation to the system. The closed-loop system is shown in Figure 6.1, ignoring the delay blocks.

Solving for $\hat{y}$ in (6.4) and (6.7), we obtain

(6.8)
$$\hat{y} = (I - \mathbf{H}\mathbf{H}^c)^{-1}\mathbf{H}_2\hat{\eta} + (I - \mathbf{H}\mathbf{H}^c)^{-1}(\mathbf{H}_1 + \mathbf{H}\mathbf{H}_1^c)\hat{f} + (I - \mathbf{H}\mathbf{H}^c)^{-1}(\mathbf{H}\hat{w} + \hat{\nu}).$$

An obvious design goal for the compensator is to make the transfer function from ω to y and the transfer function from ν to y stable, so we assume that $\mathbf{H}^c$ has been chosen so that

$$(I - \mathbf{H}\mathbf{H}^c)^{-1} \in H^\infty \qquad \text{and} \qquad (I - \mathbf{H}\mathbf{H}^c)^{-1}\mathbf{H} \in H^\infty. \tag{6.9}$$

As in [8], we consider delays in three places in the feedback loop:
(1) A delay $\varepsilon_1 \geq 0$ in the plant output $y(t)$;
(2) A delay $\varepsilon_2 \geq 0$ in the input voltage $u(t)$; and
(3) A delay $\varepsilon_3 \geq 0$ in the forcing signal $f(t)$.

These three delays are shown in Figure 6.1. When these delays are taken into account, (6.8) becomes

$$\hat{y} = (I - e^{-(\varepsilon_1+\varepsilon_2)\cdot}\mathbf{H}\mathbf{H}^c)^{-1}\mathbf{H}_2\hat{\eta} + (I - e^{-(\varepsilon_1+\varepsilon_2)\cdot}\mathbf{H}\mathbf{H}^c)^{-1}(\mathbf{H}\hat{w} + \hat{\nu})$$

$$+ (I - e^{-(\varepsilon_1+\varepsilon_2)\cdot}\mathbf{H}\mathbf{H}^c)^{-1}(\varepsilon^{-\varepsilon_3\cdot}\mathbf{H}_1 + \varepsilon^{-\varepsilon_2\cdot}\mathbf{H}\mathbf{H}_1^c)\hat{f}. \tag{6.10}$$

In the following results, when we write ε_n, we assume that it is not ε_1, ε_2, or ε_3 as defined above.

PROPOSITION 6.1. *Suppose* (6.9) *holds and* $\mathbf{H}$ *is not stable. Then there exist sequences* $\{\varepsilon_n\}$ *and* $\{p_n\}$ *with*

$$\varepsilon_n > 0, \qquad \varepsilon_n \to 0, \qquad p_n \in \mathbb{C}_0, \qquad |\mathrm{Im}(p_n)| \to \infty$$

such that for any $n \in \mathbb{N}$, p_n *is a pole of* $(I - e^{-(\varepsilon_1+\varepsilon_2)\cdot}\mathbf{H}\mathbf{H}^c)^{-1}$ *when* $\varepsilon_1 + \varepsilon_2 = \varepsilon_n$.

Proof. In sections 3 and 4 we showed that $\mathbf{H}$ is regular with feedthrough 0. Since $\mathbf{H}^c$ is well posed, $\mathbf{H}\mathbf{H}^c$ is regular with feedthrough 0. We first need to show that

$$\limsup_{s\in\mathbb{C}_0, |s|\to\infty} \|\mathbf{H}(s)\mathbf{H}^c(s)\|_{\mathcal{L}(Y)} = \infty. \tag{6.11}$$

Suppose (6.11) is not true. Then there exist $R, M > 0$ such that

$$\|I - \mathbf{H}(s)\mathbf{H}^c(s)\|_{\mathcal{L}(Y)} \le M \quad \text{for } s \in \mathbb{C}_0 \cap \{|s| > R\}. \tag{6.12}$$

Note that

$$\mathbf{H} = (I - \mathbf{H}\mathbf{H}^c)(I - \mathbf{H}\mathbf{H}^c)^{-1}\mathbf{H},$$

so (6.12) and the second equation in (6.9) imply that $\|\mathbf{H}(s)\|_{\mathcal{L}(U,Y)}$ is bounded for $s \in \mathbb{C}_0 \cap \{|s| > R\}$, which contradicts the hypothesis that $\mathbf{H}$ is unstable and the fact that $\mathbf{H}$ is analytic in $\mathbb{C}_0$. Hence (6.11) must be true. Using Lemma 6.3 in [17], we get

$$\limsup_{s\in\mathbb{C}_0, |s|\to\infty} r(\mathbf{H}(s)\mathbf{H}^c(s)) = \infty,$$

where r denotes the spectral radius. Theorem 5.3 in [17] now implies that there exist sequences $\{\varepsilon_n\}$ and $\{p_n\}$ with

$$\varepsilon_n > 0, \qquad \varepsilon_n \to 0, \qquad p_n \in \mathbb{C}_0, \qquad |\mathrm{Im}(p_n)| \to \infty,$$

such that for any $n \in \mathbb{N}$, p_n is a pole of $(I - e^{-(\varepsilon_1+\varepsilon_2)\cdot}\mathbf{H}\mathbf{H}^c)^{-1}\mathbf{H}\mathbf{H}^c$ when $\varepsilon_1 + \varepsilon_2 = \varepsilon_n$. The proposition now follows from the fact that

$$(I - e^{-(\varepsilon_1+\varepsilon_2)\cdot}\mathbf{H}\mathbf{H}^c)^{-1} = e^{-(\varepsilon_1+\varepsilon_2)\cdot}(I - e^{-(\varepsilon_1+\varepsilon_2)\cdot}\mathbf{H}\mathbf{H}^c)^{-1}\mathbf{H}\mathbf{H}^c + I. \qquad \square$$

In [8], numerical tests imply that the stability of the closed-loop system can handle small delays ε_3 in the forcing function $f(t)$, can tolerate very small delays ε_1 in $y(t)$, and becomes unstable for any delays ε_2 in $u(t)$. Direct comparison of the lack-of-robustness results in Proposition 6.1 with the results in [8] is difficult, since our results are for the full infinite dimensional system and controller, and the results in [8] are for a finite dimensional approximation of the system and controller, and also because the results in [8] are for only one initial state. Nonetheless, we can use our framework to give heuristic explanations for the results from [8].

Since in the envisioned applications $f(t)$ will be periodic, we cannot necessarily expect that $y(t)$ defined by (6.10) is in $L^2[0,\infty;Y]$, even if the transfer functions are stable. However, we will assume that the tracking component of the controller is chosen to minimize the effect of f on y. In particular, we assume that $\mathbf{H}_1^c$, which contains the tracking component of the controller, is chosen so that the term $(\mathbf{H}_1 + \mathbf{H}\mathbf{H}_1^c)$ in (6.8) has relatively small H^∞-norm. First suppose that $\varepsilon_3 = 0$. Both ε_1 and ε_2 effect $(I - e^{-(\varepsilon_1+\varepsilon_2)\cdot}\mathbf{H}\mathbf{H}^c)^{-1}$, and ε_2 also effects the term $(\mathbf{H}_1 + e^{-\varepsilon_2\cdot}\mathbf{H}\mathbf{H}_1^c)$ in (6.10). Since $e^{-\varepsilon_2 s}$ can be far from 1 when s is on the imaginary axis and $|s|$ is large, small ε_2 could easily counteract the minimizing effect of $\mathbf{H}_1^c$ on $\|\mathbf{H}_1(s) + e^{-\varepsilon_2 s}\mathbf{H}(s)\mathbf{H}_1^c(s)\|$. Thus, ε_2 should have a stronger destabilizing effect on the closed-loop system than ε_1, which has exactly the same effect on $(I - e^{-(\varepsilon_1+\varepsilon_2)\cdot}\mathbf{H}\mathbf{H}^c)^{-1}$ as ε_2 does.

Now suppose that $\varepsilon_1 = \varepsilon_2 = 0$. The delay ε_3 only effects the term $(e^{-\varepsilon_3\cdot}\mathbf{H}_1 + \mathbf{H}\mathbf{H}_1^c)$ in (6.10), and can easily counteract the effect of $\mathbf{H}_1^c$ on $\|(e^{-\varepsilon_3 s}\mathbf{H}_1(s) + \mathbf{H}(s)\mathbf{H}_1^c(s))\|$ for s on the imaginary axis. However, since $\mathcal{B}_1$ is quite smooth—for instance, in [8] it is $[0,0,0,1]^T$—it is possible that $\mathbf{H}_1(s) = \mathcal{C}R(s,\mathcal{A})\mathcal{B}_1$ is not large when s is on the imaginary axis and $|s|$ is large. This would mitigate the effect of ε_3 on $(e^{-\varepsilon_3\cdot}\mathbf{H}_1 + \mathbf{H}\mathbf{H}_1^c)$.

7. Conclusions. To obtain the results in this paper, we used a combination of techniques. As we have seen in sections 5 and 6, the effect of small delays on feedback control is best handled using a frequency domain approach. In order to use this approach we first must show that the structural acoustics system under consideration is in the class of *regular* systems; we can then readily apply results in [17] about delay robustness for regular systems. It is often a challenge to show that a given system is regular. For the structural acoustic model in a rectangular cavity, to prove regularity we used sharp PDE estimates from [1, 2, 3, 4], C_0-semigroup results from [15], and delicate estimates involving harmonic analysis; for nonrectangular cavities, microlocal analysis is required (see [5]).

Since there is extensive literature on control design for regular systems, the implications of regularity go beyond the study of robustness with respect to delays. For instance, it might be possible to use the regular systems framework to study the effect of additive perturbations on output feedback stabilization [19], to design adaptive control [18], to give Youla parametrizations for stabilizing controllers [27], or to study the relationship between input-output stabilization and exponential stabilization [21, 22].

Acknowledgment. The authors would like to thank Hartmut Logemann for assistance in the proof of Proposition 6.1.

REFERENCES

[1] G. Avalos, *Well-Posedness for a Coupled Hyperbolic/Parabolic System Seen in Structural Acoustics*, preprint 1346, Institute for Mathematics and Its Applications, Minneapolis, MN, 1995.

[2] G. Avalos and I. Lasiecka, *A differential Riccati equation for the active control of a problem in structural acoustics*, J. Optim. Theory Appl., 91 (1996), pp. 695–728.

[3] G. Avalos and I Lasiecka, *The strong stability of a semigroup arising from a coupled hyperbolic/parabolic system*, Semigroup Forum, 57 (1998), pp. 278–292.

[4] G. Avalos, *Sharp regularity estimates for solutions of the wave equation and their traces with prescribed Neumann data*, Appl. Math. Optim., 35 (1997), pp. 203–219.

[5] G. Avalos, I. Lasiecka, and R. Rebarber, *Well-posedness of a structural acoustics control model with point observation*, preprint.

[6] H.T. Banks, M.A. Demetriou, and R.C. Smith, *Robust output feedback control in a 2-D structural acoustic model with piezoceramic actuators*, in Proceedings of the 1994 International Conference on Intelligent Materials, Blacksburg, VA, 1994, C.A. Rogers, ed., Technomics, Lancaster, PA, pp. 109–127.

[7] H.T. Banks, M.A. Demetriou and R.C. Smith, *An H^∞ minmax periodic control in a 2-D structural acoustics model with piezoceramic actuators*, IEEE Trans. Automat. Control, 41 (1996), pp. 943–959.

[8] H.T. Banks, M.A. Demitriou, and R.C. Smith, *Robustness studies for H^∞ feedback control in a structural acoustics model with periodic excitation*, Internat. J. Robust and Nonlinear Control, 6 (1996), pp. 453–478.

[9] H.T. Banks, W. Fang, R.J. Silcox, and R.C. Smith, *Approximation methods for control of structural acoustics models with piezoceramic actuators*, J. Intelligent Material Systems Structures, 4 (1993), pp. 98–116.

[10] H.T. Banks and R.C. Smith, *Models for control in smart material structures*, Identification and Control in Systems Governed by Partial Differential Equations, SIAM, Philadelphia, PA, 1993, pp. 26–44.

[11] H.T. Banks and R.C. Smith, *Well-posedness of a model for structural acoustic coupling in a cavity enclosed by a thin cylindrical shell*, J. Math. Anal. Appl., 191 (1995), pp. 1–25.

[12] H.T. Banks and R.C. Smith, *Feedback control of noise in a 2-D nonlinear structural acoustics model*, Discrete Contin. Dynam. Systems, 1 (1995), pp. 119–149.

[13] G. Doetsch, *Introduction to the Theory and Application of the Laplace Transform*, Springer-Verlag, New York, 1974.

[14] P. Grisvard, *Caracterization de quelques espaces d'interpolation,* Arch. Rational Mech. Anal., 25 (1967), pp. 40–63.

[15] S.G. Krein, *Linear Differential Equations in Banach Space*, Transl. Math. Monogr., 29, AMS, Providence, RI, 1971. Translated from the Russian.

[16] I. Lasiecka and R. Triggiani, *A cosine operator approach to modeling $L_2(0,T;L_2(\Gamma))$–boundary input hyperbolic operators*, Appl. Math. Optim., 7 (1981), pp. 35–93.

[17] H. Logemann, R. Rebarber, and G. Weiss, *Conditions for robustness and nonrobustness of the stability of feedback systems with respect to small delays in the feedback loop*, SIAM J. Control Optim., 34 (1996), pp. 572–600.

[18] H. Logemann and S. Townley, *Low-gain control of uncertain regular linear systems*, SIAM J. Control Optim., 35 (1997), pp. 78–116.

[19] A.J. Pritchard and S. Townley, *A Real Stability Radii for Infinite-Dimensional Systems*, in Proc. MTNS89, Amsterdam, 1989, pp. 635–646.

[20] J. Prüss, *On the spectrum of C_0-semigroups*, Trans. Amer. Math. Soc., 24 (1984), pp. 847–857.

[21] R. Rebarber, *Conditions for the equivalence of internal and external stability for distributed parameter systems*, IEEE Trans. Automat. Control, 38 (1993) No. 6, pp. 994–998.

[22] R. Rebarber, *Exponential stability of coupled beams with dissipative joints: A frequency domain approach*, SIAM J. Control Optim., 33 (1995), pp. 1–28.

[23] R. Rebarber and S. Townley, *Robustness and continuity of the spectrum for uncertain distributed parameter systems*, Automatica, 30 (1995) pp. 1533–1546.

[24] D. Salamon, *Realization theory in Hilbert space*, Math. Systems Theory, 21 (1989), pp. 147–164.

[25] G. Weiss, *Transfer functions of regular linear systems, Part* I*: Characterizations of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.

[26] G. Weiss, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.

[27] G. Weiss and R.F. Curtain, *Dynamic stabilization of regular linear systems*, IEEE Trans. Automat. Control, 42 (1997), pp. 4–21.

[28] G. Weiss and R. Rebarber, *Optimizability and estimatability for infinite-dimensional linear systems*, preprint.

[29] M. Weiss and G. Weiss, *The spectral factorization approach to the LQ problem for regular linear systems*, Proceedings of the 1995 European Control Conference, Rome, Italy.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 5, pp. 1419–1455

EQUIVALENT CONDITIONS FOR STABILIZABILITY OF INFINITE-DIMENSIONAL SYSTEMS WITH ADMISSIBLE CONTROL OPERATORS[*]

BIRGIT JACOB[†] AND HANS ZWART[‡]

Abstract. In this paper we study the optimizability of infinite-dimensional systems with admissible control operators. We show that under a weak condition such a system is optimizable if and only if the system can be split into an exponentially stable subsystem and an unstable subsystem that is exactly controllable in finite time. The state space of the unstable subsystem equals the span of all unstable (generalized) eigenvectors of the original system. This subsystem can be infinite-dimensional. Furthermore, the unstable poles satisfy a summability condition. The state space of the exponentially stable subsystem is given by all vectors for which the action of the original C_0-semigroup is stable.

Key words. infinite-dimensional systems, stabilizability, optimizability, controllability

AMS subject classifications. 93C25, 93D15, 93B05

PII. S036301299833344X

1. Introduction.

This paper is concerned with the stabilization of infinite-dimensional systems with a finite-dimensional input space described by the following abstract differential equation:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \quad t \geq 0 \\ x(0) &= x_0, \end{aligned} \tag{1}$$

where A is the infinitesimal generator of the C_0-semigroup $T(t)$ on the separable Hilbert space H and B is the control operator.

Since the early days of infinite-dimensional systems theory, researchers have tried to characterize conditions under which (1) is stabilizable; see, e.g., the survey articles by Pritchard and Zabczyk [23] and by Russell [29]. In the early 1970s Datko [4] found a characterization in terms of the cost functional

$$J(x_0, u) := \int_0^\infty \|x(t)\|^2 + \|u(t)\|^2 \, dt. \tag{2}$$

He proved that if B is a bounded operator, then there exists a bounded feedback F such that $A + BF$ generates an exponentially stable semigroup if and only if for every $x_0 \in H$ there exists an input u such that $J(x_0, u) < \infty$.

It was only in the mid-1980s that for the class of bounded control operators other necessary and sufficient conditions were found; see Desch and Schappacher [5], Jacobson and Nett [17], and Nefedov and Sholokhovich [20]. Independently of each other, they showed the following result. Here $\mathcal{L}(X_1, X_2)$ denotes the class of linear, bounded operators from X_1 to X_2.

[*] Received by the editors January 30, 1998; accepted for publication (in revised form) October 22, 1998; published electronically July 27, 1999. This research was supported by the Human Capital and Mobility program (project number CHRX-CT 93-0402) and Volkswagen-Stiftung (RiP-program at Oberwolfach).

http://www.siam.org/journals/sicon/37-5/33344.html

[†] School of Mathematics, University of Leeds, Leeds LS2 9JT, UK (birgit@amsta.leeds.ac.uk).

[‡] Faculty of Mathematical Sciences, University of Twente, P. O. Box 217, 7500 AE Enschede, The Netherlands (H.J.Zwart@math.utwente.nl).

THEOREM 1.1. *Consider the system* (1) *with* $B \in \mathcal{L}(\mathbb{C}^m, H)$. *Then the following conditions are equivalent.*

1. *There exists an* $F \in \mathcal{L}(H, \mathbb{C}^m)$ *such that* $A + BF$ *generates an exponentially stable* C_0*-semigroup.*
2. *It is possible to split the state space as* $H = H_u \oplus H_s$ *such that* $T(t)|_{H_s}$ *is exponentially stable,* H_u *is finite-dimensional and the restriction of* (1) *to* H_u *is exactly controllable in finite time.*

Looking at the proof of this theorem, we observe that from the fact that B and F are bounded, finite-rank operators it was obtained that there are only finitely many unstable eigenvalues. Furthermore, H_u is the span over the eigenspaces corresponding to these eigenvalues.

In the eighties the first results for unbounded control operators were found. In general, the class of unbounded control operators consists of those operators B for which

$$\int_0^{t_1} T(t_1 - \rho) B u(\rho)\, d\rho$$

defines a bounded linear operator from $\boldsymbol{L}_2(0, t_1; \mathbb{C}^m)$ to H. Following Weiss [32], we call B an *admissible control operator for* $T(t)$. If B is such an admissible control operator for $T(t)$, then for every x_0 and for every $u \in \boldsymbol{L}_2^{loc}(0, \infty; \mathbb{C}^m)$, (1) has a unique solution $x(\cdot)$ which lies in H and is continuous; see Weiss [33].

Flandoli, Lasiecka, and Triggiani [8] extended the result of Datko to the class of admissible control operators for $T(t)$. Among others, they showed the following.

THEOREM 1.2. *Consider the system* (1), *where the operator* B *is an admissible control operator for* $T(t)$. *Then the following conditions are equivalent.*

1. *System* (1) *is* optimizable*; i.e., for every element* $x_0 \in H$ *there exists an input* $u \in \boldsymbol{L}_2(0, \infty; \mathbb{C}^m)$ *such that the cost functional* $J(x_0, u)$*, given by* (2), *is finite.*
2. *There exist an exponentially stable semigroup* $T_F(t)$ *with infinitesimal generator* $A_F : D(A_F) \to H$, *and an* $F \in \mathcal{L}(D(A_F), \mathbb{C}^m)$ *such that*
   (a) $FT_F(\cdot) : D(A_F) \to \boldsymbol{L}_2(0, \infty; \mathbb{C}^m)$ *extends to a bounded linear operator from* H *to* $\boldsymbol{L}_2(0, \infty; \mathbb{C}^m)$*;*
   (b) *For every* $t > 0$ *and* $x_0 \in H$ *there holds*

$$T_F(t)x_0 = T(t)x_0 + \int_0^t T(t - \rho) BFT_F(\rho) x_0\, d\rho. \tag{3}$$

The extension of Theorem 1.1 for the case that B is an admissible control operator for $T(t)$ and F is a bounded operator can be found in Rebarber [25]. The aim of this paper is to show that Theorem 1.1 can be extended to admissible control operators for $T(t)$ and to feedback operators satisfying condition 2 of Theorem 1.2. The implication part 2 to part 1 is relatively easy, so we will concentrate on the other direction.

Before we summarize our results, we have to introduce some notation. We say that the generator A of a C_0-semigroup satisfies the *spectrum decomposition assumption (SDA) at* g, $g \in \mathbb{R}$, if there exist real numbers $g_1 < g < g_2$ such that there is no spectrum of A with real part between g_1 and g_2. In particular, we say that A satisfies the SDA if A satisfies SDA at 0. Hence the SDA states that the spectrum can be split into a "stable" and an "unstable" part. The main result that we prove in this paper is the following.

THEOREM 1.3. *Consider the system* (1), *where the operator* B *is an admissible finite-rank control operator for* $T(t)$. *If the system is optimizable and* A *satisfies the SDA, then it is possible to split the state space as* $H = H_s \oplus H_u$, *where*

$$\begin{aligned} H_s &= \{x_0 \in H \mid T(t)x_0 \in \boldsymbol{L}_2(0,\infty;H)\}, \\ H_u &= \overline{\operatorname*{span}_{\lambda\in\sigma(A) with\ Re(\lambda)\geq 0} P_\lambda H}. \end{aligned}$$

Here $P_\lambda H$ *denotes the spectral subspace corresponding to the eigenvalue* λ.

Moreover, H_s *as well as* H_u *is a* $T(t)$*-invariant subspace,* $T(t)|_{H_s}$ *is exponentially stable,* $T(t)|_{H_u}$ *is a group and the restriction of* (1) *to* H_u *is exactly controllable in finite time.*

A result similar to those of Theorem 1.3 holds if the system $(A + gI, B)$ is optimizable and if $A + gI$ satisfies the SDA. So one sees that the only difference between bounded and unbounded control operators lies in the fact that H_u may be infinite-dimensional and that we have to assume the SDA.

One may interpret this theorem as follows. If the system is optimizable and the spectrum can be split into a stable and unstable part, then the state space can be split in the corresponding way. Note that for general infinite-dimensional systems such a splitting of the state space is not always possible, even if one can split the spectrum. There exist infinitesimal generators whose spectrum lies in the left-half plane and is bounded away from the imaginary axis, but they generate an unstable semigroup; see, e.g., Zabczyk [39].

It may seem that the spectrum decomposition assumption is very strong. However, in Rebarber and Zwart [28, Theorem 2.11] the following result can be found.

THEOREM 1.4. *Assume that system* (1) *is optimizable where the operator* B *is an admissible finite-rank control operator for* $T(t)$. *Then there exists an* $\varepsilon > 0$ *such that all elements in the spectrum of* A *with real part larger than* $-\varepsilon$ *are eigenvalues of* A *with finite algebraic multiplicity and all accumulation points of the spectrum of* A *have real part less than or equal to* $-\varepsilon$.

Hence this result gives a necessary condition for the system (1) to be optimizable. Looking now at the class of systems that satisfies this necessary condition, i.e., in the closed right-half plane there is only point spectrum, it is much harder to find an example that does not satisfy the SDA. All delay and partial differential equations that satisfy the necessary condition on its spectrum also satisfy the SDA at $-\varepsilon$ for some $\varepsilon > 0$.

We construct an example that is optimizable, but it is not possible to split the state space in a direct sum as given in Theorem 1.3. The generator does not satisfy the SDA at any negative number.

If A is a Riesz-spectral operator, then we are able to prove Theorem 1.3 without assuming that A satisfies SDA. The SDA at $-\varepsilon$, for some $\varepsilon > 0$, will automatically be satisfied if the system is stabilizable by a bounded feedback BF, since then there can only be finitely many eigenvalues with real part larger than $-\varepsilon$ for some $\varepsilon > 0$.

From Rebarber and Zwart [28] we have that the unstable part of the spectrum is pure point spectrum; see Theorem 1.4. Here we show an extra property concerning the distribution of this point spectrum.

THEOREM 1.5. *Assume that system* (1) *is optimizable where the operator* B *is an admissible finite-rank control operator for* $T(t)$. *Let* $\{\lambda_n\}$ *denote the set of eigenvalues with real part greater than* 0 *and let* $m_a(\lambda_n, A)$ *denote the algebraic multiplicity of* λ_n.

Then we have

$$\sum_{n\in\mathbb{N}} \frac{m_a(\lambda_n, A)}{|\lambda_n|^2} < \infty.$$

From Theorem 1.3 we obtain some consequences as easy corollaries. For instance, if the spectrum of A lies in the left-half plane and is bounded away from the imaginary axis, but A generates an unstable semigroup, then it is never optimizable, for any admissible finite-rank control operator for $T(t)$. Thus these systems will never be exactly controllable in finite time either. Examples of such generators can be found in, e.g., Zabczyk [39] and Greiner, Voigt, and Wolff [11]. This result can be found in [28] for the case of one-dimensional, not necessarily admissible, input operators.

Another consequence is the following. Suppose that A generates a C_0-group. If (1) is exactly controllable in finite time for some B where B is an admissible finite-rank control operator for $T(t)$, then the (generalized) eigenfunctions of A must span H. This follows easily from Theorem 1.3 by using the fact that $A + \gamma I$ is exactly controllable in finite time for every $\gamma \in \mathbb{R}$. Now choose γ so large that the corresponding "stable" subspace H_s is empty, which is possible since A generates a C_0-group.

A third consequence is the following. Suppose that A generates a completely unstable C_0-semigroup, that $\Sigma(A, B)$ is optimizable, and that B is an admissible finite-rank control operator for $T(t)$. Then A generates a C_0-group, and $\Sigma(A, B)$ is exactly controllable in finite time. This is an extension of the result found by Russell in [29]. Reversing the argument gives that the right-shift semigroup on $\boldsymbol{L}_2(0, \infty)$ will never be optimizable nor exactly controllable in finite time by a finite-rank admissible control operator for $T(t)$. The reason is that the right-shift semigroup is completely unstable, but it is not a C_0-group.

The organization of the paper is as follows. In the next two sections we introduce the necessary background and notation. In section 4, we show that if the state space can be split into two $T(t)$-invariant subspaces, then we can write the system (1) into two corresponding subsystems. Theorem 1.3 will be proved in section 5, apart from the characterization of H_s and H_u. That will be done in section 7. Theorem 1.5 will be the subject of section 6. Finally, in section 8 we present an example that is optimizable, but it is not possible to split the state space of this example in a direct sum as given in Theorem 1.3.

2. System description. In this section, we describe the general class of systems discussed in this paper. First we need to introduce some notation.

$\mathbb{C}_\alpha^+$	$\{z \in \mathbb{C} : \mathrm{Re}(z) > \alpha\}$, $\alpha \in \mathbb{R}$,
$\mathbb{C}_\alpha^-$	$\{z \in \mathbb{C} : \mathrm{Re}(z) < \alpha\}$, $\alpha \in \mathbb{R}$,
H	separable, complex Hilbert space,
$\sigma(A)$	spectrum of A,
$\sigma_\alpha^+(A)$	$\sigma(A) \cap \mathbb{C}_\alpha^+$,
$\sigma_p(A)$	point spectrum of A,
$\rho(A)$	resolvent set of A,
$\rho_\infty(A)$	largest connected subset of $\rho(A)$ that contains an interval of the form $[r, \infty)$,
$N(A)$	$\{x \in D(A) \mid Ax = 0\}$,

$\mathrm{Im}(A)$	$\{x \in H \mid \exists y \in D(A) : Ay = x\}$,
$\boldsymbol{H}_\infty(\mathbb{C}_\alpha^+; H)$	set of holomorphic and bounded functions from $\mathbb{C}_\alpha^+$ to H,
$\boldsymbol{H}_2(\mathbb{C}_\alpha^+; H)$	set of holomorphic functions $f : \mathbb{C}_\alpha^+ \to H$ with $\sup_{x>\alpha} \left(\int_{-\infty}^{\infty} \|f(x+iy)\|^2 dy \right)^{1/2} < \infty.$

We deal with infinite-dimensional, time-invariant systems of the following kind:

$$\dot{x}(t) = Ax(t) + Bu(t). \tag{4}$$

Here $A : D(A) \to H$ is the generator of a C_0-semigroup $T(t)$ on the separable, complex Hilbert space H, and B will be our (unbounded) control operator. The input function u is assumed to be in $\boldsymbol{L}_2^{loc}(0, \infty; \mathbb{C}^m)$.

In order to define our class of control operators we have to introduce some notation. We define the space H_{-1} to be the completion of H with respect to the norm

$$\|x\|_{-1} := \|(\beta I - A)^{-1} x\|$$

and the space H_1 to be $D(A)$ with the norm

$$\|x\|_1 := \|(\beta I - A)x\|,$$

where $\beta \in \rho(A)$, the resolvent set of A. It is easy to verify that H_{-1} and H_1 do not depend on $\beta \in \rho(A)$. Moreover, $\| \cdot \|_1$ is equivalent to the graph norm on $D(A)$, so H_1 is complete. In Weiss [33, Remark 3.4] it is shown that $T(t)$ has a restriction to a C_0-semigroup on H_1 whose generator is the restriction of A to $D(A)$, and $T(t)$ can be extended to a C_0-semigroup on H_{-1} whose generator is an extension of A with domain H. Therefore we get

$$A \in \mathcal{L}(H_1, H) \quad \text{and} \quad A \in \mathcal{L}(H, H_{-1}).$$

H_{-1} equals the dual of $D(A^*)$, where we have equipped $D(A^*)$ with the graph norm (see [33]). Following [33] we introduce admissible control operators for $T(t)$.

Definition 2.1. *Let $B \in \mathcal{L}(\mathbb{C}^m, H_{-1})$. For $t \geq 0$ we define the operator $\mathcal{B}_t : \boldsymbol{L}_2(0, \infty; \mathbb{C}^m) \to H_{-1}$ by*

$$\mathcal{B}_t u := \int_0^t T(t-\rho) B u(\rho) \, d\rho.$$

Then B is called an admissible control operator for $T(t)$, *if for some (and hence any) $t > 0$, $\mathcal{B}_t \in \mathcal{L}(\boldsymbol{L}_2(0, \infty; \mathbb{C}^m), H)$.*

By a solution of (4) with initial condition $x(0) = x_0 \in H$ we mean the function defined by the variation of parameters formula

$$x(t) = T(t)x_0 + \int_0^t T(t-\rho) B u(\rho) \, d\rho, \qquad t \geq 0.$$

Note that the admissibility of B guarantees $x(t) \in H$ for $t \geq 0$ and in [33] it is shown that x is continuous. We will denote system (4) by $\Sigma(A, B)$.

In the following the concept of admissible observation operators for $T(t)$ also will be needed (see also Weiss [34]).

DEFINITION 2.2. *Let $C \in \mathcal{L}(H_1, \mathbb{C}^p)$. Then C is called an* admissible observation operator for $T(t)$*, if for some (and hence any) $t > 0$, there is some $K > 0$ such that*

$$\|CT(\cdot)x\|_{\boldsymbol{L}_2(0,t;\mathbb{C}^p)} \leq K\|x\|, \quad x \in D(A).$$

If the semigroup $T(t)$ is exponentially stable, then the constant K does not depend on $t \geq 0$ (see Grabowski and Callier [10]). We end this section with some simple properties of admissible control and observation operators for $T(t)$.

REMARK 2.3. *B is an admissible control operator for the C_0-semigroup $T_\tau(t)$ given by*

$$T_\tau(t) := e^{\tau t}T(t)$$

where $\tau \in \mathbb{R}$ is arbitrary. Moreover, if the operator $C \in \mathcal{L}(H_1, \mathbb{C}^p)$ is an admissible observation operator for $T(t)$, then C also is an admissible observation operator for $T_\tau(t)$.

3. Optimizability, LQ-stabilizing feedbacks, and exact controllability. In this section we introduce the notions of optimizability, LQ-stabilizing feedbacks, and exact controllability. Furthermore, we investigate relationships between these notions. With the system $\Sigma(A, B)$ we associate the cost functional

$$J(x_0, u) := \int_0^\infty \|x(t)\|^2 + \|u(t)\|^2 \, dt. \tag{5}$$

DEFINITION 3.1. *We call the system $\Sigma(A, B)$* g-optimizable*, $g \in \mathbb{R}$, if for every $x_0 \in H$ there exists an input $u \in \boldsymbol{L}_2^{loc}(0, \infty; \mathbb{C}^m)$ such that*

$$J^g(x_0, u) := \int_0^\infty \left[\|e^{-gt}x(t)\|^2 + \|e^{-gt}u(t)\|^2\right] \, dt < \infty. \tag{6}$$

In particular, we call the system $\Sigma(A, B)$ optimizable*, if it is 0-optimizable.*

Thus the system $\Sigma(A, B)$ is optimizable if for every $x_0 \in H$ there exists an input $u \in \boldsymbol{L}_2(0, \infty; \mathbb{C}^m)$ such that the cost functional $J(x_0, u)$ is finite. Optimizability is also known as the *finite cost condition*. Note further that if the system $\Sigma(A, B)$ is g-optimizable for some $g \in \mathbb{R}$, then the system is g'-optimizable for every $g' \geq g$.

DEFINITION 3.2. *Let $F : D(F) \mapsto \mathbb{C}^m$. We call F an* LQ-stabilizing feedback at *g for the system $\Sigma(A, B)$ if there exists a C_0-semigroup $T_F(t)$ on H with generator $A_F : D(A_F) \to H$ such that $F \in \mathcal{L}(D(A_F), \mathbb{C}^m)$, F is an admissible observation operator for $T_F(t)$,*

$$T_F(t)x_0 = T(t)x_0 + \int_0^t T(t-\rho)BFT_F(\rho)x_0 \, d\rho \tag{7}$$

for every $t \geq 0$ and any $x_0 \in H$, and $\|T_F(t)\| \leq Me^{(g-\delta)t}$ for all $t \geq 0$ and some $M, \delta > 0$. We call the feedback F LQ-stabilizing*, if it is LQ-stabilizing at zero.*

Since F is assumed to be an admissible observation operator for $T_F(t)$, we have that $FT_F(\cdot)x \in \boldsymbol{L}_2^{loc}(0, \infty; \mathbb{C}^m)$ for all $x \in H$. Thus the integral in (7) is well-defined.

If an operator F is stabilizing in the sense given in Rebarber [26] and Weiss and Curtain [37], then it is easy to see that it is also an LQ-stabilizing feedback. The difference between their definition and our Definition 3.2 is that we do not assume that F is an admissible observation operator for $T(t)$. Therefore, in general we do not have the second perturbation equation

$$T_F(t)x = T(t)x + \int_0^t T_F(t-\rho)BFT(\rho)x\,d\rho, \quad t \geq 0, x \in H,$$

which holds if F is stabilizing in their sense; see Weiss [36]. At the present it is unknown whether LQ-stabilizing feedbacks and stabilizing feedbacks in the sense of Rebarber are equivalent.

The following theorem shows that in our situation the system $\Sigma(A,B)$ is g-optimizable if and only if there exists a feedback which is LQ-stabilizing at g. So, we have equivalence between the boundedness of the LQ-cost functional (5) and our notion of stability. This inspired us to call our notion "LQ-stabilizing."

THEOREM 3.3. *The following statements are equivalent.*

(1) *The system $\Sigma(A,B)$ is g-optimizable.*

(2) *There exists a feedback which is LQ-stabilizing at g for the system $\Sigma(A,B)$.*

Proof. By choosing $u(t) = FT_F(t)x_0$ it is easy to see that part (2) implies part (1). Thus it remains to prove that part (1) implies part (2). For g equals zero, this implication is proved in Flandoli, Lasiecka, and Triggiani [8, Corollary 4.3 and Theorem 4.4]; see also Zwart [41]. For general $g \in \mathbb{R}$, it is easy to see that the cost $J^g(x_0,u)$ for the system $\Sigma(A,B)$ equals the cost $J(x_0, e^{-g\cdot}u)$ for the system $\Sigma(A-gI,B)$. Hence, the system $\Sigma(A,B)$ is g-optimizable if and only if the system $\Sigma(A-gI,B)$ is optimizable. For the system $\Sigma(A-gI,B)$ we use the result that we already have, and thus there exist an exponentially stable C_0-semigroup $T_F(t)$ and an admissible observation operator F for $T_F(t)$ such that

$$\begin{aligned} T_F(t)x_0 &= e^{-gt}T(t)x_0 + \int_0^t e^{-g(t-\rho)}T(t-\rho)BFT_F(\rho)\,d\rho \\ &= e^{-gt}\left[T(t)x_0 + \int_0^t T(t-\rho)BFe^{g\rho}T_F(\rho)\,d\rho\right]. \end{aligned}$$

This proves that F is an LQ-stabilizing feedback at g for the system $\Sigma(A,B)$. □

For a C_0-semigroup $T(t)$ we define by $g_b(T)$ the *growth bound* of $T(t)$, i.e.,

$$g_b(T) := \lim_{t\to\infty} \frac{1}{t}\log\|T(t)\|.$$

If the system $\Sigma(A,B)$ is LQ-stabilizable, then we have the existence of an exponentially stable C_0-semigroup $T_F(t)$. The growth bound of this semigroup is denoted by g_b^F, i.e.,

$$g_b^F := g_b(T_F). \tag{8}$$

We call g_b^F the *closed-loop growth bound* of the LQ-stabilizing feedback. If the system is optimizable, then g_b^F depends on the LQ-stabilizing feedback F. However, g_b^F will always be negative, and this is the essential property of the closed-loop growth bound that we will use in the following.

PROPOSITION 3.4. *Assume that F is an LQ-stabilizing feedback for the system $\Sigma(A,B)$, and let g_b^F be the corresponding closed-loop growth bound. Then we have the following.*

(1) *For every $\varepsilon \in [0, -g_b^F)$ the feedback F is LQ-stabilizing for the system $\Sigma(A + \varepsilon I, B)$.*

(2) *For every $\varepsilon \in [0, -g_b^F)$ the feedback F is LQ-stabilizing at $-\varepsilon$ for the system $\Sigma(A, B)$.*

Proof. Let $\varepsilon \in [0, -g_b^F)$. By Remark 2.3 B is an admissible control operator for $e^{\varepsilon t}T(t)$. Let F and $T^F(t)$ be as in Definition 3.2. Then Remark 2.3 implies that F is an admissible observation operator for $e^{\varepsilon t}T^F(t)$. By the definition of g_b^F, the C_0-semigroup $T_\varepsilon^F(t) := e^{\varepsilon t}T^F(t)$ is exponentially stable.

1. From (7) we have that

$$\begin{aligned} T_\varepsilon^F(t)x_0 &= e^{\varepsilon t}T_F(t)x_0 \\ &= e^{\varepsilon t}T(t)x_0 + \int_0^t e^{\varepsilon(t-\rho)}T(t-\rho)BFe^{\varepsilon\rho}T_F(\rho)\,d\rho \\ &= e^{\varepsilon t}T(t)x_0 + \int_0^t e^{\varepsilon(t-\rho)}T(t-\rho)BFT_\varepsilon^F(\rho)\,d\rho. \end{aligned}$$

Since $A + \varepsilon I$ is the infinitesimal generator of the C_0-semigroup $e^{\varepsilon t}T(t)$, we see that F is LQ-stabilizing for the system $\Sigma(A + \varepsilon I, B)$;

2. For all $\gamma > 0$, there exists a $M_\gamma > 0$ such that

$$\|T_F(t)\| \le M_\gamma e^{(g_b^F+\gamma)t}.$$

We choose $\gamma > 0$ such that $g_b^F + \gamma < -\varepsilon$ holds, and we define $\delta := -\varepsilon - g_b^F - \gamma$. Then $\delta > 0$ and $g_b^F + \gamma = -\varepsilon - \delta$, and hence by Definition 3.2 the feedback F is LQ-stabilizing at $-\varepsilon$ for the system $\Sigma(A, B)$. □

Next, we show that the spectral subset $\sigma^+_{g_b^F}(A)$ is of a special form if F is a LQ-stabilizing feedback.

Proposition 3.5. *Assume that F is an LQ-stabilizing feedback for the system $\Sigma(A, B)$, and let g_b^F be the closed-loop growth bound. Then the spectral subset $\sigma^+_{g_b^F}(A)$ consists only of point spectrum with finite algebraic multiplicity, and the possible finite accumulation points of the set $\sigma^+_{g_b^F}(A)$ must lie on the line with real part equal to g_b^F.*

Proof. Let $\tau \in (g_b^F, 0)$ and $\varepsilon \in (-\tau, -g_b^F)$. Then Proposition 3.4 and Theorem 3.3 imply that system $\Sigma(A, B)$ is ε-optimizable. Now Rebarber and Zwart [28, Theorem 2.11] imply that the spectral subset $\sigma^+_\tau(A)$ consists only of point spectrum with finite algebraic multiplicity and contains no finite accumulation point. Since $\tau \in (g_b^F, 0)$ is arbitrary, this implies the statement. □

Note that it is possible that the spectrum of A possesses an accumulation point λ_0 with $\mathrm{Re}(\lambda_0) = g_b^F$.

In the following proposition we rewrite LQ-stabilizability in an equivalent way in the frequency domain. We show that LQ-stabilizability implies a useful representation of $x_0 \in H$, the so-called (ξ, ω)-representation. This was introduced for finite-dimensional systems by Hautus [12] and for infinite-dimensional systems by Zwart [40].

Proposition 3.6.

1. *If for every $x_0 \in H$ there exist $\xi \in \boldsymbol{H}_2(\mathbb{C}_0^+; H)$ and $\omega \in \boldsymbol{H}_2(\mathbb{C}_0^+; \mathbb{C}^m)$ such that*

$$x_0 = (sI - A)\xi(s) - B\omega(s), \qquad s \in \mathbb{C}_0^+, \tag{9}$$

then the system $\Sigma(A,B)$ is optimizable, and hence there exists an LQ-stabilizing feedback.

2. *Assume that there exists an LQ-stabilizing feedback for the system $\Sigma(A,B)$, and let g_b^F be its closed-loop growth bound. Then for every $\sigma \in (g_b^F, 0)$ there exists a constant $C > 0$ such that for every $x_0 \in H$ there exist $\xi \in \boldsymbol{H}_\infty(\mathbb{C}_\sigma^+; H) \cap \boldsymbol{H}_2(\mathbb{C}_\sigma^+; H)$ and $\omega \in \boldsymbol{H}_\infty(\mathbb{C}_\sigma^+; \mathbb{C}^m) \cap \boldsymbol{H}_2(\mathbb{C}_\sigma^+; \mathbb{C}^m)$ with*

$$x_0 = (sI - A)\xi(s) - B\omega(s), \qquad s \in \mathbb{C}_\sigma^+, \tag{10}$$

and

$$\max\{\|\xi\|_{\boldsymbol{H}_\infty(\mathbb{C}_\sigma^+;H)}, \|\xi\|_{\boldsymbol{H}_2(\mathbb{C}_\sigma^+;H)}, \|\omega\|_{\boldsymbol{H}_\infty(\mathbb{C}_\sigma^+;\mathbb{C}^m)}, \|\omega\|_{\boldsymbol{H}_2(\mathbb{C}_\sigma^+;\mathbb{C}^m)}\} \leq C\|x_0\|_H.$$

Proof. By [33, Remark 3.12] it follows that for every $u \in \boldsymbol{L}_2(0,\infty;\mathbb{C}^m)$ the function $\Phi_u(t) := \int_0^t T(t-\rho)Bu(\rho)\,d\rho$ is Laplace-transformable with Laplace transform

$$\hat{\Phi}_u(s) = (sI - A)^{-1}B\omega(s), \qquad \mathrm{Re}(s) \geq r \text{ for some } r \in \mathbb{R},$$

where ω is the Laplace transform of u.

1. Let $x_0 \in H$ be arbitrary. Since $\xi \in \boldsymbol{H}_2(\mathbb{C}_0^+; H)$ and $\omega \in \boldsymbol{H}_2(\mathbb{C}_0^+; \mathbb{C}^m)$, by the Paley–Wiener theorem (see, e.g., [3, Theorem A.6.21]) there exist unique $u \in \boldsymbol{L}_2(0,\infty;\mathbb{C}^m)$ and $x \in \boldsymbol{L}_2(0,\infty;H)$ such that ω is the Laplace transform of u and ξ is the Laplace transform of x. Applying Laplace transforms to the function

$$T(t)x_0 + \int_0^t T(t-\rho)Bu(\rho)\,d\rho, \qquad t \geq 0,$$

we get

$$(sI - A)^{-1}x_0 + (sI - A)^{-1}B\omega(s), \qquad \mathrm{Re}(s) \geq r \text{ for some } r \in \mathbb{R}.$$

Since $\xi(s) = (sI-A)^{-1}x_0 + (sI-A)^{-1}B\omega(s)$, $\mathrm{Re}(s) > \max\{r, 0\}$, the uniqueness of the Laplace transform implies

$$x(t) = T(t)x_0 + \int_0^t T(t-\rho)Bu(\rho)\,d\rho.$$

This proves that x is the state trajectory corresponding to x_0 and u and thus the system $\Sigma(A,B)$ is optimizable. By Theorem 3.3 we have the existence of an LQ-stabilizing feedback.

2. Let F be a LQ-stabilizing feedback from Definition 3.2. Taking the Laplace transform of (7), we get

$$(sI - A_F)^{-1}x_0 = (sI - A)^{-1}x_0 + (sI - A)^{-1}BF(sI - A_F)^{-1}x_0. \tag{11}$$

Since $T_F(t)$ has growth bound g_b^F, it is easy to see that

$$\|(sI - A_F)^{-1}x_0\|_{\boldsymbol{H}_\infty(\mathbb{C}_\sigma^+;H)} \leq C_1\|x_0\|_H,$$

and by using the Paley–Wiener theorem [3, Theorem A.6.21] we get

$$\|(sI - A_F)^{-1}x_0\|_{\boldsymbol{H}_2(\mathbb{C}_\sigma^+;H)} \leq C_2\|x_0\|_H.$$

Since F is an admissible observation operator for $T_F(t)$, similar results hold for $F(sI-A_F)^{-1}x_0$. Define $\xi(s) := (sI-A_F)^{-1}x_0$ and $\omega(s) := F(sI-A_F)^{-1}x_0$. Multiplying (11) with $(sI-A)$ and using the definition of ξ and ω, we obtain (10). ◻

We end this section with the definition and some simple properties of exact controllability in finite time.

DEFINITION 3.7. *The system $\Sigma(A, B)$ is called* exactly controllable in finite time *if there exists a time $t_0 \in (0, \infty)$ such that for every $x \in H$ we can find an input $u \in \boldsymbol{L}_2(0, t_0; \mathbb{C}^m)$ such that*

$$x = \int_0^{t_0} T(t_0 - \rho) B u(\rho) \, d\rho.$$

LEMMA 3.8. *Assume that the system $\Sigma(A, B)$ is exactly controllable in finite time. Then for every $\tau \in \mathbb{R}$ the system $\Sigma(A + \tau I, B)$ also is exactly controllable in finite time.*

The following proposition shows that exact controllability in finite time implies g-optimizability for every $g \in \mathbb{R}$.

PROPOSITION 3.9. *If the system $\Sigma(A, B)$ is exactly controllable in finite time, then it also is g-optimizable for every $g \in \mathbb{R}$.*

The proof of Lemma 3.8 and Proposition 3.9 is quite easy and left to the reader. We conclude this section by proving that the spectrum of an exactly controllable system is of a very special form.

THEOREM 3.10. *If the system $\Sigma(A, B)$ is exactly controllable in finite time, then the spectrum of A is pure point spectrum and it contains no (finite) accumulation point.*

Proof. Let $\gamma \in \mathbb{R}$ be arbitrary. Then by Lemma 3.8 we see that the system $\Sigma(A + \gamma I, B)$ is exactly controllable in finite time as well. Proposition 3.9 implies that $\Sigma(A + \gamma I, B)$ is optimizable. Now from Theorem 3.3 and Proposition 3.5 we obtain that $\sigma_0^+(A + \gamma I)$ consists only of point spectrum with no finite accumulation point. Since $\sigma_{-\gamma}^+(A) = \sigma_0^+(A + \gamma I)$, and since γ is arbitrary, the assertion follows. □

4. Decompositions of C_0-semigroups.

For the formulation of Theorem 5.3 we need to write our system into two subsystems. Properties of such a decomposition are given next.

DEFINITION 4.1. *We call a closed subspace V of H $T(t)$*-invariant, *if $T(t)V \subset V$, for all $t \geq 0$.*

In Kurtz [19] it is shown that a closed subspace V is $T(t)$-invariant if and only if $(\lambda I - A)^{-1} V \subset V$ for all $\lambda \in \rho_\infty(A)$; see also Curtain and Zwart [3, Lemma 2.5.6].

LEMMA 4.2. *Let V be a $T(t)$-invariant subspace of H and let $P_V \in \mathcal{L}(H)$ be a projection from H onto V. Define $W := N(P_V)$ and $P_W := I - P_V$. Then the following hold:*

1. *The operator*

$$T_V(t)x := T(t)x, \qquad t \geq 0, x \in V,$$

defines a C_0-semigroup on V with its generator A_V given by

$$A_V x := Ax, \qquad x \in D(A_V) := D(A) \cap V.$$

Furthermore, $\rho_\infty(A) \subset \rho(A_V)$, and for every $v \in V$ and $s \in \rho_\infty(A)$ we have $(sI - A_V)^{-1} v = (sI - A)^{-1} v$.

2. *Let V_{-1} be the completion of V with respect to the norm $\|(\beta I - A_V)^{-1} \cdot \|_V$. Then $V_{-1} \subset H_{-1}$ and $\|v\|_{V_{-1}} = \|v\|_{H_{-1}}$ for every $v \in V_{-1}$.*

3. *The operator*

$$T_W(t)x := P_W T(t)x, \qquad t \geq 0, x \in W,$$

defines a C_0-semigroup on W.

Proof.

1. Follows immediately from Pazy [22, page 123] and [19].
2. From the first part it follows that $\|(sI - A_V)^{-1}v\| = \|(sI - A)^{-1}v\|$ for every $v \in V$. Now the definition of a completion of a normed vector space shows that $V_{-1} \subset H_{-1}$ and $\|v\|_{V_{-1}} = \|v\|_{H_{-1}}$ for every $v \in V_{-1}$.
3. It is easy to see that $T_W(0) = I_W$ holds. Since V is a $T(t)$-invariant subspace, we have $P_W T(t) P_V = 0$. Thus, for $t, s \geq 0$ and $x \in W$ we obtain

$$\begin{aligned} T_W(t)T_W(s)x = P_W T(t) P_W T(s)x &= P_W T(t)T(s)x - P_W T(t)(I - P_W)T(s)x \\ &= P_W T(t+s)x - P_W T(t) P_V T(s)x = T_W(s+t)x. \end{aligned}$$

Since the strong continuity of $T(t)$ immediately implies that $T_W(t)$ is strongly continuous, $T_W(t)$ is a C_0-semigroup on W. □

LEMMA 4.3. *Let V be a $T(t)$-invariant subspace of H, and let $P_V \in \mathcal{L}(H)$ be a projection from H onto V. Define $W := N(P_V)$ and $P_W := I - P_V$. Then there exists a unique admissible control operator $B_W \in \mathcal{L}(\mathbb{C}^m, W_{-1})$ for $T_W(t)$ on W, where $T_W(t)x := P_W T(t)x$ for $t \geq 0$ and $x \in W$, with the property that*

$$\int_0^t T_W(t-\rho) B_W u(\rho)\, d\rho = P_W \int_0^t T(t-\rho) B u(\rho)\, d\rho \tag{12}$$

for every $t \geq 0$ and $u \in \boldsymbol{L}_2(0, t; \mathbb{C}^m)$. This unique operator B_W is given by

$$B_W u := \lim_{\tau \to 0} \frac{1}{\tau} P_W \int_0^\tau T(\tau - \rho) B u\, d\rho, \qquad u \in \mathbb{C}^m, \tag{13}$$

where the limit is taken in W_{-1}.

Additionally, if $\Sigma(A, B)$ is exactly controllable in finite time, then $\Sigma(A_W, B_W)$ also is exactly controllable in finite time. Here A_W is the generator of the C_0-semigroup $T_W(t)$.

Proof. Let $\mathcal{B}_t \in \mathcal{L}(\boldsymbol{L}_2(0, \infty; \mathbb{C}^m), H)$ be the operator as introduced in Definition 2.1. For $u, v \in L_2(0, \infty; \mathbb{C}^m)$ we define $u \underset{\tau}{\diamond} v$ as

$$(u \underset{\tau}{\diamond} v)(t) = \begin{cases} u(t), & t \in [0, \tau], \\ v(t - \tau), & t > \tau. \end{cases}$$

The admissibility of B for the C_0-semigroup $T(t)$ implies (see [33]) that

$$\mathcal{B}_{\tau + t}(u \underset{\tau}{\diamond} v) = T(t)\mathcal{B}_\tau u + \mathcal{B}_t v.$$

Define $\mathcal{B}_t^W \in \mathcal{L}(\boldsymbol{L}_2(0, \infty; \mathbb{C}^m), W)$, $t \geq 0$, by

$$\mathcal{B}_t^W := P_W \mathcal{B}_t.$$

Since V is a $T(t)$-invariant subspace of H, we have that $P_W T(t) P_V = 0$, $t \geq 0$. Thus for every $\tau, t \geq 0$ and every $u, v \in \boldsymbol{L}_2(0, \infty; \mathbb{C}^m)$ we obtain

$$\begin{aligned} \mathcal{B}_{\tau+t}^W(u \underset{\tau}{\diamond} v) = P_W \mathcal{B}_{\tau+t}(u \underset{\tau}{\diamond} v) &= P_W T(t)\mathcal{B}_\tau u + P_W \mathcal{B}_t v = P_W T(t)\mathcal{B}_\tau u + \mathcal{B}_t^W v \\ &= P_W T(t) P_W \mathcal{B}_\tau u + P_W T(t) P_V \mathcal{B}_\tau u + \mathcal{B}_t^W v = T_W(t)\mathcal{B}_\tau^W u + \mathcal{B}_t^W v. \end{aligned}$$

Now [33, Theorem 3.9] implies that there exists a unique $B_W \in \mathcal{L}(\mathbb{C}^m, W_{-1})$ such that

$$\mathcal{B}_t^W u = \int_0^t T_W(t-\rho) B_W u(\rho)\, d\rho,$$

for $t \geq 0$ and $u \in \boldsymbol{L}_2(0, \infty; \mathbb{C}^m)$, and this unique operator B_W is given by (13). Thus B_W is an admissible control operator for $T_W(t)$ and (12) holds.

If $\Sigma(A, B)$ is exactly controllable in finite time, then there exists a time $t_0 > 0$ such that for every $x \in W$ there is an input $u \in \boldsymbol{L}_2(0, t_0; \mathbb{C}^m)$ such that

$$x = \int_0^{t_0} T(t_0 - \rho) B u(\rho)\, d\rho.$$

From (12) it follows directly that the system $\Sigma(A_W, B_W)$ is exactly controllable in finite time as well. □

If P is a projection that commutes with the C_0-semigroup $T(t)$, we have $H = N(P) \oplus \mathrm{Im}(P)$ and both $N(P)$ and $\mathrm{Im}(P)$ are closed $T(t)$-invariant subspaces. In the following we denote by $T_u(t)$ and $T_s(t)$ the restricted C_0-semigroups on $H_u := N(P)$ and $H_s := \mathrm{Im}(P)$, respectively. Moreover, the generators of the C_0-semigroups $T_u(t)$ and $T_s(t)$ are denoted by A_u and A_s. By $[H_s]_{-1}$ ($[H_u]_{-1}$) we denote the completion of H_s (H_u) with respect to the norm $\|(\beta I - A_s)^{-1} \cdot \|_{H_s}$ ($\|(\beta I - A_u)^{-1} \cdot \|_{H_u}$).

The existence of such a projection P is very helpful since it reduces the semigroup into two (possibly simpler) semigroups $T_s(t)$ and $T_u(t)$.

LEMMA 4.4. *Let $T(t)$ be a C_0-semigroup on H and $B \in \mathcal{L}(\mathbb{C}^m, H_{-1})$ an admissible control operator for $T(t)$. Moreover, let $P \in \mathcal{L}(H)$ be a projection that commutes with $T(t)$, $t \geq 0$. Then we have the following.*

1. *P has a unique continuous extension $\tilde{P}$ in $\mathcal{L}(H_{-1})$ with $\mathrm{Im}(\tilde{P}) = [H_s]_{-1}$ and $N(\tilde{P}) = [H_u]_{-1}$. Moreover, $\tilde{P}$ is a projection and $\tilde{P}$ commutes with $T(t)$ and A.*
2. *$\tilde{P}B \in \mathcal{L}(\mathbb{C}^m, [H_s]_{-1})$ is an admissible control operator for $T_s(t)$ on H_s with the property*

$$\int_0^t T_s(t-\rho) \tilde{P} B u(\rho)\, d\rho = P \int_0^t T(t-\rho) B u(\rho)\, d\rho, \tag{14}$$

for $t \geq 0$ and $u \in \boldsymbol{L}_2(0, t; \mathbb{C}^m)$.

3. *$(I - \tilde{P})B \in \mathcal{L}(\mathbb{C}^m, [H_u]_{-1})$ is an admissible control operator for $T_u(t)$ on H_u with the property*

$$\int_0^t T_u(t-\rho)(I - \tilde{P}) B u(\rho)\, d\rho = (I - P) \int_0^t T(t-\rho) B u(\rho)\, d\rho, \tag{15}$$

for $t \geq 0$ and $u \in \boldsymbol{L}_2(0, t; \mathbb{C}^m)$.

Proof.

1. Choose $\beta \in \rho(A)$ with $\mathrm{Re}\,\beta > g_b(T)$, where $g_b(T)$ is the growth bound of $T(t)$. Then we have

$$(\beta I - A)^{-1} x = \int_0^\infty e^{-\beta t} T(t) x\, dt.$$

Thus P commutes with $(\beta I - A)^{-1}$, and hence also with A. Now [33, Proposition 3.3] shows that P has a unique continuous extension $\tilde{P}$ in $\mathcal{L}(H_{-1})$ and that $\tilde{P}$ commutes with $T(t)$ and A. It is now easy to see that $\tilde{P}$ again is a projection.

In order to prove $\operatorname{Im}(\tilde{P}) = [H_s]_{-1}$, we first choose $x \in [H_s]_{-1}$. Then there exists a sequence $\{x_n\}_n \subset H_s$ with $x_n \to x$, as $n \to \infty$, in $[H_s]_{-1}$. Since $\tilde{P}$ is continuous, we get $\tilde{P}x_n \to \tilde{P}x$. On the other hand, $\tilde{P}x_n = Px_n = x_n \to x$, and thus $x \in \operatorname{Im}(\tilde{P})$. Conversely, we choose $x \in \operatorname{Im}(\tilde{P})$. Then there exists an element $z \in H_{-1}$ with $\tilde{P}z = x$ and a sequence $\{z_n\}_n \subset H$ with $z_n \to z$, as $n \to \infty$, in H_{-1}. Since $Pz_n \in H_s$ and $Pz_n = \tilde{P}z_n \to \tilde{P}z = x$, we get $x \in [H_s]_{-1}$. This proves $\operatorname{Im}(\tilde{P}) = [H_s]_{-1}$. Finally, $N(\tilde{P}) = [H_u]_{-1}$ can be proved in a similar manner.

2. By Lemma 4.3 there is a unique admissible control operator $B_s \in \mathcal{L}(\mathbb{C}^m, [H_s]_{-1})$ for $T_s(t)$ with the property that

$$\int_0^t T_s(t-\rho)B_s u(\rho)\,d\rho = P\int_0^t T(t-\rho)Bu(\rho)\,d\rho, \tag{16}$$

for every $t \geq 0$ and $u \in \boldsymbol{L}_2(0,t;\mathbb{C}^m)$. Using [33], we see that B is given by

$$Bu = \lim_{\tau\to\infty}\frac{1}{\tau}\int_0^t T(t-\rho)Bu\,d\rho, \qquad u \in \mathbb{C}^m,$$

where the limit is taken in H_{-1}. Since $\tilde{P} \in \mathcal{L}(H_{-1})$, we get

$$\tilde{P}Bu = \lim_{\tau\to\infty}\frac{1}{\tau}P\int_0^t T(t-\rho)Bu\,d\rho, \qquad u \in \mathbb{C}^m.$$

By Lemma 4.2, $[H_s]_{-1} \subset H_{-1}$ and the norms coincide. Now by Lemma 4.3 we get $B_s = \tilde{P}B$, which completes the proof.

3. The proof for $T_u(t)$ is similar to the proof of part 2. □

5. Equivalent conditions for optimizability. In this section, we develop equivalent conditions for optimizability. We will see that under a weak assumption the system $\Sigma(A,B)$ is optimizable if and only if we can split it into a part that is exponentially stable and a part that is exactly controllable in finite time. This sufficient condition of optimizability was obtained fairly early in 1975 by Triggiani [31] for infinite-dimensional systems with a bounded control operator, under the extra assumption of finitely many unstable eigenvalues. We now show that this holds for admissible control operators for $T(t)$.

THEOREM 5.1. *Assume that there exists a projection $P \in \mathcal{L}(H)$ such that*

(1) *P commutes with the semigroup $T(t)$;*

(2) *$T_s(t) := PT(t)$ is an exponentially stable C_0-semigroup on $H_s := \operatorname{Im}(P)$;*

(3) *$\Sigma(A_u, (I-\tilde{P})B)$ is exactly controllable in finite time. Here A_u is the generator of the C_0-semigroup $T_u(t) := (I-P)T(t)$ on $H_u := N(P)$.*

Then the system $\Sigma(A,B)$ is optimizable.

Proof. Let $x_0 \in H$ be arbitrary. Since $\Sigma(A_u,(I-\tilde{P})B)$ is exactly controllable in finite time there exist a number $t_0 > 0$ and an input $\tilde{u} \in \boldsymbol{L}_2(0,t_0;\mathbb{C}^m)$ such that

$$-T_u(t_0)(I-P)x_0 = \int_0^{t_0} T_u(t_0-\rho)(I-\tilde{P})B\tilde{u}(\rho)\,d\rho. \tag{17}$$

Define $u \in \boldsymbol{L}_2(0,\infty;\mathbb{C}^m)$ by

$$u(s) := \begin{cases} \tilde{u}(s), & s \leq t_0, \\ 0, & s > t_0. \end{cases}$$

Then for $t \geq t_0$ the state trajectory corresponding to the initial state x_0 and the input function u satisfies

$$\begin{aligned} x(t) \quad &= \quad T(t)x_0 + \int_0^t T(t-\rho)Bu(\rho)\,d\rho \\ &\overset{(14)\text{ and }(15)}{=} T_u(t)(I-P)x_0 + T_s(t)Px_0 \\ &\quad + \int_0^t T_u(t-\rho)(I-\tilde{P})Bu(\rho)\,d\rho + \int_0^t T_s(t-\rho)\tilde{P}Bu(\rho)\,d\rho \\ &= \quad T_u(t-t_0)T_u(t_0)(I-P)x_0 + T_s(t)Px_0 \\ &\quad + T_u(t-t_0)\int_0^{t_0} T_u(t_0-\rho)(I-\tilde{P})B\tilde{u}(\rho)\,d\rho \\ &\quad + T_s(t-t_0)\int_0^{t_0} T_s(t_0-\rho)\tilde{P}B\tilde{u}(\rho)\,d\rho \\ &\overset{(17)}{=} \quad T_s(t)Px_0 + T_s(t-t_0)\int_0^{t_0} T_s(t_0-\rho)\tilde{P}B\tilde{u}(\rho)\,d\rho. \end{aligned}$$

Now the exponential stability of $T_s(t)$ implies $x \in \boldsymbol{L}_2(0,\infty;H)$, and thus the system $\Sigma(A,B)$ is optimizable. □

To show that the converse of Theorem 5.1 holds, we need an extra assumption.

Definition 5.2. *We say that the operator A satisfies the* spectrum decomposition assumption at g (SDA(g)), *if there exist numbers $g_1 < g < g_2$ such that*

$$\mathbb{C}^+_{g_1} \cap \mathbb{C}^-_{g_2} \subset \rho(A). \tag{18}$$

We say that A satisfies the SDA, if it satisfies SDA(0).

If an operator A satisfies the SDA(g) in the sense given in Curtain and Zwart [3], then A satisfies the SDA at g in the sense of Definition 5.2. The difference between their definition and Definition 5.2 is that we do not assume that $\sigma^+_g(A)$ is compact.

The result that we want to prove in this section is the following.

Theorem 5.3. *Assume that the system $\Sigma(A,B)$ is optimizable and that A satisfies SDA. Then there exists a projection P such that*

(1) *P commutes with the semigroup $T(t)$,*
(2) *$T_s(t) := PT(t)$ is an exponentially stable C_0-semigroup on $H_s := \mathrm{Im}(P)$,*
(3) *$T_u(t) := (I-P)T(t)$ is a C_0-group on $H_u := N(P)$,*
(4) *$-A_u$ generates an exponentially stable C_0-semigroup,*
(5) *$\Sigma(A_u,(I-\tilde{P})B)$ is exactly controllable in finite time.*

Remark 5.4. *There are examples of optimizable systems $\Sigma(A,B)$ for which A does not satisfy the SDA, but for every $g < 0$ there exists an $\varepsilon \in (0,-g)$ such that A satisfies the SDA at $-\varepsilon$. If this is the case, then using Theorem* 3.3, *Proposition* 3.4, *and Theorem* 5.3, *we get a result similar to that in Theorem* 5.3; *namely, there exists a projection P such that* 1–3 *and* 5 *of Theorem* 5.3 *are satisfied and additionally $-A_u - \varepsilon I$ generates an exponentially stable C_0-semigroup.*

A projection $P \in \mathcal{L}(H)$ satisfying properties (1)–(4) of Theorem 5.3 is also known as dichotomic projection; see [24]. The proof of Theorem 5.3 is given at the end of this section.

Desch and Schappacher [5], Jacobson and Nett [17] and Nefedov and Sholokhovich [20] have proved Theorem 5.3 for bounded control operators B, i.e., $B \in \mathcal{L}(\mathbb{C}^m, H)$. In this situation, for every (LQ)-stabilizing feedback F, the spectral subset $\sigma^+_{g^F_b}(A)$

consists only of finitely many points. Hence, the SDA is satisfied or at least SDA(g) is satisfied for g arbitrarily close to zero. Example 5.11 shows that the assumptions made in Theorem 5.1 do not imply that A satisfies the SDA at g for some $g \leq 0$. In the following we will see that the assumption that A satisfies the SDA is not very restrictive, because there are a lot of situations where this assumption holds automatically; see Proposition 5.9.

Definition 5.5. *We call the C_0-semigroup $T(t)$* completely unstable *if*

$$\{x \in H : T(\cdot)x \in \boldsymbol{L}_2(0,\infty;H)\} = \{0\}.$$

For example, if A generates a C_0-group $T(t)$ and $-A$ generates a bounded C_0-semigroup, then $T(t)$ is completely unstable.

As a corollary of Theorem 5.3 we show that under the assumption that A generates a completely unstable C_0-semigroup, the system $\Sigma(A,B)$ is optimizable if and only if the system $\Sigma(A,B)$ is exactly controllable in finite time.

Corollary 5.6. *If A generates a completely unstable C_0-semigroup, then the following statements are equivalent.*

(1) *The system $\Sigma(A,B)$ is optimizable.*

(2) *There exists an LQ-stabilizing feedback for the system $\Sigma(A,B)$.*

(3) *The system $\Sigma(A,B)$ is exactly controllable in finite time.*

Moreover, if one of these statements holds, then $T(t)$ is a C_0-group on H, and A satisfies the SDA at any point between the closed-loop growth bound and zero.

Proof. The equivalence between part 1 and part 2 is proved in Theorem 3.3. Proposition 3.9 shows that part 3 implies part 1. We now prove that part 2 implies part 3. Let F be an LQ-stabilizing feedback for the system $\Sigma(A,B)$. Proposition 3.5 implies that the spectral subset $\sigma^+_{g_b^F}(A)$ consists only of point spectrum. Let us assume that there is a $\lambda \in \mathbb{C}^+_{g_b^F} \cap \mathbb{C}^-_0 \cap \sigma(A)$ and x_λ a corresponding eigenvector. In [22, p. 46] it is shown that $e^{\lambda t}$ is an eigenvalue of $T(t)$ with eigenvector x_λ. This implies that $T(\cdot)x_\lambda = e^{\lambda t}x_\lambda \in \boldsymbol{L}_2(0,\infty;H)$, which is in contradiction with the fact that $T(t)$ is completely unstable. Thus $\mathbb{C}^+_{g_b^F} \cap \mathbb{C}^-_0 \subset \rho(A)$ and so A satisfies the SDA($-\varepsilon$) for any $\varepsilon \in (0, -g_b^F)$. Remark 5.4 now implies the existence of a projection P such that 1–3 and 5 of Theorem 5.3 are satisfied. Since $T_s(t) = PT(t) = T(t)P$, we get $T(\cdot)x \in L_2(0,\infty;H)$ for every $x \in H_s$. Using the complete instability of our system this can only happen if $H_s = 0$. Therefore, $H_u = H$ and $T(t)$ equals $T_u(t)$ which implies that $T(t)$ is a C_0-group and that the system $\Sigma(A,B)$ is exactly controllable in finite time. □

The result presented in the previous corollary is closely related to Russell's principle on exact controllability. Russell [29] proved exact controllability for a system governed by the wave equation by showing that the system was stabilizable and backwards stabilizable. The definition of stabilizability used in [29] is stronger than the definition used in this paper. Rebarber and Weiss [27] extended Russell's result to our notion of stabilizability. Their result is closely related to the corollary above, but it is neither a consequence nor a generalization of our result.

Example 5.7. *Consider the right-shift semigroup $T(t)$ on $\boldsymbol{L}_2(0,\infty)$ with generator A. This semigroup satisfies $\|T(t)x_0\| = \|x_0\|$ for every $x_0 \in \boldsymbol{L}_2(0,\infty)$. Hence the right shift semigroup on $\boldsymbol{L}_2(0,\infty)$ is completely unstable. Since this C_0-semigroup is not a C_0-group, Corollary* 5.6 *implies that there cannot exist an admissible finite-rank control operator B for $T(t)$ such that the system $\Sigma(A,B)$ is optimizable.*

Definition 5.8. *We call a C_0-semigroup $T(t)$* eventually norm continuous *if there exists a constant $t' \geq 0$ such that the function $t \to T(t)$ from (t', ∞) into $\mathcal{L}(H)$ is operator-norm continuous.*

Examples for eventually norm continuous C_0-semigroups are holomorphic, differentiable, norm continuous, and compact C_0-semigroups; see Arendt et al. [1, p. 41].

Proposition 5.9. *Assume that F is an LQ-stabilizing feedback for the system $\Sigma(A, B)$, and let g_b^F be its closed-loop growth bound. If the C_0-semigroup $T(t)$ is eventually norm continuous, then for every $\tau \in (g_b^F, 0)$, the set $\sigma_\tau^+(A)$ contains only finitely many points, and thus A satisfies SDA(g) for all but finitely many $g > g_b^F$.*

Proof. By Proposition 3.5 the spectral subset $\sigma_{g_b^F}^+(A)$ contains no finite accumulation point. Let $\tau \in (g_b^F, 0)$. Since by [1, A-II Theorem 1.20] the set $\{\lambda \in \sigma(A) | \mathrm{Re}(\lambda) \geq \tau\}$ is bounded, we have that $\sigma_\tau^+(A)$ contains only finitely many points. Thus the assertion follows. □

In the special case where A is a Riesz-spectral operator we are able to prove Theorem 5.3 without the assumption that A satisfies the SDA. For the definition of a Riesz-spectral operator we refer the reader to [3, Definition 2.3.4].

Theorem 5.10. *Assume that F is an LQ-stabilizing feedback for the system $\Sigma(A, B)$ and that A is a Riesz-spectral operator. Let g be negative and larger than the closed-loop growth bound. Then there exists a projection $P \in \mathcal{L}(H)$ such that* 1–3 *and* 5 *of Theorem* 5.3 *are satisfied and additionally the growth bound of the C_0-semigroup generated by $-A_u$ is at most $-g$.*

Proof. By the definition of a Riesz-spectral operator, there exist sequences $(\lambda_n)_n \subset \mathbb{C}$, $(\phi_n)_n \subset H$, and $(\psi_n)_n \subset H$, with $\langle \phi_n, \psi_m \rangle = \delta_{n,m}$, such that

$$Az := \sum_{n \in \mathbb{N}} \lambda_n \langle z, \psi_n \rangle \phi_n, \qquad z \in H,$$

$$D(A) := \left\{ z \in H | \sum_{n \in \mathbb{N}} |\lambda_n|^2 |\langle z, \psi_n \rangle|^2 < \infty \right\}.$$

Let $T_F(t)$ and F be the operators given by Definition 3.2. Defining $S := \{n \in \mathbb{N} \mid \mathrm{Re}(\lambda_n) \leq g\}$, it is easy to see that

$$Pz := \sum_{n \in S} \langle z, \psi_n \rangle \phi_n$$

is a projection operator on H which commutes with A and $T(t)$, $t \geq 0$. Moreover, by [3, Theorem 2.3.5] we get that

(1) $T_s(t) := PT(t)$ is an exponentially stable C_0-semigroup on $H_s := \mathrm{Im}(P)$,

(2) $T_u(t) := (I - P)T(t)$ is a C_0-group on $H_u := N(P)$,

(3) $\|T_u(t)x\| \leq Me^{gt}\|x\|$ for all $x \in H_u$, $t \leq 0$,

where $M > 0$ is independent of x and t. Lemma 4.4 now shows that $(I - \tilde{P})B$ is an admissible control operator for $T_u(t)$. Thus it remains to prove that $\Sigma(A_u, (I - \tilde{P})B)$ is exactly controllable in finite time. Since $T_{-g}^F(t) := e^{-gt}T_F(t)$ is exponentially stable and $T_u(t)$ satisfies the inequality in outcome (3), the operator $T_u(-t)(I-P)T_F(t)|_{H_u} \in \mathcal{L}(H_u)$ satisfies the estimate

$$\|T_u(-t)(I - P)T_F(t)|_{H_u}\| = \|e^{gt}T_u(-t)(I - P)T_{-g}^F(t)|_{H_u}\| \leq Ce^{-\tau t}, \qquad t \geq 0,$$

for some constants $\tau > 0$ and $C > 0$, which are independent of t. Using the positivity of τ, we can choose a $t_0 > 0$ (sufficiently large) such that $\|T_u(-t_0)(I-P)T_F(t_0)|_{H_u}\| <$

1, and thus the operator $I|_{H_u} - T_u(-t_0)(I-P)T_F(t_0)|_{H_u}$ is invertible in $\mathcal{L}(H_u)$. Let $x_0 \in H_u$ be arbitrary and define

$$\begin{aligned} z_0 &:= \left[I - T_u(-t_0)(I-P)T_F(t_0)|_{H_u}\right]^{-1} T_u(-t_0)x_0, \\ u(\rho) &:= -FT_F(\rho)z_0. \end{aligned}$$

Thus $u \in \boldsymbol{L}_2(0,t_0;\mathbb{C}^m)$. By (7) and (15) we get

$$\begin{aligned} (I-P)T_F(t_0)z_0 &= T_u(t_0)z_0 - (I-P)\int_0^{t_0} T(t_0-\rho)Bu(\rho)\,d\rho \\ &= T_u(t_0)z_0 - \int_0^{t_0} T_u(t_0-\rho)(I-\tilde{P})Bu(\rho)\,d\rho, \end{aligned}$$

which is equivalent to

$$T_u(t_0)\left[z_0 - T_u(-t_0)(I-P)T_F(t_0)z_0\right] = \int_0^{t_0} T_u(t_0-\rho)(I-\tilde{P})Bu(\rho)\,d\rho$$

and thus

$$x_0 = \int_0^{t_0} T_u(t_0-\rho)(I-\tilde{P})Bu(\rho)\,d\rho.$$

Consequently, $\Sigma(A_u,(I-\tilde{P})B)$ is exactly controllable in finite time. □

The following example shows that there exist optimizable systems $\Sigma(A,B)$ for which A does not satisfy SDA(g) at any $g \le 0$, but there is a projection P that fulfills (1)–(3) of Theorem 5.1.

Example 5.11. Let $\{e_n\}_{n\in\mathbb{Z}} \cup \{f_n\}_{n\in\mathbb{Z}}$ be an orthonormal basis of ℓ_2, $\{\alpha_n\}_{n\in\mathbb{Z}} = \mathbb{Q}\cap[-1,0)$, $\{\beta_n\}_{n\in\mathbb{Z}} = \mathbb{Q}\cap(-\infty,-1)$, $\mu_n := \alpha_n + in$ and $\nu_n := \beta_n + in$, $n\in\mathbb{Z}$. We define $A : D(A) \to H$ by

$$\begin{aligned} A &:= \sum_{n\in\mathbb{Z}} \mu_n\langle\cdot,e_n\rangle e_n + \sum_{n\in\mathbb{Z}} \nu_n\langle\cdot,f_n\rangle f_n, \\ D(A) &:= \{x\in H |\sum_{n\in\mathbb{Z}} |\mu_n|^2|\langle x,e_n\rangle|^2 + \sum_{n\in\mathbb{Z}} |\nu_n|^2|\langle x,f_n\rangle|^2 < \infty\} \end{aligned}$$

and $B : \mathbb{C} \to H_{-1}$ by

$$B := \sum_{n\in\mathbb{Z}} e_n.$$

Clearly, A is a Riesz-spectral operator, and A does not satisfy SDA(g) for any $g \le 0$.

A is the generator of the C_0-semigroup $T(t)$ given by

$$T(t) := \sum_{n\in\mathbb{Z}} e^{\mu_n t}\langle\cdot,e_n\rangle e_n + \sum_{n\in\mathbb{Z}} e^{\nu_n t}\langle\cdot,f_n\rangle f_n.$$

Weiss [32] shows that B is an admissible control operator for $T(t)$. It is easy to see that

$$Pz := \sum_{n\in\mathbb{Z}} \langle z,f_n\rangle f_n$$

is a projection operator on H which commutes with A and $T(t)$, $t \ge 0$. Moreover, by [3, Theorem 2.3.5] we get that

(1) $T_s(t) := PT(t)$ is an exponentially stable C_0-semigroup on $H_s := \mathrm{Im}(P)$,
(2) $T_u(t) := (I - P)T(t)$ is a C_0-group on $H_u := N(P)$,
(3) $\|T_u(t)x\| \leq Me^{-t}\|x\|$ for all $x \in H_u$, $t \leq 0$,
where $M > 0$ is independent of x and t. Lemma 4.4 now shows that $(I - \tilde{P})B$ is an admissible control operator for $T_u(t)$.

Next we prove that $\Sigma(A_u, (I - \tilde{P})B)$ is exactly controllable in finite time. By Young [38, Corollary 2, p. 196] we get that $\{e^{\mu_n t}\}_{n\in\mathbb{Z}}$ forms a Riesz basis of $\boldsymbol{L}_2(-\pi, \pi)$, and thus it is easy to show that $\{e^{\mu_n t}\}_{n\in\mathbb{Z}}$ forms a Riesz basis for $\boldsymbol{L}_2(0, 2\pi)$. This implies that $\{g_n\}_{n\in\mathbb{Z}}$, $g_n(t) := e^{\mu_n(2\pi - t)}$ is a Riesz basis for $\boldsymbol{L}_2(0, 2\pi)$ as well. Let $\{h_n\}_{n\in\mathbb{Z}}$ be the biorthogonal sequence of $\{g_n\}_{n\in\mathbb{Z}}$ in $\boldsymbol{L}_2(0, 2\pi)$ [38, p. 29], which again is a Riesz basis of $\boldsymbol{L}_2(0, 2\pi)$ [38, p. 36]. By definition, the biorthogonality of $\{g_n\}_{n\in\mathbb{Z}}$ and $\{h_n\}_{n\in\mathbb{Z}}$ implies that

$$\int_0^{2\pi} g_m(t)\overline{h_n(t)}\,dt = \delta_{nm}, \qquad n, m \in \mathbb{Z}.$$

Thus we get

$$\begin{aligned}\int_0^{2\pi} T_u(2\pi - \rho)B\overline{h_n(\rho)}\,d\rho &= \int_0^{2\pi} \sum_{m\in\mathbb{Z}} e^{\mu_m(2\pi-\rho)}\overline{h_n(\rho)}e_m\,d\rho \\ &= \sum_{m\in\mathbb{Z}} \int_0^{2\pi} g_m(\rho)\overline{h_n(\rho)}\,d\rho\, e_m \qquad (19) \\ &= e_n.\end{aligned}$$

Let $x \in H_u$ be arbitrary. Then $x = \sum_{n\in\mathbb{Z}} \langle x, e_n\rangle e_n$ and $\{\langle x, e_n\rangle\}_{n\in\mathbb{Z}} \in \ell_2$. Defining $u \in \boldsymbol{L}_2(0, 2\pi)$ by

$$u := \sum_{n\in\mathbb{Z}} \langle x, e_n\rangle \overline{h_n},$$

(19) implies that

$$\int_0^{2\pi} T(2\pi - \rho)Bu(\rho)\,d\rho = x.$$

This shows that system $\Sigma(A_u, (I - \tilde{P})B)$ is exactly controllable in finite time. Finally, Theorem 5.1 implies that system $\Sigma(A, B)$ is optimizable.

For the proof of Theorem 5.3 we need the following lemma.

LEMMA 5.12. *Assume that system $\Sigma(A, B)$ is optimizable. Let $(s_n)_n$ be a sequence in $\rho(A) \cap \overline{\mathbb{C}_0^+}$ and $\delta > 0$ be a constant such that*

$$\{s \in \mathbb{C} \mid |s - s_n| < \delta\} \subset \rho(A) \tag{20}$$

for every $n \in \mathbb{N}$. Then

$$\sup_{n\in\mathbb{N}} \|(s_n I - A)^{-1}B\| < \infty \qquad \text{and} \qquad \sup_{n\in\mathbb{N}} \|(s_n I - A)^{-1}\| < \infty.$$

Proof. First we prove that $\sup_{n\in\mathbb{N}} \|(s_n I - A)^{-1}B\| < \infty$ holds. Since the system is optimizable, there exists an LQ-stabilizing feedback F. Let g_b^F be the (negative) closed-loop growth bound, let $\sigma \in (g_b^F, 0)$, and define $\tau := \frac{1}{2}\min\{-\sigma, \delta\}$. If

$\sup_{n\in\mathbb{N}} \|(s_nI-A)^{-1}B\| = \infty$, then the principle of uniform boundedness implies that there exist a vector $u \in \mathbb{C}^m$ and an element $y \in H$ such that

$$\sup_{n\in\mathbb{N}} |\langle (s_nI-A)^{-1}Bu, y\rangle| = \infty. \tag{21}$$

Since B is an admissible control operator for $T(t)$, there exists a constant $\gamma > \tau$ such that

$$(\cdot I-A)^{-1}B \in \boldsymbol{H}_\infty(\mathbb{C}^+_\gamma; H) \tag{22}$$

[35, Proposition 2.3]. Since $s_n \in \overline{\mathbb{C}^+_0}$, we get $s_n + 2\gamma \in \mathbb{C}^+_\gamma$.

We write B as $[b_1,\dots,b_m]$. By Proposition 3.6, for every $n \in \mathbb{N}$ and all $k \in \{1,\dots,m\}$ there exist $\xi_{s_n,k} \in \boldsymbol{H}^\infty(\mathbb{C}^+_\sigma; H)$ and $\omega_{s_n,k} \in \boldsymbol{H}^\infty(\mathbb{C}^+_\sigma; \mathbb{C}^m)$ such that

$$((s_n+2\gamma)I-A)^{-1}b_k = (sI-A)\xi_{s_n,k}(s) - B\omega_{s_n,k}(s), \qquad s \in \mathbb{C}^+_\sigma.$$

Moreover, using (22) and Proposition 3.6, the functions $\xi_{s_n,k}$ and $\omega_{s_n,k}$ can be chosen in such a way that there exists a constant $M > 0$ with

$$\|\xi_{s_n,k}\|_{\boldsymbol{H}_\infty(\mathbb{C}^+_\sigma;H)} \le M \qquad \text{and} \qquad \|\omega_{s_n,k}\|_{\boldsymbol{H}_\infty(\mathbb{C}^+_\sigma;\mathbb{C}^m)} \le M$$

for $n \in \mathbb{N}$ and $k \in \{1,\dots,m\}$. Defining $\Xi_{s_n}(s) := [\xi_{s_n,1}(s),\dots,\xi_{s_n,m}(s)]$ and $\Omega_{s_n}(s) := [\omega_{s_n,1}(s),\dots,\omega_{s_n,m}(s)]$, we obtain

$$((s_n+2\gamma)I-A)^{-1}B = (sI-A)\Xi_{s_n}(s) - B\Omega_{s_n}(s), \qquad s \in \mathbb{C}^+_\sigma. \tag{23}$$

The functions Ξ_{s_n} and Ω_{s_n} also are holomorphic on $\mathbb{C}^+_\sigma$ and there exists a number $\tilde{M} > 0$ with

$$\|\Xi_{s_n}\|_{\boldsymbol{H}_\infty(\mathbb{C}^+_\sigma;H^m)} \le \tilde{M} \qquad \text{and} \qquad \|\Omega_{s_n}\|_{\boldsymbol{H}_\infty(\mathbb{C}^+_\sigma;\mathbb{C}^{m\times m})} \le \tilde{M}. \tag{24}$$

Now for $s \in \mathbb{C}^+_\sigma$, we define the functions

$$\widetilde{\Xi}_{s_n}(s) := \Xi_{s_n}(s_n+s) \qquad \text{and} \qquad \widetilde{\Omega}_{s_n}(s) := \Omega_{s_n}(s_n+s).$$

Since $\mathrm{Re}(s_n) \ge 0$, the functions $\widetilde{\Xi}_{s_n}$ and $\widetilde{\Omega}_{s_n}$ also are holomorphic and bounded (with the same estimates) on $\mathbb{C}^+_\sigma$. Multiplying equation (23) (with $s_n + s$ instead of s) by $((s_n+s)I-A)^{-1}$, $|s| < \tau$, and using the resolvent identity, we get

$$((s_n+s)I-A)^{-1}B\left[\widetilde{\Omega}_{s_n}(s) - \frac{I}{s-2\gamma}\right] = \widetilde{\Xi}_{s_n}(s) - \frac{1}{s-2\gamma}((s_n+2\gamma)I-A)^{-1}B. \tag{25}$$

Note that the existence of $((s_n+s)I-A)^{-1}$, $|s| < \tau$, is guaranteed by (20). Since $\{\widetilde{\Omega}_{s_n}\}_n$ is a uniformly bounded set in $\boldsymbol{H}_\infty(\mathbb{C}^+_\sigma;\mathbb{C}^{m\times m})$, there exists a subsequence $\{\widetilde{\Omega}_{s_{n_j}}\}_j$ of $\{\widetilde{\Omega}_{s_n}\}_n$ which converges uniformly on compact subsets of $\mathbb{C}^+_\sigma$ to a holomorphic function $\widetilde{\Omega} \in \boldsymbol{H}_\infty(\mathbb{C}^+_\sigma;\mathbb{C}^{m\times m})$ (see Hille and Phillips [13, Theorem 3.14.2]). We denote this subsequence again by $\{\widetilde{\Omega}_{s_n}\}_n$.

For $\beta \in (0,\tau]$ we define $B(\beta) := \{s \in \mathbb{C} : |s| \le \beta\}$. We now prove that there exists a number $\tilde{\tau} \in (0,\tau]$ such that

$$\det\left(\widetilde{\Omega}(s) - \frac{I}{s-2\gamma}\right) \neq 0$$

for all $s \in B(\tilde{\tau})\backslash\{0\}$. If it were not true, then there would exist a sequence $(\delta_n)_n \subset B(\tau)\backslash\{0\}$, tending to 0 as n tends to ∞, such that

$$\det\left(\widetilde{\Omega}(\delta_n) - \frac{I}{\delta_n - 2\gamma}\right) = 0, \quad n \in \mathbb{N}.$$

Now

$$\det\left(\widetilde{\Omega}(\delta_n) - \frac{I}{\delta_n - 2\gamma}\right) = \det\Big((\delta_n - 2\gamma)\widetilde{\Omega}(\delta_n) - I\Big)\det\left(\frac{I}{\delta_n - 2\gamma}\right)$$

implies that

$$\det\Big((\delta_n - 2\gamma)\widetilde{\Omega}(\delta_n) - I\Big) = 0, \qquad n \in \mathbb{N}.$$

In other words, $\det((s-2\gamma)\widetilde{\Omega}(s) - I)$ would have a converging sequence of zeros in $B(\tau)$. By the holomorphicity of $\widetilde{\Omega}$ on $\mathbb{C}_\sigma^+$ and of the determinant on $\mathbb{C}^{m\times m}$, we get that $\det((s-2\gamma)\widetilde{\Omega}(s) - I)$ is holomorphic on $\mathbb{C}_\sigma^+$. Since $-\tau > \sigma$, this would imply that the function $\det((s-2\gamma)\widetilde{\Omega}(s) - I)$ would be equal to zero everywhere on $\mathbb{C}_\sigma^+$. Taking $s = 2\gamma$ (note that $2\gamma \in \mathbb{C}_\sigma^+$), we get a contradiction. Therefore, there exists a number $\tilde{\tau} \in (0, \tau]$ such that

$$\det\left(\widetilde{\Omega}(s) - \frac{I}{s - 2\gamma}\right) \neq 0 \qquad \text{for all } s \in B(\tilde{\tau})\backslash\{0\}. \tag{26}$$

On $B(\tilde{\tau})$ we now define the functions

$$f_n(s) := \langle((s+s_n)I - A)^{-1}Bu, y\rangle, \qquad n \in \mathbb{N},$$

where u and y are the vectors from (21). Since $\tilde{\tau} < \delta$, we get by (20) that f_n is holomorphic, and so by the maximum principle there exists a number $\theta_n \in [0, 2\pi)$ such that

$$|f_n(\tilde{\tau}e^{i\theta_n})| \geq |f_n(0)|. \tag{27}$$

Since (by (21)) $\sup_{n\in\mathbb{N}} |f_n(0)| = \infty$, equation (27) implies

$$\sup_{n\in\mathbb{N}} |f_n(\tilde{\tau}e^{i\theta_n})| = \infty. \tag{28}$$

There now exists a subsequence of $\{\theta_n\}_{n=1}^\infty$ which converges to $\theta \in [0, 2\pi]$. Again we rename the subsequence $\{\theta_n\}_{n=1}^\infty$. By the choice of $\tilde{\tau}$, we have that

$$\det\left(\widetilde{\Omega}(\tilde{\tau}e^{i\theta}) - \frac{I}{\tilde{\tau}e^{i\theta} - 2\gamma}\right) \neq 0.$$

The set of invertible matrices is open and $\widetilde{\Omega}_{s_n}(\tilde{\tau}e^{i\theta_n}) - \frac{I}{\tilde{\tau}e^{i\theta_n} - 2\gamma}$ converges to $\widetilde{\Omega}(\tilde{\tau}e^{i\theta}) - \frac{I}{\tilde{\tau}e^{i\theta} - 2\gamma}$. Thus there exists a number $N \in \mathbb{N}$ such that

$$\det\left(\widetilde{\Omega}_{s_n}(\tilde{\tau}e^{i\theta_n}) - \frac{I}{\tilde{\tau}e^{i\theta_n} - 2\gamma}\right) \neq 0, \qquad n \geq N.$$

Multiplying (25) (with $s := \tilde{\tau} e^{i\theta_n}$) by $[\widetilde{\Omega}_{s_n}(\tilde{\tau} e^{i\theta_n}) - \frac{I}{\tilde{\tau} e^{i\theta_n} - 2\gamma}]^{-1} u$ and taking the inner product with y, we obtain

$$\begin{aligned} f_n(\tilde{\tau} e^{i\theta_n}) = \Bigg\langle & \left[\widetilde{\Xi}_{s_n}(\tilde{\tau} e^{i\theta_n}) - \frac{1}{\tilde{\tau} e^{i\theta_n} - 2\gamma} ((s_n + 2\gamma) I - A)^{-1} B \right] \\ & \cdot \left[\widetilde{\Omega}_{s_n}(\tilde{\tau} e^{i\theta_n}) - \frac{I}{\tilde{\tau} e^{i\theta_n} - 2\gamma} \right]^{-1} u, y \Bigg\rangle . \end{aligned} \tag{29}$$

Now $[\widetilde{\Omega}_{s_n}(\tilde{\tau} e^{i\theta_n}) - \frac{I}{\tilde{\tau} e^{i\theta_n} - 2\gamma}]^{-1}$ converges to $[\widetilde{\Omega}(\tilde{\tau} e^{i\theta}) - \frac{I}{\tilde{\tau} e^{i\theta} - 2\gamma}]^{-1}$. Thus using (22) and (24), we see that the right-hand side of (29) is bounded. However, this is in contradiction with (28) and so

$$\sup_{n \in \mathbb{N}} \|(s_n I - A)^{-1} B\| < \infty. \tag{30}$$

It now remains to prove that $\sup_{n \in \mathbb{N}} \|(s_n I - A)^{-1}\| < \infty$. Assume that $\sup_{n \in \mathbb{N}} \|(s_n I - A)^{-1}\| = \infty$. By the principle of uniform boundedness there exists a $z \in H$ such that

$$\sup_{n \in \mathbb{N}} \|(s_n I - A)^{-1} z\| = \infty. \tag{31}$$

By Proposition 3.6 there exist $\omega_z \in \boldsymbol{H}_\infty(\mathbb{C}_\sigma^+; \mathbb{C}^m)$ and $\xi_z \in \boldsymbol{H}_\infty(\mathbb{C}_\sigma^+; H)$ such that we have

$$\xi_z(s_n) = (s_n I - A)^{-1} z + (s_n I - A)^{-1} B \omega_z(s_n), \qquad n \in \mathbb{N}.$$

Since $\{\|\xi_z(s_n)\|\}_n$ and $\{\|\omega_z(s_n)\|\}_n$ are bounded sets, it follows with (31) that

$$\sup_{n \in \mathbb{N}} \|(s_n I - A)^{-1} B\| = \infty$$

which is in contradiction with (30). This completes the proof. □

COROLLARY 5.13. *Assume that* $\Sigma(A, B)$ *is optimizable and let* A *satisfy the SDA. Then*

$$(\cdot I - A)^{-1} \in \boldsymbol{L}_\infty(i\mathbb{R}; \mathcal{L}(H)) \quad \text{ and } \quad (\cdot I - A)^{-1} B \in \boldsymbol{L}_\infty(i\mathbb{R}; \mathcal{L}(\mathbb{C}^m, H)).$$

Proof. Since A satisfies the SDA, there exists a $\delta > 0$ such that $\mathbb{C}_{-\delta}^+ \cap \mathbb{C}_\delta^- \subset \rho(A)$. Now it is easy to see that $(\cdot I - A)^{-1} : i\mathbb{R} \to \mathcal{L}(H)$ and $(\cdot I - A)^{-1} B : i\mathbb{R} \to \mathcal{L}(\mathbb{C}^m, H)$ are continuous. If the statement does not hold, then there exists a sequence $(s_n)_n \in i\mathbb{R}$ such that

$$\lim_{n \to \infty} \|(s_n I - A)^{-1}\| = \infty \quad \text{ or } \quad \lim_{n \to \infty} \|(s_n I - A)^{-1} B\| = \infty,$$

which is in contradiction with Lemma 5.12. □

By $g_{\sigma_p}(A)$ we denote the *bound for the point spectrum of* A, i.e.,

$$g_{\sigma_p}(A) := \sup\{\operatorname{Re}(\lambda) \mid \lambda \in \sigma_p(A)\}.$$

It is easy to see that $g_{\sigma_p}(A) \le g_b(T)$. The following theorem shows that for an optimizable system $\Sigma(A, B)$ we cannot have $g_{\sigma_p}(A) < 0 \le g_b(T)$. For one-dimensional control operators this result also can be found in [28].

THEOREM 5.14. *Assume that $\Sigma(A, B)$ is optimizable and that $\sigma_p(A) \subset \mathbb{C}^-_{-\tau}$ for some $\tau > 0$. Then $T(t)$ is exponentially stable.*

Proof. If $T(t)$ is not exponentially stable, then Huang [14] shows that $\|(sI-A)^{-1}\|$ is not bounded in $\mathbb{C}^+_0$. Therefore, there exists a sequence $(s_n)_n \subset \mathbb{C}^+_0$ such that

$$\lim_{n\to\infty} \|(s_nI - A)^{-1}\| = \infty. \tag{32}$$

Since $\Sigma(A, B)$ is optimizable, by Theorem 3.3 there exists an LQ-stabilizing feedback F. By Proposition 3.5 we get that $\sigma^+_{g^F_b}(A)$ consists of only point spectrum. Thus, for $-\delta := \max\{-\tau, g^F_b\}$ we obtain $\sigma^+_{-\delta}(A) = \emptyset$. From the fact $\sigma^+_{-\delta}(A) = \emptyset$, we see that the sequence $(s_n)_n$ satisfies equation (20). Hence equation (32) is in contradiction with the result of Lemma 5.12. □

REMARK 5.15. *There are many cases where a generator satisfies $g_{\sigma_p}(A) < 0$ and $g_b(T) \geq 0$. This can happen, for example, when the spectrum consists solely of eigenvalues, but the multiplicity of the eigenvalues is not bounded* (*see Zabczyk* [39]) *or for certain shift semigroups* (*see Greiner, Voigt, and Wolff* [11] *or Curtain and Zwart* [3, *Example* 5.1.4]). *This means that for such generators A there exists no admissible finite-rank control operator B for $T(t)$ such that the system $\Sigma(A, B)$ is optimizable.*

We are now going to prove the main result of this section.

Proof of Theorem 5.3. Since $\Sigma(A, B)$ is optimizable, we have from Theorem 3.3 the existence of an LQ-stabilizing feedback F. Let $T_F(t)$ be as in Definition 3.2, and let g^F_b be the (negative) growth bound of $T_F(t)$. By the SDA we have the existence of $g_1 < 0 < g_2$ such that (18) holds. Since (18) remains valid if we increase g_1 (g_1 must remain negative), without loss of generality we may assume that $g^F_b < g_1$ holds. Let $g \in (g_1, g_2)$ be arbitrary. By Theorem 3.3 and Proposition 3.4 the system $\Sigma(A-gI, B)$ is optimizable. The choice of g implies the existence of $\delta > 0$ such that

$$\mathbb{C}^+_{-\delta} \cap \mathbb{C}^-_{\delta} \subset \rho(A - gI),$$

and thus Corollary 5.13 shows

$$((\cdot + g)I - A)^{-1} \in \boldsymbol{L}_\infty(i\mathbb{R}; \mathcal{L}(H)).$$

Therefore Prüss [24, Corollary 5] shows that

$$\{\lambda \in \mathbb{C} \mid |\lambda| = e^g\} \subset \rho(T(1))$$

holds. Since $g \in (g_1, g_2)$ is arbitrary, this implies

$$\{\lambda \in \mathbb{C} \mid e^{g_1} < |\lambda| < e^{g_2}\} \subset \rho(T(1)).$$

Let $c \in (g_1, g_2)$ be arbitrary and define $P \in \mathcal{L}(H)$ by

$$P := \int_{|s|=e^c} (sI - T(1))^{-1}\, ds.$$

It is easy to see that $P^2 = P$ holds, i.e., P is a projection, and that P commutes with the C_0-semigroup $T(t)$. Thus part (1) is satisfied. Define $H_s := \operatorname{Im}(P)$ and $H_u := N(P)$. Then the spectrum of the C_0-semigroup $T_s(t) := PT(t)$ on H_s satisfies

$$\sigma(T_s(1)) \subset \{\lambda \in \mathbb{C} \mid |\lambda| < e^{g_1}\},$$

and thus we get that the spectral radius of $T_s(1)$ is less than 1. Therefore, $T_s(t)$ is an exponentially stable C_0-semigroup on H_s, which proves part 2. Similarly, the spectrum of the C_0-semigroup $T_u(t) := (I-P)T(t)$ on H_u satisfies

$$\sigma(T_u(1)) \subset \{\lambda \in \mathbb{C} \mid e^{g_2} < |\lambda|\},$$

and thus $0 \notin \sigma(T_u(1))$. Then Pazy [22, Theorem 6.5] shows that $T_u(t)$ can be embedded in a C_0-group, and so part 3 follows. Since

$$\sigma(T_u(-1)) = \sigma((T_u(1))^{-1}) \subset \{\lambda \in \mathbb{C} \mid |\lambda| < e^{-g_2}\},$$

we get that the spectral radius of $T_u(-1)$ is less than 1. Thus $-A_u$ generates an exponentially stable C_0-semigroup on H_u, and so part 4 follows.

Lemma 4.4 now shows that $(I-\tilde{P})B$ is an admissible control operator for $T_u(t)$. Thus it remains to prove that $\Sigma(A_u, (I-\tilde{P})B)$ is exactly controllable in finite time. Since $T_F(t)$ and $T_u(-t)$ are exponentially stable, the operator $T_u(-t)(I-P)T_F(t)|_{H_u} \in \mathcal{L}(H_u)$ satisfies the estimate

$$\|T_u(-t)(I-P)T_F(t)|_{H_u}\| \leq Ce^{-\tau t}, \quad t \geq 0,$$

for some $C, \tau > 0$. The proof that $\Sigma(A_u, (I-\tilde{P})B)$ is exactly controllable in finite time now exactly follows the proof of Theorem 5.10. □

6. More information on the spectrum of A.

In this section, we show that the spectrum of the generator A has to be of a special form for the system $\Sigma(A,B)$ to be optimizable. In Proposition 3.5 we saw already that for any LQ-stabilizing feedback F, the spectrum of A in the right-half plane $\mathbb{C}^+_{g_b^F}$ consists only of point spectrum with finite multiplicity and contains no finite accumulation point. We now prove that the eigenvalues cannot go too slowly to infinity if the system $\Sigma(A,B)$ is optimizable.

Let λ_0 be an isolated point of $\sigma(A)$ and an eigenvalue of A. Then $P_{\lambda_0} \in \mathcal{L}(H)$, given by

$$P_{\lambda_0}x := \frac{1}{2\pi i}\int_\Gamma (\lambda I - A)^{-1}x\, d\lambda,$$

where Γ is a simple closed contour in $\mathbb{C}$ with λ_0 the only point of $\sigma(A)$ in its interior and no points of $\sigma(A)$ on Γ, is a projection onto the spectral subspace corresponding to λ_0; see [3, Lemma 2.5.7]. The dimension of $P_{\lambda_0}H$ is called the *algebraic multiplicity*, denoted by $m_a(\lambda_0, A)$, and the *geometric multiplicity* is given by $m_g(\lambda_0, A) := \dim(N(\lambda_0 I - A))$. If there exists a number $p(\lambda_0, A) > 0$ such that $(A - \lambda_0 I)^{p(\lambda_0,A)-1}P_{\lambda_0} \neq 0$, while $(A-\lambda_0 I)^p P_{\lambda_0} = 0$ for all $p \geq p(\lambda_0, A)$, then the point λ_0 is called *pole of* $(\cdot I - A)^{-1}$ *of order* $p(\lambda_0, A)$. In [1, p. 73] it is shown that

$$\max\{m_g(\lambda_0,A), p(\lambda_0,A)\} \leq m_a(\lambda_0,A) \leq p(\lambda_0,A)m_g(\lambda_0,A) \tag{33}$$

holds. The main result of this section is the following theorem.

THEOREM 6.1. *Assume that the system $\Sigma(A,B)$ is optimizable. Then for any LQ-stabilizing feedback F, the spectral subset $\sigma^+_{g_b^F}(A) = \{\lambda_n\}_{n\in\mathbb{N}}$ satisfies*

$$m_g(\lambda_n, A) \leq m, \qquad n \in \mathbb{N},$$

where m is the dimension of the input space, and for every $g > g_b^F$ we have

$$\sum_{n\in\mathbb{N},\mathrm{Re}(\lambda_n)>g} \frac{m_a(\lambda_n, A)}{|\lambda_n - g|^2} < \infty. \tag{34}$$

Moreover, the finite-dimensional systems $\Sigma(A\,|_{P_{\lambda_n}H}, \tilde{P}_{\lambda_n}B)$ are controllable.

In [7], Fattorini proved a result that is similar to equation (34). He considers systems $\Sigma(A, B)$ with bounded control operators and assumes that there exists a nonzero holomorphic function f, being the Laplace transform of a function with compact support, such that for every $x \in \mathrm{Im}\,(f(A))$ there exists an input u, which may be a σ-additive measure with compact support, such that

$$x = \int_0^t T(t-\rho)Bu(d\rho).$$

In this situation Fattorini proved that

$$\sum_{\lambda\in\sigma(A)} \frac{m_a(\lambda, A)\mathrm{Re}(\lambda)}{1+|\lambda|^2} < \infty.$$

For the proof of Theorem 6.1 we need the following lemma.

Lemma 6.2. *Let λ be an isolated point of $\sigma(A)$ and an eigenvalue of A. Then $\overline{\lambda}$ is an isolated point of $\sigma(A^*)$ and an eigenvalue of A^* with $m_g(\overline{\lambda}, A^*) = m_g(\lambda, A)$, $m_a(\overline{\lambda}, A^*) = m_a(\lambda, A)$, and $p(\overline{\lambda}, A^*) = p(\lambda, A)$.*

Proof. The proof follows directly from Kato [18, Remark III 6.23]. □

Proof of Theorem 6.1. In Proposition 3.5 it is shown that $\sigma^+_{g_b^F}(A)$ consists only of point spectrum with finite multiplicity and contains no finite accumulation point in $\sigma^+_{g_b^F}(A)$. Thus we can write $\sigma^+_{g_b^F}(A)$ as $\{\lambda_n\}_{n\in\mathbb{N}}$. That the finite-dimensional systems $\Sigma(A\,|_{P_{\lambda_n}H}, \tilde{P}_{\lambda_n}B)$ are controllable follows directly from [28].

1. We prove that $m_g(\lambda_n, A) \le m$ for all $n \in \mathbb{N}$. Let $n \in \mathbb{N}$. By Proposition 3.6, for every $x_0 \in H$ there exist a vector $\xi \in H$ and $\omega \in \mathbb{C}^m$ such that

$$x_0 = (\lambda_n I - A)\xi - B\omega.$$

Since $\tilde{P}_{\lambda_n}$ commutes with A, this implies

$$P_{\lambda_n}x_0 = (\lambda_n I - A)P_{\lambda_n}\xi - \tilde{P}_{\lambda_n}B\omega$$

and therefore $(\lambda_n I - A)P_{\lambda_n}H + \tilde{P}_{\lambda_n}B\mathbb{C}^m \supset P_{\lambda_n}H$. Thus we have

$$\begin{aligned} m_a(\lambda_n, A) &= \dim(P_{\lambda_n}H) \\ &\le \dim((\lambda_n I - A)P_{\lambda_n}H) + \dim(\tilde{P}_{\lambda_n}B\mathbb{C}^m) \\ &\le m_a(\lambda_n, A) - m_g(\lambda_n, A) + m, \end{aligned}$$

which shows $m_g(\lambda_n, A) \le m$.

2. We prove

$$\sum_{n\in\mathbb{N},\mathrm{Re}(\lambda_n)>g} \frac{m_a(\lambda_n, A)(\mathrm{Re}(\lambda_n) - g)}{1+|\lambda_n - g|^2} < \infty$$

for every $g > g_b^F$. Let $g > g_b^F$ and $\tau \in (g_b^F, g)$. Then there exists $\alpha \in \rho(A)$ with $\mathrm{Re}(\alpha) \in (\tau, g)$. We write B as $B = [b_1, \dots, b_m]$. By Proposition 3.6, for every $k \in \{1, \dots, m\}$ there exist $\xi_k \in \boldsymbol{H}_\infty(\mathbb{C}_\tau^+; H)$ and $\omega_k \in \boldsymbol{H}_\infty(\mathbb{C}_\tau^+; \mathbb{C}^m)$ such that

$$(\alpha I - A)^{-1} b_k = (sI - A)\xi_k(s) - B\omega_k(s), \qquad s \in \mathbb{C}_\tau^+.$$

Defining $\Xi(s) := [\xi_1(s), \dots, \xi_m(s)]$ and $\Omega(s) := [\omega_1(s), \dots, \omega_m(s)]$ we obtain

$$(\alpha I - A)^{-1} B = (sI - A)\Xi(s) - B\Omega(s), \qquad s \in \mathbb{C}_\tau^+. \tag{35}$$

The functions Ξ and Ω also are holomorphic and bounded on $\mathbb{C}_\tau^+$.
Multiplying (35) by $(sI - A)^{-1}$ and using the resolvent identity, we obtain

$$\frac{1}{s-\alpha}(\alpha I - A)^{-1} B = \Xi(s) - (sI - A)^{-1} B \left[\Omega(s) - \frac{1}{s-\alpha} I\right], \qquad s \in \mathbb{C}_g^+ \backslash \sigma(A).$$

Since the term on the left-hand side and the first term of the right-hand side are holomorphic on $\mathbb{C}_g^+$, the function $(sI - A)^{-1} B[\Omega(s) - \frac{1}{s-\alpha} I]$ also is holomorphic on $\mathbb{C}_g^+$. Defining

$$F(s) := \left[\Omega(s) - \frac{1}{s-\alpha} I\right], \qquad s \in \mathbb{C}_g^+, \tag{36}$$

it is easy to see that $F(s)^{-1} \det(F(s))$ is holomorphic on $\mathbb{C}_g^+$. Thus

$$\begin{aligned} G(s) &:= (sI - A)^{-1} B \det(F(s)) \\ &= (sI - A)^{-1} B \left[\Omega(s) - \frac{1}{s-\alpha} I\right] \cdot F(s)^{-1} \det(F(s)) \end{aligned}$$

is holomorphic on $\mathbb{C}_g^+$.
Next we show that every $\lambda \in \sigma_g^+(A)$ is a zero of $\det(F(s))$ with order at least $p(\lambda, A)$. Let us assume that this is not true. Then there is one $\lambda_0 \in \sigma_g^+(A)$ such that λ_0 is no zero of $\det(F(s))$ or λ_0 is a zero of $\det(F(s))$, but the order of λ_0 as zero of $\det(F(s))$ is less than $p := p(\lambda_0, A)$. By $f^{(i)}$ we denote the ith derivative of f. Let $k_0 \in \{0, \dots, p-1\}$ be such that $\det(F(\lambda_0))^{(i)} = 0$, $i = 0, \dots, k_0 - 1$, while $\det(F(\lambda_0))^{(k_0)} \neq 0$ and let $\delta > 0$ be such that $B(\lambda_0) := \{s \in \mathbb{C} \mid 0 < |s - \lambda_0| < \delta\} \subset \rho(A)$. Since $\tilde{P}_{\lambda_0}$ commutes with the resolvent $(\cdot I - A)^{-1}$, by Lemma 4.2 we obtain that

$$\begin{aligned} G_{\lambda_0}(s) &:= P_{\lambda_0} G(s) \\ &= (sI - A|_{P_{\lambda_0} H})^{-1} \tilde{P}_{\lambda_0} B \det(F(s)) \end{aligned}$$

is holomorphic on $\mathbb{C}_g^+$. Since $(sI - A|_{P_{\lambda_0} H})^{-1} = (sI - A)^{-1}$ on $P_{\lambda_0} H$ (see Lemma 4.2), we get $p = p(\lambda_0, A|_{P_{\lambda_0} H})$. Expressing $(\cdot I - A|_{P_{\lambda_0} H})^{-1}$ in its Laurent series, we obtain

$$G_{\lambda_0}(s) = \tilde{G}_{\lambda_0}(s) + \sum_{k=0}^{p-1} \frac{\det(F(s))}{(s - \lambda_0)^{k+1}} (\lambda_0 I - A|_{P_{\lambda_0} H})^k \tilde{P}_{\lambda_0} B, \qquad s \in B(\lambda_0),$$

where $\tilde{G}_{\lambda_0}(s)$ is a holomorphic function on $B(\lambda_0) \cup \{\lambda_0\}$. Since λ_0 is a zero of $\det(F(s))$ with order k_0, we get

$$\sum_{k=0}^{k_0-1} \frac{\det(F(s))}{(s-\lambda_0)^{k+1}} (\lambda_0 I - A|_{P_{\lambda_0}H})^k \tilde{P}_{\lambda_0} B$$

is holomorphic at λ_0. Since $G_{\lambda_0}(s)$ is holomorphic at λ_0, this would imply that

$$\sum_{k=k_0}^{p-1} \frac{\det(F(s))}{(s-\lambda_0)^{k+1}} (\lambda_0 I - A|_{P_{\lambda_0}H})^k \tilde{P}_{\lambda_0} B$$

is holomorphic on $\mathbb{C}_g^+$. Thus using the definition of a pole, we get that

$$\begin{aligned}&(\lambda_0 I - A|_{P_{\lambda_0}H})^{p-k_0-1} \sum_{k=k_0}^{p-1} \frac{\det(F(s))}{(s-\lambda_0)^{k+1}} (\lambda_0 I - A|_{P_{\lambda_0}H})^k \tilde{P}_{\lambda_0} B \\ &= \frac{\det(F(s))}{(s-\lambda_0)^{k_0+1}} (\lambda_0 I - A|_{P_{\lambda_0}H})^{p-1} \tilde{P}_{\lambda_0} B\end{aligned}$$

is holomorphic at λ_0. However, this can only happen if $(\lambda_0 I - A|_{P_{\lambda_0}H})^{p-1}\tilde{P}_{\lambda_0}B = 0$. By the definition of a pole there is a $x \in P_{\lambda_0}H$ such that $(\lambda_0 I - A|_{P_{\lambda_0}H})^{p-1}x \neq 0$. The optimizability of system $\Sigma(A,B)$ implies that there are elements $y \in H$ and $u \in \mathbb{C}^m$ such that

$$x = (\lambda_0 I - A)y - Bu.$$

Since $x \in P_{\lambda_0}H$, this implies $x = (\lambda_0 I - A|_{P_{\lambda_0}H})P_{\lambda_0}y - \tilde{P}_{\lambda_0}Bu$. Multiplying both sides with $(\lambda_0 I - A|_{P_{\lambda_0}H})^{p-1}$, we get

$$0 \neq (\lambda_0 I - A|_{P_{\lambda_0}H})^{p-1}x = (\lambda_0 I - A|_{P_{\lambda_0}H})^{p}y - (\lambda_0 I - A|_{P_{\lambda_0}H})^{p-1}\tilde{P}_{\lambda_0}Bu = 0,$$

which is a contradiction. Thus

$$f(s) := \det(F(s)), \quad s \in \mathbb{C}_g^+, \tag{37}$$

is holomorphic and bounded on $\mathbb{C}_g^+$ and every $\lambda \in \mathbb{C}_g^+ \cap \sigma(A)$ is a zero of f with order at least $p(\lambda, A)$.

Next we show that f is not identically zero. Assume that f is identical to zero on $\mathbb{C}_g^+$. Using (36), (37),

$$f(s) = \det\left((s-\alpha)\Omega(s) - I\right) \det\left(\frac{1}{s-\alpha} I\right), \quad s \in \mathbb{C}_g^+,$$

it would imply that

$$\det\left((s-\alpha)\Omega(s) - I\right) = 0, \quad s \in \mathbb{C}_g^+.$$

Now the holomorphicity of $\det\left((s-\alpha)\Omega(s) - I\right)$ on $\mathbb{C}_\tau^+$ implies

$$0 = \det\left((\alpha-\alpha)\Omega(\alpha) - I\right) = \det(I) = 1,$$

which is a contradiction. Thus $f \not\equiv 0$ on $\mathbb{C}_g^+$.
Applying now Duren [6, Theorem 11.3] to $f \in \boldsymbol{H}_\infty(\mathbb{C}_g^+)$ shows

$$\sum_{n\in\mathbb{N},\mathrm{Re}(\lambda_n)>g} \frac{p(\lambda_n, A)(\mathrm{Re}(\lambda_n) - g)}{1 + |\lambda_n - g|^2} < \infty.$$

Therefore, using (33) and $m_g(\lambda_n, A) \leq m$, we get

$$\sum_{n\in\mathbb{N},\mathrm{Re}(\lambda_n)>g} \frac{m_a(\lambda_n, A)(\mathrm{Re}(\lambda_n) - g)}{1 + |\lambda_n - g|^2} < \infty.$$

3. It remains to prove that

$$\sum_{n\in\mathbb{N},\mathrm{Re}(\lambda_n)>g} \frac{m_a(\lambda_n, A)}{|\lambda_n - g|^2} < \infty$$

for every $g > g_b^F$. Let $k_n := m_a(\lambda_n, A)$, $g > g_b^F$ and $\tilde{g} \in (g_b^F, g)$. By part (2), we have

$$\sum_{n\in\mathbb{N},\mathrm{Re}(\lambda_n)>g} \frac{k_n(\mathrm{Re}(\lambda_n) - g)}{1 + |\lambda_n - g|^2} < \infty \quad \text{and} \quad \sum_{n\in\mathbb{N},\mathrm{Re}(\lambda_n)>g} \frac{k_n(\mathrm{Re}(\lambda_n) - \tilde{g})}{1 + |\lambda_n - \tilde{g}|^2} < \infty.$$

For any $\lambda \in \mathbb{C}$ with $|\lambda - g| \geq 1$ we get

$$\begin{aligned}
&\frac{1}{1 + |\lambda - g|^2} \\
&= \frac{1}{1 + |\lambda - \tilde{g}|^2} \frac{1 + |\lambda - \tilde{g}|^2}{1 + |\lambda - g|^2} \leq \frac{1}{1 + |\lambda - \tilde{g}|^2} \frac{1 + (|\lambda - g| + |g - \tilde{g}|)^2}{1 + |\lambda - g|^2} \\
&\leq \frac{1}{1 + |\lambda - \tilde{g}|^2} \frac{1 + |\lambda - g|^2(1 + g - \tilde{g})^2}{1 + |\lambda - g|^2} = \frac{(1 + g - \tilde{g})^2}{1 + |\lambda - \tilde{g}|^2}.
\end{aligned} \tag{38}$$

Let $J_1 := \{n \in \mathbb{N} \mid \lambda_n \in \sigma_g^+(A) \text{ and } |\lambda_n - g| \geq 1\}$ and $J_2 := \{n \in \mathbb{N} \mid \lambda_n \in \sigma_g^+(A) \text{ and } |\lambda_n - g| < 1\}$. Since J_2 consists only of finitely many points, we get

$$\sum_{n\in J_2} \frac{k_n}{|\lambda_n - g|^2} < \infty.$$

Thus the statement follows directly from the calculation

$$\begin{aligned}
&\sum_{n\in J_1} \frac{k_n}{|\lambda_n - g|^2} \\
&\leq 2 \sum_{n\in J_1} \frac{k_n}{1 + |\lambda_n - g|^2} \\
&= \frac{2}{g - \tilde{g}} \left[\sum_{n\in J_1} k_n \frac{\mathrm{Re}(\lambda_n - \tilde{g})}{1 + |\lambda_n - g|^2} - \sum_{n\in J_1} k_n \frac{\mathrm{Re}(\lambda_n - g)}{1 + |\lambda_n - g|^2} \right] \\
&\overset{(38)}{\leq} \frac{2(1 + g - \tilde{g})^2}{g - \tilde{g}} \sum_{n\in J_1} k_n \frac{\mathrm{Re}(\lambda_n - \tilde{g})}{1 + |\lambda_n - \tilde{g}|^2} - \frac{2}{g - \tilde{g}} \sum_{n\in J_1} k_n \frac{\mathrm{Re}(\lambda_n - g)}{1 + |\lambda_n - g|^2} \\
&< \infty. \qquad \square
\end{aligned}$$

The next proposition gives a simple necessary condition for optimizability of a system $\Sigma(A,B)$.

PROPOSITION 6.3. *Assume that the system $\Sigma(A,B)$ is optimizable, and let F be an LQ-stabilizing feedback. Then for every $g > g_b^F$ there exist constants $M_1, M_2 > 0$ such that*

$$M_1 \leq \|B^*\psi\| \leq M_2$$

for every $\psi \in N(\overline{\lambda}I - A^*)$ with $\|\psi\| = 1$, where λ is an eigenvalue of A with $\mathrm{Re}(\lambda) > g$.

Proof. Let $g > g_b^F$. Using Proposition 3.6 there exists $M_1' > 0$, such that for every $\psi \in H$ with $\|\psi\| = 1$ and every $\lambda \in \mathbb{C}_g^+$ there is a $\xi_\psi \in H$ and $\omega_\psi \in \mathbb{C}^m$ such that

$$\psi = (\lambda I - A)\xi_\psi - B\omega_\psi \tag{39}$$

and $\|\xi_\psi\|_H, \|\omega_\psi\| \leq M_1'$. We now choose $\psi \in N(\overline{\lambda}I - A^*)$ with $\|\psi\| = 1$, where λ is an eigenvalue of A with $\mathrm{Re}(\lambda) > g$. Note that by Lemma 6.2 there exists such a ψ. Taking the inner product of (39) with ψ, we obtain

$$1 = \langle \xi_\psi, (\overline{\lambda}I - A^*)\psi\rangle - \langle B\omega_\psi, \psi\rangle = -\langle \omega_\psi, B^*\psi\rangle$$

and thus

$$1 \leq M_1'\|B^*\psi\|.$$

Next we prove that there exists a constant $M_2 > 0$ such that $\|B^*\psi\| \leq M_2$. Since B is an admissible control operator for $T(t)$, there exist $t_1 > 0$ and $M > 0$ such that

$$\left\| \int_0^{t_1} T(t_1 - \rho)Bu(\rho)\, d\rho \right\| \leq M\|u\|_{\boldsymbol{L}_2(0,t_1;\mathbb{C}^m)}$$

for every $u \in \boldsymbol{L}_2(0,t_1;\mathbb{C}^m)$. Let $K := \sup_{\lambda\in\sigma(A)} \mathrm{Re}(\lambda) < \infty$ and define

$$f(\mu) := \begin{cases} \frac{1-e^{-2\mu t_1}}{2\mu}, & \mu \in [g,K]\backslash\{0\}, \\ t_1, & \mu = 0. \end{cases}$$

By l'Hôspital's rule, f is continuous on $[g,K]$ and so

$$\sup_{\mu\in[g,K]} |f(\mu)| < \infty.$$

Let $\psi \in N(\overline{\lambda}I - A^*)$ with $\|\psi\| = 1$, where λ is an eigenvalue of A with $\mathrm{Re}(\lambda) > g$. We define $u_\lambda \in \boldsymbol{L}_2(0,t_1;\mathbb{C})$ by

$$u_\lambda(\rho) := \frac{1}{t_1}e^{-\lambda(t_1-\rho)}, \qquad \rho \in [0,t_1].$$

This implies

$$\begin{aligned}\|B^*\psi\| &= \sup_{\|u\|=1} |\langle u, B^*\psi\rangle| = \sup_{\|u\|=1} |\langle Bu, \psi\rangle| = \sup_{\|u\|=1} \left| \int_0^{t_1} u_\lambda(\rho)e^{\lambda(t_1-\rho)}\, d\rho \langle Bu, \psi\rangle \right| \\ &= \sup_{\|u\|=1} \left| \int_0^{t_1} \langle Buu_\lambda(\rho), e^{\overline{\lambda}(t_1-\rho)}\psi\rangle\, d\rho \right| = \sup_{\|u\|=1} \left| \int_0^{t_1} \langle Buu_\lambda(\rho), T^*(t_1-\rho)\psi\rangle\, d\rho \right|\end{aligned}$$

$$
\begin{aligned}
&= \sup_{\|u\|=1} \left| \int_0^{t_1} \langle T(t_1-\rho)Buu_\lambda(\rho), \psi\rangle \, d\rho \right| = \sup_{\|u\|=1} \left| \left\langle \int_0^{t_1} T(t_1-\rho)Buu_\lambda(\rho)\, d\rho, \psi \right\rangle \right| \\
&\le \sup_{\|u\|=1} \left\| \int_0^{t_1} T(t_1-\rho)Buu_\lambda(\rho)\, d\rho \right\| \le \sup_{\|u\|=1} M\|uu_\lambda\|_{\boldsymbol{L}_2(0,t_1;\mathbb{C}^m)} \\
&\le \frac{M}{t_1} f(\mathrm{Re}(\lambda))^{1/2} \le \frac{M}{t_1} \sup_{\mu\in[g,K]} |f(\mu)|^{1/2} := M_2. \qquad \square
\end{aligned}
$$

7. More information on the spaces $\boldsymbol{H_u}$ and $\boldsymbol{H_s}$. In Theorem 5.3 we saw that every optimizable system $\Sigma(A,B)$ with A satisfying SDA can be split into an exponentially stable part and an exactly controllable part. We now give a precise description of the spaces H_u and H_s appearing in Theorem 5.3. Note that by the definition of H_u and H_s we have $H = H_u \oplus H_s$. The main result of this section is the following theorem.

THEOREM 7.1. *Suppose $\Sigma(A,B)$ is optimizable and let A satisfy the SDA. Then the spaces H_u and H_s in Theorem* 5.3 *are given by*

$$
\begin{aligned}
H_u &= \overline{\mathrm{span}_{n\in\mathbb{N}} P_{\lambda_n} H}, \\
H_s &= \{x \in H | T(\cdot)x \in \boldsymbol{L}_2(0,\infty;H)\},
\end{aligned}
$$

where $\sigma(A)\cap\mathbb{C}_0^+ = \{\lambda_n\}_{n\in\mathbb{N}}$ (see Theorem 6.1).

In order to prove this theorem we need a series of lemmas.

LEMMA 7.2. *Suppose that $\Sigma(A,B)$ is optimizable and let A satisfy the SDA. Then there exists a constant $C>0$ such that*

$$\|(\cdot I - A)^{-1}x\|_{\boldsymbol{L}_2(i\mathbb{R};H)} \le C\|x\|$$

for every $x\in H$.

Proof. Let F be an LQ-stabilizing feedback, and choose a negative σ larger than g_b^F. Then by Proposition 3.6 there exists a constant $\tilde{C}>0$ such that for every $x\in H$ there exist $\xi_x \in \boldsymbol{H}_2(\mathbb{C}_\sigma^+, H)$ and $\omega_x \in \boldsymbol{H}_2(\mathbb{C}_\sigma^+, \mathbb{C}^m)$ such that

$$(itI-A)^{-1}x = \xi_x(it) - (itI-A)^{-1}B\omega_x(it), \quad t\in\mathbb{R}, \tag{40}$$

$\|\xi_x\|_{\boldsymbol{H}_2(\mathbb{C}_\sigma^+,H)} \le \tilde{C}\|x\|_H$ and $\|\omega_x\|_{\boldsymbol{H}_2(\mathbb{C}_\sigma^+,\mathbb{C}^m)} \le \tilde{C}\|x\|_H$. Using Corollary 5.13, we obtain $(\cdot I-A)^{-1}B \in \boldsymbol{L}_\infty(i\mathbb{R};\mathcal{L}(\mathbb{C}^m,H))$ and thus for $x\in H$ (40) implies

$$
\begin{aligned}
&\|(\cdot I - A)^{-1}x\|_{\boldsymbol{L}_2(i\mathbb{R};H)} \\
&\le \|\xi_x(\cdot)\|_{\boldsymbol{L}_2(i\mathbb{R};H)} + \|(\cdot I-A)^{-1}B\|_{\boldsymbol{L}_\infty(i\mathbb{R};\mathcal{L}(U,H))}\|\omega_x(\cdot)\|_{\boldsymbol{L}_2(i\mathbb{R};\mathbb{C}^m)} \\
&\le C\|x\|_H
\end{aligned}
$$

for some $C>0$ independent of x. $\square$

LEMMA 7.3. *Suppose that $\Sigma(A,B)$ is exactly controllable in finite time and that A generates a C_0-group $T(t)$ on H. Then*

$$H = \overline{\mathrm{span}_{n\in\mathbb{N}} P_{\lambda_n} H},$$

where λ_n are the eigenvalues of A.

Proof. For a C_0-group there is a real constant γ such that $-A+\gamma I$ generates an exponentially stable C_0-semigroup. Since the assumptions made in the theorem also hold for $A-\gamma I$, we may without loss of generality assume that $-A$ generates an exponentially stable C_0-semigroup. By Theorem 3.10 we have that the spectrum of A is purely point spectrum without (finite) accumulation points. We denote this spectrum by $\{\lambda_n\}$.

Define $V := \overline{\operatorname{span}_{n\in\mathbb{N}} P_{\lambda_n} H}$ and $W := V^\perp$, and assume that $W \neq \emptyset$. It is easy to see that V is $T(t)$-invariant. By Lemmas 4.2 and 4.3 we obtain that the system $\Sigma(A_W, B_W)$ is exactly controllable in finite time. The corresponding C_0-semigroup on W is given by

$$T_W(t)x = P_W T(t)x, \qquad x \in W. \tag{41}$$

Since W is the orthogonal complement of V, it follows that W is $T(t)^*$-invariant, and

$$T_W(t)^* x = T(t)^* x, \qquad x \in W.$$

The generator of $T_W(t)$ is given by $A_W^* := A^*|_W$. From the fact that the spectrum of A consists of only point spectrum with finite multiplicity and without an accumulation point, it follows that $\sigma(A^*) = \{\bar{\lambda}_n\}_n$, and $\rho_\infty(A^*) = \rho(A^*)$. Using part (1) of Lemma 4.2 and (41), we have that

$$\sigma(A_W^*) \subset \sigma(A^*) = \{\bar{\lambda}_n\}_n.$$

Let $\varepsilon > 0$ be arbitrary. If $\sigma_{-\varepsilon}^+(A_W^*) \neq \emptyset$, then there exist a $x \in W$ and an $n \in \mathbb{N}$ such that $(\cdot I - A_W^*)^{-1}x$ is not holomorphic at $\overline{\lambda}_n$ [18, p. 174]. Since $(sI - A_W^*)^{-1}x = (sI - A^*)^{-1}x$, $s \in \rho(A^*)$, we get $(\cdot I - A^*)^{-1}x$ is not holomorphic at $\overline{\lambda}_n$. This implies that

$$\int_\Gamma (sI - A^*)^{-1} x \, ds \neq 0,$$

where Γ is a simple closed contour in $\mathbb{C}$ with only one point of $\sigma(A^*)$ in its interior, namely $\overline{\lambda}_n$, and no points of $\sigma(A^*)$ on Γ. Now there is a $y \in H$ such that

$$\langle P_{\lambda_n} y, x\rangle = \left\langle y, \int_\Gamma (sI - A^*)^{-1} x \, ds \right\rangle \neq 0.$$

This is in contradiction with $P_{\lambda_n} y \in V$ and $x \in W = V^\perp$. Thus $\sigma_{-\varepsilon}^+(A_W^*) = \emptyset$, and therefore $\sigma_{-\varepsilon}^+(A_W) = \emptyset$. Since $\Sigma(A_W, B_W)$ is exactly controllable in finite time it is LQ-stabilizable as well. Thus Theorem 5.14 shows that there is a constant $M > 0$ such that

$$\|T_W(t)\| \leq M, \qquad t \geq 0.$$

This implies

$$\|T(-t)\| \geq \|T_W(-t)\| \geq \frac{1}{\|T_W(t)\|} \geq M^{-1}, \qquad t \geq 0,$$

which is in contradiction with the exponential stability of $-A$. Thus the assumption $W \neq \emptyset$ does not hold, and so we get $H = \overline{\operatorname{span}_{n\in\mathbb{N}} P_{\lambda_n} H}$. □

Since $\boldsymbol{H}_2(\mathbb{C}_0^+; H)$ is a closed subset of $\boldsymbol{L}_2(i\mathbb{R}; H)$, we can write

$$\boldsymbol{L}_2(i\mathbb{R}; H) = \boldsymbol{H}_2(\mathbb{C}_0^+; H) \oplus \boldsymbol{H}_2(\mathbb{C}_0^+; H)^\perp, \tag{42}$$

where $\boldsymbol{H}_2(\mathbb{C}_0^+;H)^\perp$ is the orthogonal complement in $\boldsymbol{L}_2(i\mathbb{R};H)$ of $\boldsymbol{H}_2(\mathbb{C}_0^+;H)$. In [3, Theorem A.6.22] it is shown that

$$\boldsymbol{H}_2(\mathbb{C}_0^+;H)^\perp = \left\{f : \mathbb{C}_0^- \to H | f(\,-\,\cdot) \in \boldsymbol{H}_2(\mathbb{C}_0^+;H)\right\}. \tag{43}$$

Proof of Theorem 7.1. By Theorem 5.3 there exists a projection $P \in \mathcal{L}(H)$ such that (1)–(5) of Theorem 5.3 are satisfied. Lemma 7.2 shows that $(\cdot I - A)^{-1}x \in \boldsymbol{L}_2(i\mathbb{R};H)$ for every $x \in H$. If we define

$$\tilde{H}_s = \{x \in H \mid (\cdot I - A)^{-1}x \in \boldsymbol{H}_2(\mathbb{C}_0^+;H)\} \quad \text{and}$$
$$\tilde{H}_u = \{x \in H \mid (\cdot I - A)^{-1}x \in \boldsymbol{H}_2(\mathbb{C}_0^+;H)^\perp\},$$

then (42) shows that

$$\tilde{H}_s \cap \tilde{H}_u = \{0\}.$$

1. We now prove that $\tilde{H}_s$ and $\tilde{H}_u$ are closed, linear subspaces of H. We will prove this result only for $\tilde{H}_s$, as the proof for $\tilde{H}_u$ is very similar. That $\tilde{H}_s$ is a linear subspace of H follows immediately from the fact that $\boldsymbol{H}_2(\mathbb{C}_0^+,H)$ is a linear vector space. In order to prove the closedness we choose a sequence $\{z_n\}_n$ in $\tilde{H}_s$ which converges to $z \in H$. By Lemma 7.2 we get

$$\lim_{n\to\infty} \|(\cdot I - A)^{-1}z_n - (\cdot I - A)^{-1}z\|_{\boldsymbol{L}_2(i\mathbb{R},H)} \le C \lim_{n\to\infty} \|z_n - z\|_H = 0.$$

Since $\boldsymbol{H}_2(\mathbb{C}_0^+;H)$ is a closed subspace of $\boldsymbol{L}_2(i\mathbb{R},H)$, this implies that $(\cdot I - A)^{-1}z \in \boldsymbol{H}_2(\mathbb{C}_0^+;H)$, and so $\tilde{H}_s$ is closed.
2. We now prove that $H_u \subset \tilde{H}_u$. Since $\sigma^+_{g_b^F}(A)$ consists only of point spectrum with no (finite) accumulation point, we have that

$$\rho_\infty(A) \cap \mathbb{C}^+_{g_b^F} = \rho(A) \cap \mathbb{C}^+_{g_b^F}.$$

Since the imaginary axis is in the resolvent set, by Lemma 4.2 we have for all $t \in \mathbb{R}$ and $x \in H_u$

$$(itI - A)^{-1}x = (itI - A_u)^{-1}. \tag{44}$$

The operator $-A_u$ generates an exponentially stable semigroup on H_u, and hence for $x \in H_u$ we get that $(itI - A_u)^{-1}x$ is the Fourier transform of $-T_u(t)x$, $t \in (-\infty,0)$. This last function is in $\boldsymbol{L}_2(-\infty,0;H)$. By the Paley–Wiener theorem [3, Theorem A.6.21], (43), and (44), this implies that $(\cdot I - A)^{-1}x \in \boldsymbol{H}_2(\mathbb{C}_0^+;H)^\perp$, and thus $x \in \tilde{H}_u$.
3. We now prove that $H_s \subset \tilde{H}_s$. Let $x \in H_s$ arbitrary. Then $T(\cdot)x \in \boldsymbol{L}_2(0,\infty;H)$ and thus the Paley–Wiener theorem [3, Theorem A.6.21] implies that its Laplace transform $(\cdot I - A)^{-1}x$ is in $\boldsymbol{H}_2(\mathbb{C}_0^+;H)$. Thus $x \in \tilde{H}_s$.
4. We now prove $H_u = \tilde{H}_u$ and $H_s = \tilde{H}_s$. We prove this result only for $\tilde{H}_s$, as the proof for $\tilde{H}_u$ is very similar. Assume that $H_s = \tilde{H}_s$ does not hold. Then there exists an element $x \in \tilde{H}_s \backslash H_s$. We can write x as $x = x_u + x_s$ with $x_u \in H_u$ and $x_s \in H_s$. Since $x_u = x - x_s \in \tilde{H}_s$, we obtain $x_u \in H_u \cap \tilde{H}_s \subset \tilde{H}_u \cap \tilde{H}_s = \{0\}$, which is in contradiction with $x \in \tilde{H}_s \backslash H_s$.

5. We now prove that $V \subset H_u$, where
$$V := \overline{\operatorname{span}_{n\in\mathbb{N}} P_{\lambda_n} H},$$
and $\{\lambda_n\}_n = \sigma(A) \cap \mathbb{C}_0^+$. First of all we choose $x \in P_{\lambda_n} H$ for some $n \in \mathbb{N}$. Then in [3, p. 99] it has been proven that
$$(itI - A)^{-1}x = \sum_{j=0}^{m_a(\lambda_n,A)} (-1)^j \frac{(\lambda_n I - A)^j x}{(it-\lambda_n)^{j+1}}, \qquad t \in \mathbb{R},$$
where $m_a(\lambda_n, A)$ is the algebraic multiplicity of the eigenvalue λ_n. Since $\mathrm{Re}(\lambda_n) > 0$, we have
$$\frac{1}{(\cdot - \lambda_n)^j} \in \boldsymbol{H}_2(\mathbb{C}_0^+; H)^\perp,$$
and so we obtain that $x \in \tilde{H}_u$. Now $H_u = \tilde{H}_u$ and V are closed, linear subspaces of H and thus the statement is proved.
6. In order to prove the theorem it now remains to show that $V = H_u$ holds. The system $\Sigma(A_u, (I - \tilde{P})B)$ is exactly controllable in finite time. Thus Lemma 7.3 proves $V = H_u$. □

8. An example. We now construct an example that is optimizable, but it is not possible to split the state space in a direct sum as given in Theorem 1.3. The generator in this example does not satisfy the SDA at any negative number. For the construction of this example the following lemmas will be useful. For background information on Carleson measures we refer to Garnett [9, p. 31].

LEMMA 8.1. *Let* $\{q_n\}_{n\in\mathbb{N}} = \mathbb{Q} \cap (\frac{1}{2}, \infty)$ *be chosen such that*
$$q_{k+\frac{l(l-1)}{2}} \in [k-1, k], \qquad l \in \mathbb{N}, k \in \{1, \dots, l\},$$
and define $\{\gamma_n\}_{n\in\mathbb{N}} \subset \mathbb{C}_0^+$ *by*
$$\gamma_n := q_n + in^4, \qquad n \in \mathbb{N}.$$
Then $\mu := \sum_{n\in\mathbb{N}} \mathrm{Re}\gamma_n \delta_{\gamma_n}$ *is a Carleson measure.*

Proof. In order to prove this we define
$$S(h, y_0) := \{z \in \mathbb{C}_0^+ \mid 0 < \mathrm{Re} z < h,\ y_0 < \mathrm{Im}\, z < y_0 + h\}, \qquad h > 0, y_0 \in \mathbb{R}.$$
Now μ is a Carleson measure if and only if there exists a constant $A > 0$ such that $\mu(S(h, y_0)) \le Ah$ for every $h > 0$ and all $y_0 \in \mathbb{R}$ (see, for example, [9, p. 31]).

Clearly, $\mu(S(h, y_0)) = 0$ if $h \in [0, \frac{1}{2}]$. For $h > \frac{1}{2}$ and $y_0 < -h$, we have that $\mu(S(h, y_0)) = 0$ as well.

We now consider the situation $h > \frac{1}{2}$ and $y_0 \ge h^2$. Then for every $n \in \mathbb{N}$ with $y_0 < n^4 < y_0 + h$, we get
$$(n+1)^4 > n^4 + n^2 > y_0 + \sqrt{y_0} \ge y_0 + h.$$
Thus $\#\{\{\gamma_n\}_{n\in\mathbb{N}} \cap S(h, y_0)\} \le 1$. Combining this with the fact that if $\gamma_n \in S(h, y_0)$, then q_n is less than h, shows that $\mu(S(h, y_0)) \le h$.

Finally, we have to consider the situation $h > \frac{1}{2}$, $y_0 \in (-h, h^2)$. Then

$$S(h, y_0) \subset R(h) := \left\{z \in \mathbb{C}_0^+ \mid 0 < \mathrm{Re} z < h,\ -h < \mathrm{Im}\, z < h^2 + h\right\}.$$

Using $h > \frac{1}{2}$, we see that

$$h^2 + h \leq h^2 + 2h^2 < 16h^2.$$

From this we get $\#\{\{\gamma_n\}_{n\in\mathbb{N}} \cap R(h)\} \leq 2\sqrt{h}$. We now choose $l_h \in \mathbb{N}$ such that

$$\frac{l_h(l_h-1)}{2} < 2\sqrt{h} < \frac{l_h(l_h+1)}{2}.$$

This implies immediately

$$l_h \leq 1 + 4h^{1/4} < 6h^{1/4},$$

where we have used that $h > \frac{1}{2}$. Thus

$$\mu(S(h, y_0)) \leq \mu(R(h)) \leq \sum_{n=1}^{[2\sqrt{h}]} q_n \leq \sum_{l=1}^{l_h} \sum_{k=1}^{l} k = \frac{1}{6} l_h(l_h+1)(l_h+2) \leq Ah^{3/4},$$

where $A > 0$ is independent of h. Thus μ is a Carleson measure. □

Lemma 8.2. *Let* $\{q_n\}_{n\in\mathbb{N}} = \mathbb{Q} \cap (\frac{1}{2}, \infty)$ *be chosen as in Lemma* 8.1 *and define* $\{\beta_n\}_{n\in\mathbb{N}}$ *by*

$$\beta_{2n} := \gamma_n, \qquad \beta_{2n-1} := \gamma_n + \frac{1}{n}, \qquad n \in \mathbb{N}.$$

Moreover, let $\Theta \in \boldsymbol{H}_\infty(\mathbb{C}_0^+)$ *be the Blaschke product corresponding to the zeros* $\{\beta_n\}$, *i.e.,*

$$\Theta(s) := \prod_{n=1}^{\infty} \frac{|1-\beta_n^2|}{1-\beta_n^2} \frac{s-\beta_n}{s+\bar{\beta}_n}, \quad s \in \mathbb{C}_0^+.$$

Then

$$\sup_{0<\mathrm{Re}s<1/4} |\Theta^{-1}(s)| < \infty. \tag{45}$$

Proof. For $x \in (0, \frac{1}{4})$, $y \in \mathbb{R}$, we get

$$\begin{aligned}|\Theta(x+iy)^{-2}| &= \prod_{n\in\mathbb{N}} \frac{(x+\mathrm{Re}\beta_n)^2 + (y-\mathrm{Im}\,\beta_n)^2}{(x-\mathrm{Re}\beta_n)^2 + (y-\mathrm{Im}\,\beta_n)^2} \\ &= \prod_{n\in\mathbb{N}} \left(1 + \frac{4x\mathrm{Re}\beta_n}{(x-\mathrm{Re}\beta_n)^2 + (y-\mathrm{Im}\,\beta_n)^2}\right).\end{aligned}$$

Thus (45) is equivalent to

$$\sup_{x\in(0,\frac{1}{4})} \sup_{y\in\mathbb{R}} \sum_{n\in\mathbb{N}} \ln\left(1 + \frac{4x\mathrm{Re}\beta_n}{(x-\mathrm{Re}\beta_n)^2 + (y-\mathrm{Im}\,\beta_n)^2}\right) < \infty. \tag{46}$$

Using $\ln(1+z) \leq z$, $z \geq 0$, and

$$0 \leq \frac{4x\mathrm{Re}\beta_n}{(x-\mathrm{Re}\beta_n)^2 + (y - \mathrm{Im}\,\beta_n)^2},$$

it is easy to see that

$$\sup_{x\in(0,\frac{1}{4})} \sup_{y\in\mathbb{R}} \sum_{n\in\mathbb{N}} \frac{x\mathrm{Re}\beta_n}{(x-\mathrm{Re}\beta_n)^2 + (y - \mathrm{Im}\,\beta_n)^2} < \infty \tag{47}$$

implies (46). Thus it remains to prove that

$$h(x,y) := \sum_{n\in\mathbb{N}} \frac{x\mathrm{Re}\beta_n}{(x-\mathrm{Re}\beta_n)^2 + (y - \mathrm{Im}\,\beta_n)^2}$$

is uniformly bounded on $\mathbb{C}_0^+ \cap \mathbb{C}_{1/4}^-$.

Let $x + iy \in \mathbb{C}_0^+ \cap \mathbb{C}_{1/4}^-$ be arbitrary. Define $z := \frac{1}{3} - x$. From $\mathrm{Re}\beta_n > \frac{1}{2}$, we get

$$\begin{aligned} h(x,y) &= \sum_{n\in\mathbb{N}} \frac{(\frac{1}{3}-z)\mathrm{Re}\beta_n}{(\frac{1}{3}-z-\mathrm{Re}\beta_n)^2 + (y - \mathrm{Im}\,\beta_n)^2} \\ &\leq 3 \sum_{n\in\mathbb{N}} \frac{(\frac{1}{3}-z)(\mathrm{Re}\beta_n - \frac{1}{3})}{(\frac{1}{3}-z-\mathrm{Re}\beta_n)^2 + (y - \mathrm{Im}\,\beta_n)^2} \\ &\leq 3 \sum_{n\in\mathbb{N}} \mathrm{Re}\gamma_n |g_{z,y}(\gamma_n)|^2, \end{aligned}$$

where $\gamma_n := \beta_n - \frac{1}{3} \in \mathbb{C}_0^+$ and

$$g_{z,y}(s) = \frac{\sqrt{\frac{1}{3}-z}}{s+z-iy} \in \boldsymbol{H}_2(\mathbb{C}_0^+).$$

By Lemma 8.1 $\sum_{n\in\mathbb{N}} \mathrm{Re}\beta_{2n}\delta_{\beta_{2n}}$ is a Carleson measure. Similarly, it can be shown that $\sum_{n\in\mathbb{N}} \mathrm{Re}\beta_{2n+1}\delta_{\beta_{2n+1}}$ is a Carleson measure. From this it follows easily that $\sum_{n\in\mathbb{N}} \mathrm{Re}\gamma_n\delta_{\gamma_n}$ also is a Carleson measure. Hence, Avdonin and Ivanov [2, p. 56] or Garnett [9, Theorem 3.9] implies

$$h(x,y) \leq C\|g_{z,y}\|^2_{\boldsymbol{H}_2(\mathbb{C}_0^+)} = C\left\| e^{-(z-iy)\cdot}\sqrt{\frac{1}{3}-z}\right\|^2_{\boldsymbol{L}_2(0,\infty)} = \frac{C}{2}\frac{x}{\frac{1}{3}-x} \leq \frac{3C}{2}.$$

Thus $|\Theta(s)^{-1}|$ is uniformly bounded on $\mathbb{C}_0^+ \cap \mathbb{C}_{1/4}^-$. □

The following realization result can be found in Salamon [30] and in Ober and Wu [21]. The proof in the notation of this paper is documented in Jacob and Zwart [16].

Lemma 8.3. *Let Θ be given as in Lemma 8.2. Then there exist a Hilbert space V, an exponentially stable C_0-semigroup $T(t)$ on V with generator A, and an admissible control operator B for $T(t)$ such that the system $\Sigma(A,B)$ is exactly controllable in finite time, every $-\overline{\beta}_n$, $n \in \mathbb{N}$, is an eigenvalue of A, and the function $x_n(s) := \frac{1}{s+\overline{\beta}_n}$ is a corresponding eigenvector.*

EXAMPLE 8.4. We now consider the system $\Sigma(A,B)$ given as in Lemma 8.3. Since $\Sigma(A,B)$ is exactly controllable in finite time, by Lemma 3.8 and Proposition 3.9 the system $\Sigma(A+I,B)$ is optimizable. Lemma 8.3 shows that $\{\lambda_n\}_{n\in\mathbb{N}} \subseteq \sigma(A+I)$ holds, where $\lambda_n := -\bar{\beta}_n + 1$, and $x_{\lambda_n}(s) := \frac{1}{s-\lambda_n}$ is an eigenvector corresponding to the eigenvalue λ_n. Thus the spectrum of $A+I$ does not satisfy the SDA at any negative number.

Next, we show that it is not possible to decompose the state space V as $V = V_s \oplus V_u$, where

$$\begin{aligned} V_s &:= \{x_0 \in V \mid e^{-g\cdot}T(\cdot)x_0 \in \boldsymbol{L}_2(0,\infty;V)\},\\ V_u &:= \overline{\operatorname*{span}_{\lambda\in\sigma(A+I) \text{ with } \mathrm{Re}(\lambda)\geq g} P_\lambda V} \end{aligned}$$

for some $g<0$. Let $g<0$ be arbitrary. It is easy to see that $x_{\lambda_n} \in V_u$ if $\mathrm{Re}(\lambda_n) \geq g$. Next, we show that $x_{\lambda_n} \in V_s$ holds if $\mathrm{Re}(\lambda_n) < g$. Let $n \in \mathbb{N}$ with $\mathrm{Re}(\lambda_n) < g$ be arbitrary. Since x_{λ_n} is an eigenvector of A with eigenvalue λ_n, we have that

$$T(t)x_{\lambda_n} = e^{\lambda_n t}x_{\lambda_n}.$$

This shows $e^{-g\cdot}T(\cdot)x_{\lambda_n} \in \boldsymbol{L}_2(0,\infty;V)$.

By the construction of the sequence $\{\beta_n\}_{n\in\mathbb{N}}$ there now exist subsequences $\{\mu_n\}_{n\in\mathbb{N}}$, $\{\nu_n\}_{n\in\mathbb{N}} \subseteq \{\lambda_n\}_{n\in\mathbb{N}}$ such that

(1) $\mathrm{Re}\mu_n < g < \mathrm{Re}\nu_n$, $n \in \mathbb{N}$,

(2) $\nu_n = \mu_n + \alpha_n$, $n \in \mathbb{N}$, with $\alpha_n > 0$ and $\lim_{n\to\infty} \alpha_n = 0$,

(3) $\lim_{n\to\infty} \mathrm{Re}\mu_n = g$ and $\lim_{n\to\infty} \mathrm{Im}\,\mu_n = \infty$.

Thus $x_{\mu_n} \in V_s$ and $x_{\nu_n} \in V_u$ for every $n \in \mathbb{N}$. It is now easy to see that

$$\|x_{\nu_n}\|^2 = \frac{\pi}{|\mathrm{Re}\nu_n|}, \qquad n \in \mathbb{N},$$

and

$$\lim_{n\to\infty} \|x_{\mu_n} - x_{\nu_n}\| = 0.$$

Let us now assume that $V = V_s \oplus V_u$ is satisfied. Then there would exist a projection $P \in \mathcal{L}(V)$ with $\ker P = V_s$ and $\mathrm{Im}\,P = V_u$. Thus

$$\begin{aligned} \frac{\pi}{|g|} = \lim_{n\to\infty} \|x_{\nu_n}\|^2 = \lim_{n\to\infty} \|Px_{\nu_n}\|^2 &= \lim_{n\to\infty} \|P(x_{\nu_n} - x_{\mu_n})\|^2 \\ &\leq \|P\|^2 \lim_{n\to\infty} \|x_{\nu_n} - x_{\mu_n}\|^2 = 0, \end{aligned}$$

which is a contradiction. Hence $V = V_s \oplus V_u$ is not satisfied.

9. Conclusion. For an infinite-dimensional optimizable system with a finite-rank admissible control operator we showed that the system can be decomposed into an exponentially stable subsystem and an exactly controllable subsystem. For the proof we needed the spectrum to be able to be decomposed into a stable and an unstable part. The example in section 8 shows that without this SDA Theorem 1.3 does not hold. Since this example is very technical, we feel that every optimizable system encountered in practice satisfies the SDA. Note that Theorem 1.4 already shows that the unstable part of the spectrum of an optimizable system consists of isolated points.

We gave characterizations of the state spaces of the subsystems. The state space of the unstable part equals the span of all unstable (generalized) eigenvectors and the state space of the exponentially stable part is given by all vectors for which the action of the original semigroup is stable.

From our results we derived easy necessary conditions for a system to be optimizable; see, e.g., Example 5.7, Theorem 6.1, and Proposition 6.3 or Jacob and Zwart [15].

REFERENCES

[1] W. Arendt, A. Grabosch, G. Greiner, U. Groh, H. P. Lotz, U. Moustakas, R. Nagel, F. Neubrander, and U. Schlotterbeck, *One Parameter Semigroups of Positive Operators*, Lecture Notes in Math. 1184, Springer-Verlag, Berlin, 1986.

[2] S. A. Avdonin and S. A. Ivanov, *Families of Exponentials: The Method of Moments in Controllability Problems for Distributed Parameter Systems*, Cambridge University Press, Cambridge, UK, 1995.

[3] R. F. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Texts Appl. Math. 21, Springer-Verlag, New York, 1995.

[4] R. Datko, *A linear control problem in abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–359.

[5] W. Desch and W. Schappacher, *Spectral properties of finite-dimensional perturbed linear semigroups*, J. Differential Equations, 59 (1985), pp. 80–102.

[6] P. L. Duren, *Theory of H^p Spaces*, Pure Appl. Math. 38, Academic Press, San Diego, 1970.

[7] H. O. Fattorini, *Exact controllability of linear systems in infinite dimensional spaces*, in Partial Differential Equations and Related Topics, Tulane Univ., New Orleans, Lecture Notes in Math. 446, Springer-Verlag, New York, 1975, pp. 166–183.

[8] F. Flandoli, I. Lasiecka, and R. Triggiani, *Algebraic Riccati equations with non-smoothing observation arising in hyperbolic and Euler-Bernoulli boundary control problems*, Ann. Mat. Pura Appl., CLIII (1988), pp. 307–382.

[9] J. B. Garnett, *Bounded Analytic Functions*, Academic Press, New York, 1981.

[10] P. Grabowski and F. M. Callier, *Admissible observation operators, duality of observation and control*, Technical report 94-27, Département de Mathématiques, Facultés Universitaires de Namur, Namur, Belgium, 1994.

[11] G. Greiner, J. Voigt, and M. Wolff, *On the spectral bound of the generator of semigroups with positive operators*, J. Operator Theory, 5 (1981), pp. 245–256.

[12] M. L. J. Hautus, *(A, B)-invariant and stabilizability subspace, a frequency domain description*, Automatica J., 16 (1980), pp. 703–707.

[13] E. Hille and R. S. Phillips, *Functional Analysis and Semi–Groups*, Amer. Math. Soc. Colloq. Publ. 31, AMS, Providence, RI, 1957.

[14] F. Huang, *Characteristic conditions for exponential stability of linear dynamical systems in Hilbert spaces*, Ann. Differential Equations, 1 (1985), pp. 43–55.

[15] B. Jacob and H. Zwart, *On lack of optimizability*, in Proceedings of the European Control Conference, Conference ID 976, Timing FR-A-E-3, Brussels, 1997.

[16] B. Jacob and H. Zwart, *Realization of inner functions*, Technical report 1998-17, School of Mathematics, University of Leeds, Leeds, UK, 1998.

[17] C. A. Jacobson and C. N. Nett, *Linear state space systems in infinite-dimensional space: The role and characterization of joint stabilizability/detectability*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 541–550.

[18] T. Kato, *Perturbation Theory for Linear Operators*, Classics Math., Springer-Verlag, Berlin, 1995.

[19] M. G. Kurtz, *A general theorem on the convergence of operator semigroups*, Trans. Amer. Math. Soc., 148 (1980), pp. 23–32.

[20] S. A. Nefedov and F. A. Sholokhovich, *A criterion for the stability of dynamical systems with finite-dimensional input*, Differentsial'nye Uravneniya, 22 (1986), pp. 163–166.

[21] R. J. Ober and Y. Wu, *Infinite-dimensional continuous-time linear systems: Stability and structure analysis*, SIAM J. Control Optim., 34 (1996), pp. 757–812.

[22] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, 1983.

[23] A. J. Pritchard and J. Zabczyk, *Stability and stabilizability of infinite dimensional systems*, SIAM Rev., 23 (1981), pp. 25–52.

[24] J. Prüss, *On the spectrum of C_0-semigroups*, Trans. Amer. Math. Soc., 284 (1984), pp. 847–857.

[25] R. Rebarber, *Necessary conditions for exponential stability of distributed parameter systems with infinite dimensional unbounded feedback*, Systems Control Lett., 14 (1990), pp. 241–248.

[26] R. Rebarber, *Conditions for the equivalence of internal and external stability for distributed parameter systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 994–998.

[27] R. Rebarber and G. Weiss, *An extension of Russell's principle on exact controllability*, in European Control Conference, Conference ID 736, Timing TH-E-E-2, Brussels, 1997.

[28] R. Rebarber and H. J. Zwart, *Open loop stabilizability of infinite-dimensional systems*, Math. Control Signals Systems, 11 (1998), pp. 129–160.

[29] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open problems*, SIAM Rev., 20 (1978), pp. 639–739.

[30] D. Salamon, *Realization theory in Hilbert space*, Math. Systems Theory, 21 (1989), pp. 147–164.

[31] R. Triggiani, *On the stabilizability problem in Banach spaces*, J. Math. Anal. Appl., 52 (1975), pp. 383–403.

[32] G. Weiss, *Admissibility of input elements for diagonal semigroups on l^2*, Systems Control Lett., 10 (1988), pp. 79–82.

[33] G. Weiss, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.

[34] G. Weiss, *Admissible observation operators for linear semigroups*, Israel J. Math., 65 (1989), pp. 17–43.

[35] G. Weiss, *Two conjectures on the admissibility of control operators*, in Estimation and Control of Distributed Parameter Systems, W. Desch, F. Kappel, and K. Kunisch, eds., Birkhäuser Verlag, Basel, 1991, pp. 367–378.

[36] G. Weiss, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.

[37] G. Weiss and R. F. Curtain, *Dynamical stabilization of regular linear systems*, IEEE Trans. Automat. Control, 42 (1997), pp. 4–21.

[38] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

[39] J. Zabczyk, *A note on C_0-semigroups*, Bull. Polish Acad. Sci. Math., 23 (1975), pp. 895–898.

[40] H. J. Zwart, *Equivalence between open-loop and closed-loop invariance for infinite-dimensional systems: A frequency domain approach*, SIAM J. Control Optim., 26 (1988), pp. 1175–1199.

[41] H. J. Zwart, *Linear quadratic optimal control for abstract linear systems*, in Conference on Modelling and Optimization of Distributed Parameter Systems Applications to Engineering, Warsaw, Poland, Chapman & Hall, New York, 1996, pp. 175–182

SIAM J. CONTROL OPTIM.
Vol. 37, No. 5, pp. 1456–1482

ASYMPTOTIC BEHAVIOR OF A MARKOVIAN STOCHASTIC ALGORITHM WITH CONSTANT STEP*

JEAN-CLAUDE FORT† AND GILLES PAGÈS‡

Abstract. We first derive from abstract results on Feller transition kernels that, under some mild assumptions, a Markov stochastic algorithm with constant step size ε usually has a tight family of invariant distributions ν^ε, $\varepsilon \in (0, \varepsilon_0]$, whose weak limiting distributions as $\varepsilon \downarrow 0$ are all flow-invariant for its ODE. Then the main part of the paper deals with a kind of converse: what are the possible limiting distributions among all flow-invariant distributions of the ODE? We first show that no repulsive invariant (thin) set can belong to their supports. When the ODE is a stochastic pseudogradient descent, these supports cannot contain saddle or spurious equilibrium points either, so that they are eventually supported by the set of local minima of their potential. Such results require only the random perturbation to lie in L^2. Various examples are treated, showing that these results yield some weak convergence results for the ν^ε's, sometimes toward a saddle point when the algorithm is not a pseudogradient.

Key words. stochastic algorithm, stationary distribution, Markov chain, gradient descent, saddle point

AMS subject classifications. 62L20, 60J27

PII. S0363012997328610

Introduction. The use of recursive stochastic algorithms is widespread for solving optimization problems. This is due to the new simulation facilities brought by modern scientific computation. Stochastic algorithms are especially encountered in situations where some on-line parameter estimation has to be performed. Such algorithms are recursive, the current estimate being updated at every new observation of a process according to some time-varying deterministic (or predictable) parameter ε_t called *step* or *gain* parameter. In most practical situations, this parameter is gradually decreased so as to finally reach some lower bound ε_∞. Generally ε_∞ is nonzero to prevent any "freezing" of the algorithm at some metastable state or to track some possible slow change of the target.

Hence, this work addresses constant step algorithms, i.e., $\varepsilon_t = \varepsilon_\infty > 0$.

On the other hand, partially because of the recent developments of artificial neural network procedures, it turns out that many learning algorithms share the same features: a "gentle" white noise perturbing the "sophisticated" dynamics of a deterministic "average" differential system $ODE_h \equiv \dot{x} = -h(x)$ (known as the *ordinary differential equation* of the system) so that it reads

$$X^{t+1} = X^t - \varepsilon H(X^t, \omega^{t+1}),\ X^t \in \mathbb{R}^d,$$
$$\omega^t \text{ independently and identically distributed (i.i.d.)},\quad h(x) := \mathbb{E}(H(x, \omega^1)).$$

In such a framework, an algorithm proves to be a homogeneous Markov chain parametrized by its step ε. Under some reasonable conditions, for every small enough

*Received by the editors October 10, 1997; accepted for publication (in revised form) October 15, 1998; published electronically July 27, 1999.
http://www.siam.org/journals/sicon/37-5/32861.html

†Institut E. Cartan, Université Nancy I, B.P. 239, F-54506 Vandœuvre-Lès-Nancy Cedex, France, and SAMOS, Université Paris I, UFR 27, 90 rue de Tolbiac, F-75634 Paris Cedex 13, France (fortjc@iecn.u-nancy.fr).

‡Laboratoire de Probabilité, URA 224, Université P. and M. Curie, 4, Pl. Jussieu, F-75252 Paris Cedex 05, France, and Université Paris XII, UFR Sciences, 61 av. du Gal de Gaulle, 94010 Créteil Cedex, France (gpa@ccr.jussieu.fr).

ε the chain owns (at least) one invariant distribution ν^ε. Let $\mathcal{I}_\varepsilon$ be the set of invariant distributions of the algorithm with step ε. The set $\cup_{\varepsilon \in (0,\varepsilon_0]} \mathcal{I}_\varepsilon$ is tight for small enough $\varepsilon_0 > 0$, i.e., weakly relatively compact. On the other hand, the differential system ODE_h being a (deterministic) homogeneous Markov chain, usually owns some invariant distribution(s) as well, characterized by a flow-invariance property. Let $\mathcal{I}_h$ be the set of such flow-invariant distributions. Actually, in most practical situations, $\mathcal{I}_h$ is infinite: whenever ODE_h has two equilibrium points x_1^* and x_2^*, any combination $\alpha\delta_{x_1^*} + (1-\alpha)\delta_{x_2^*}$ is an invariant distribution. The set $\mathcal{I}_h$ can be much more sophisticated: for instance, an attracting cycle surrounding a repulsive equilibrium, etc.

It is a rather classical result that the set $\mathcal{I}_{0^+} := \cap_{\varepsilon\to 0}\overline{\cup_{0<\eta\le\varepsilon}\mathcal{I}_\eta}$ of the *possible weak limiting distributions* of $\nu^\varepsilon \in \mathcal{I}_\varepsilon$ as $\varepsilon \downarrow 0$ is contained in $\mathcal{I}_h$ (see [15] in one dimension, [20], [29] in the case of a unique attracting equilibrium x^* or, more recently, [18] in the slightly different framework of perturbed dynamical systems ; see also [9] for similar results for abstract continuous-time Markov processes).

However, whenever $\mathcal{I}_h \neq \{\delta_{x^*}\}$, the inclusion generally does not hold as an equality: not any flow-invariant distribution $\nu \in \mathcal{I}_h$ is a limiting value of $\nu^{\varepsilon_p} \in \mathcal{I}_{\varepsilon_p}$ as $\varepsilon_p \downarrow 0$. Thus, it seems quite natural that a repulsive equilibrium of ODE_h cannot be asymptotically weighted by any invariant distribution ν^ε of the algorithm when $\varepsilon \downarrow 0$, provided that there is noise enough at this point to push it away. Indeed, similar problems have been investigated in the decreasing step setting, see, e.g., [7], [21], or [28], for repulsive or saddle points leading to the conclusion that such points cannot be limit points for the algorithm. The question is to state whether or not such results still hold in the constant step setting.

These problems are deeply connected to the field of randomly perturbed systems which has been extensively investigated by several authors. Freidlin and Wentzell [13] deal with an ordinary differential equation perturbed by a vanishing standard Brownian motion. They show, using large deviations techniques, that such a diffusion converges in probability to some absolute minima of a suitably defined potential function. Y. Kifer, in several papers and in his book (see, e.g., [18, Chap. 2]), treats the case of randomly perturbed discrete dynamical systems when the perturbation fades. He obtains some general results, still based on a large deviation approach, that prove that the only possible limiting distributions for the invariant distributions ν^ε as $\varepsilon \downarrow 0$ are supported by "quasi-attractors" which turn out to be classical attractors under some natural assumptions. In fact, these works are more in connection with algorithms with decreasing step, since in all cases the perturbation fades as time goes by. In the field of Markov stochastic algorithms with constant step, new results by large deviation techniques have been obtained by M. Benaïm in [3]. This work, carried out independently from ours, was originally motivated by urn processes where the random perturbation term is naturally bounded.

Our aim in this paper is of the same sort but with different techniques. It consists, still within this Markov framework, of investigating the set $\mathcal{I}_{0^+}$ but under some low moment assumption on the random perturbation term (essentially L^2). Thus, a comparison shows what results are or are not moment dependent. Our methods rely on the local or global existence of a smooth Lyapunov function for the algorithm. We show, by twisting this function somewhat, that the distributions in $\mathcal{I}_{0^+}$ never weight the repulsive equilibrium points or the *thin* invariant repelling sets provided that they are excited enough by the noise. Furthermore, in the *pseudogradient* setting, i.e., when there exists a *global* Lyapunov function V such that $(\nabla V|h) \ge 0$ and

$\{(\nabla V|h) = 0\} = \{\nabla V = 0\}$, the whole set $\mathcal{I}_{0^+}$ is supported by the local minima of V if other critical points are excited enough. The case of *spurious* points is also investigated (x^* is spurious if $h(x^*) = 0$, whereas $\nabla V(x^*) \neq 0$).

The paper is organized as follows. In section 1 some existence and tightness results for invariant distributions of Feller homogeneous Markov chains are recalled; these yield some tractable criteria for constant gain stochastic algorithms. The aim is to show that our framework is quite realistic from a practical point of view. We do not care about the possible uniqueness of the invariant distribution for a given positive step ε as this question seems not to interfere with the asymptotic behavior of these distributions.

Section 2 is essentially devoted to the flow-invariance theorem and its first applications. This theorem proves that, under a mild uniform integrability assumption, $\mathcal{I}_{0^+} \subset \mathcal{I}_h$. An additional result is provided when the existence of the flow fails, provided there is some global Lyapunov function. In these two sections, we give only some very short proofs, along with references for the classical technical points.

Section 3 is the main part of the work. It relies on a second-order Taylor expansion in ε of the homogeneous Markov transitions $P^\varepsilon(x, dy)$. First, this yields that no excited enough repulsive critical point of the average function h can be asymptotically weighted by the invariant distributions of the algorithm as $\varepsilon \downarrow 0$. This result extends to repulsive *thin invariant sets.* Two subsections deal with important examples: the Lemniscate example and the periodic/quasi-cycle example, which illustrate both the apparent dissimilarity between the constant and the decreasing step versions of the same stochastic algorithm.

Second, the case of excited enough saddle points of ODE_h is investigated. It is first pointed out that such critical points can be asymptotically weighted when the algorithm is not a stochastic pseudogradient. The Lemniscate example stresses this situation: all the mass of the ν^ε concentrates at some saddle point of h. At this stage, a whole subsection is devoted to stochastic pseudogradient algorithms under the classical assumption in differential geometry that all saddle points are of Morse type. Finally, some further types of critical points are eliminated: the spurious equilibrium points. This latter result solves the case of inflection points in one dimension. Several examples illustrate these results.

Notation.

- $|\,.\,|$ denotes the canonical Euclidean norm on any $\mathbb{R}^d$ space unless noted otherwise;
- $(.|.)$ denotes the canonical inner product;
- $B(x, r)$ denotes the Euclidean open ball with center x and radius $r > 0$;
- $\mathcal{C}(0; 1)$ denotes the unit circle;
- $u \in \mathbb{R}^d$ denotes a row vector and ${}^t u$ denotes its (horizontal) transpose;
- $C_b(\mathbb{R}^d, \mathbb{R}) := \{f : \mathbb{R}^d \longrightarrow \mathbb{R},$ continuous and bounded$\}$;
- $C_0(\mathbb{R}^d, \mathbb{R}) := \{f : \mathbb{R}^d \longrightarrow \mathbb{R},$ continuous subject to (s.t.) $\lim_{|x|\to+\infty} f(x) = 0\}$;
- $\overset{(\mathcal{T})}{\Longrightarrow}$ denotes the weak convergence of probability measures on the topological space X endowed with the topology $\mathcal{T}$; when this topology is obvious (e.g., on $X = \mathbb{R}^d$) the notation $\overset{(X)}{\Longrightarrow}$ will be preferred;
- The letter ν—with or without superscript—always denotes a *probability* measure.

1. Background: Existence and tightness of the invariant distributions of a family of Markov chains.

1.1. General results. Let $(P^\varepsilon(x, dy))_{x\in\mathbb{R}^d}$, $\varepsilon \in (0, \varepsilon_0]$ be a family of Markov transitions on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. We propose in Theorem 1.1 below several criteria for the existence and tightness of a family of stationary distributions in a Feller setting.

A distribution ν^ε is P^ε stationary (or P^ε invariant) if it satisfies $\nu^\varepsilon P^\varepsilon = \nu^\varepsilon$, i.e., $\int \nu^\varepsilon(dx)P^\varepsilon(x,A) = \nu^\varepsilon(A)$ for every Borel set $A \in \mathcal{B}(\mathbb{R}^d)$.

A probability transition $P(x,dy)$ is Feller if $P(f)(x) := \int f(y)P(x,dy)$ maps $C_b(\mathbb{R}^d,\mathbb{R})$ into itself.

These criteria are based on the existence of a Lyapunov function. They are adapted from classical criteria mentioned, e.g., in [9] or [25]. They turn out to be especially well fitted for stochastic approximation with constant step.

THEOREM 1.1. *Let $(P^\varepsilon)_{\varepsilon\in(0,\varepsilon_0]}$ be a family of transitions.*

A. PAKES–HAS'MINSKII SETTING. *Assume that $(P^\varepsilon)_{\varepsilon\in(0,\varepsilon_0]}$ satisfies the Pakes–Has'minskii assumption*

$$(H_b) \equiv \exists\, V : \mathbb{R}^d \to \mathbb{R}_+,\ \psi : \mathbb{R}^d \to \mathbb{R}\ s.t. \begin{cases} \text{(a)}\ \ \psi \le C\ \ and\ \lim\limits_{|x|\to+\infty} \psi(x) = -\infty, \\ \text{(b)}\ \ \begin{cases} \forall\,\varepsilon \in (0,\varepsilon_0],\ \exists\,\lambda(\varepsilon) > 0,\ \mu(\varepsilon) \in \mathbb{R} \\ s.t.\ P^\varepsilon V - V \le \lambda(\varepsilon)\psi + \mu(\varepsilon). \end{cases} \end{cases}$$

(a) *If the transitions $P^\varepsilon(x,dy)$ are Feller for every $\varepsilon \in (0,\varepsilon_0]$, there exists a family $(\nu^\varepsilon)_{\varepsilon\in(0,\varepsilon_0]}$ of P^ε-invariant distributions.*

(b) *If $\sup_{\varepsilon\in(0,\varepsilon_0]} \frac{\mu(\varepsilon)}{\lambda(\varepsilon)} < +\infty$, then the set of invariant distributions $\{\nu^\varepsilon \,/\, \nu^\varepsilon P^\varepsilon = \nu^\varepsilon,\ \varepsilon \in (0,\varepsilon_0]\}$ is tight.*

The function V is called a Lyapunov function for the family.

B. HAJEK SETTING. *Assume that $(P^\varepsilon)_{\varepsilon\in(0,\varepsilon_0]}$ satisfies the Hajek assumption*

$$(H_c) \equiv \exists\, V : \mathbb{R}^d \to \mathbb{R}_+,\ \ s.t. \begin{cases} \text{(a)}\quad \lim\limits_{|x|\to+\infty} V(x) = +\infty, \\ \text{(b)}\quad \begin{cases} \forall\,\varepsilon \in (0,\varepsilon_0],\ \exists\,\alpha(\varepsilon) \in (0,1),\ \exists\,\beta(\varepsilon) \in \mathbb{R}_+ \\ s.t. \quad P^\varepsilon V \le \alpha(\varepsilon)V + \beta(\varepsilon). \end{cases} \end{cases}$$

Then, (H_b) is fulfilled with $\lambda(\varepsilon) := 1-\alpha(\varepsilon)$, $\mu(\varepsilon) := \beta(\varepsilon)$, and $\psi := -V$. If, moreover,

$$L := \sup_{\varepsilon\in(0,\varepsilon_0]} \frac{\beta(\varepsilon)}{1-\alpha(\varepsilon)} < +\infty,\ \textit{then}\ \sup\{\nu^\varepsilon(V),\ \nu^\varepsilon P^\varepsilon = \nu^\varepsilon,\ \varepsilon \in (0,\varepsilon_0]\} \le L.$$

Proof (sketch). Setting A is a parametrized version of a classical criterion of existence of a stationary distribution for Feller–Markov chains (see, e.g., [9, p. 272]). The boundedness of the $\nu^\varepsilon(V)$'s in setting B derives from the following inequalities:

$$\begin{aligned} \frac{1}{n}\sum_{k=0}^{n-1} P^{\varepsilon,k}(V \wedge K)(x) &\le \left(\frac{1}{n}\sum_{k=0}^{n-1} P^{\varepsilon,k}(V)(x)\right) \wedge K \\ &\le \left(\frac{\beta(\varepsilon)}{1-\alpha(\varepsilon)} + \alpha(\varepsilon)^n V(x)\right) \wedge K. \end{aligned}$$

Then, integrating with respect to ν^ε, letting first n go to $+\infty$ and then K, completes the proof. □

This theorem is well suited for stochastic algorithms with constant step (see subsection 1.2 below). It is important for applications to emphasize two features of these criteria: the tightness conclusion holds for the *whole set* $\{\nu^\varepsilon \,/\, \nu^\varepsilon P^\varepsilon = \nu^\varepsilon,\ \varepsilon \in (0,\varepsilon_0]\}$

- without a uniqueness assumption of the invariant distribution at a given ε (such a property is not always true and is usually difficult to fulfill even when it holds);
- without a Feller assumption, once the existence of a family $(\nu^\varepsilon)_{\varepsilon\in(0,\varepsilon_0]}$ is granted.

Remarks. Note that (H_b) implies that the Lyapunov function V satisfies

$$\lim_{|x|\to+\infty} V(x) = +\infty.$$

Following, e.g., [8], the Hajek setting implies, for every $\varepsilon \in (0, \varepsilon_0]$, the stability of the chain $(X^{\varepsilon,t})_{t\in\mathbb{N}}$ in the following sense: for every starting value $x \in \mathbb{R}^d$, $\mathbb{P}_x$-almost surely (a.s.), the sequence of empirical distributions $(\frac{1}{t}\sum_{0\le s\le t-1}\delta_{X^{\varepsilon,s}})_{t\ge1}$ is tight and all its weak limiting distributions are P^ε-invariant.

If $P^\varepsilon(x,dy) = p_\varepsilon(x,y)\theta(dy)$ where $\theta(dy)$ denotes a σ-finite nonnegative measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $p^\varepsilon(x,y) > 0$, $\theta(dy)$-almost everywhere (a.e.) for every x, then uniqueness of the invariant distributions ν^ε holds (see, e.g., [8]). So, in the Hajek setting, $\mathbb{P}_x$-a.s., the empirical distribution weakly converges to this invariant distribution ν^ε.

1.2. The case of stochastic algorithms with constant gain. In this paper, a stochastic algorithm with constant gain parameter is a family of homogeneous Markov chains, depending on a parameter $\varepsilon > 0$, satisfying the general recursive equation

$$X^{\varepsilon,t+1} = X^{\varepsilon,t} - \varepsilon H(X^{\varepsilon,t}, \omega^{t+1}), \tag{1.1}$$

where $H : \mathbb{R}^d \times E \longrightarrow \mathbb{R}^d$ is a Borel function, $(\omega^t)_{t\ge1}$ is an i.i.d. sequence with common distribution μ on a measurable space $(E, \mathcal{B}(E))$ and ε is a positive real number. The transition $(P^\varepsilon(x,dy))_{x\in\mathbb{R}^d}$ is obviously defined on nonnegative or bounded Borel functions

$$P^\varepsilon(f)(x) := \int f\left(x - \varepsilon H(x,\omega)\right)\mu(d\omega).$$

The first assumption to be made ensures that the above transitions are Feller: There exists some $\rho \in (0,1]$ s.t.

$$(C^\mu_\rho) \equiv \left\{\begin{array}{ll} \text{(i)} & \forall x \in \mathbb{R}^d,\ y \mapsto H(y,\omega) \text{ is } \mu(d\omega)\text{-a.s. continuous at } x, \\ \text{(ii)} & \left\{\begin{array}{l} \forall x \in \mathbb{R}^d,\ \omega \mapsto H(x,\omega) \text{ belongs to } L^{1+\rho}(\mu) \text{ and} \\ x \mapsto \displaystyle\int |H(x,\omega)|^{1+\rho}\mu(d\omega) \text{ is bounded on compact sets.} \end{array}\right. \end{array}\right.$$

Note that if (C^μ_ρ) holds for some $\rho \in (0,1]$, the so-called mean function $h(x) := \int H(x,\omega)\mu(d\omega)$ of the algorithm is continuous (due to uniform integrability).

Most Lyapunov functions V that will be used further on fulfill the following smoothness assumption:

$$(L_\rho) \equiv\ V \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}_+) \text{ and } \nabla V \text{ is } \rho\text{-Hölder with coefficient } [\nabla V]_\rho.$$

Of course, some nonnegativity assumption on $(\nabla V|h)$ will be added when necessary throughout the text. From a practical point of view, finding a function V that satisfies assumption (L_ρ) for some $\rho \in (0,1)$, instead of (L_1), allows a decrease in the integrability requirement on the noise.

Theorem 1.1 applied to Markov stochastic algorithms with constant step straightforwardly yields the following global existence and tightness result for the P^ε-invariant distributions ν^ε.

Proposition 1.2. A. Pakes–Has'minskii setting. *Let $\rho \in (0,1]$ and $V : \mathbb{R}^d \to \mathbb{R}_+$. Assume that*

$$(1.2) \quad (Has)_\rho \equiv \begin{cases} \text{(a)} & \textit{the function } H \textit{ satisfies} \\ & (C^\mu_\rho) \textit{ and } \textit{ the function } V \textit{ satisfies } (L_\rho), \\ \text{(b)} & \exists\, A > 0 \textit{ s.t.} \\ & \displaystyle\lim_{|x|\to+\infty}\left((\nabla V|h)(x) - \frac{1}{A}\int |H(x,\omega)|^{1+\rho}\mu(d\omega)\right) = +\infty; \end{cases}$$

then assumption (H_b) holds.

B. Hajek setting. *Assume that there is some $\rho \in (0,1]$ such that*

$$(1.3) \quad (Haj)_\rho \equiv \begin{cases} \text{(a)} & \textit{the function } H \textit{ satisfies } C^\mu_\rho, \\ \text{(b)} & \textit{the function } V \textit{ satisfies } (L_\rho) \textit{ and } \displaystyle\lim_{|x|\to+\infty} V(x) = +\infty, \\ \text{(c)} & \exists\, A > 0 \textit{ s.t.} \\ & \displaystyle V(x) + \int |H(x,\omega)|^{1+\rho}\mu(d\omega) \le A((\nabla V|h)(x) + 1). \end{cases}$$

Then assumption H_c is fulfilled.

B$'$. *Sometimes, rather than the above assumption $(Haj)_\rho$(c), it may be more convenient to check the slightly more stringent*

$$(1.4) \quad (\widetilde{Haj})_\rho \equiv \begin{cases} \text{(a), (b),} \\ \text{(c)} \;\; \exists\, \tilde{A} > 0, \\ \qquad \displaystyle\frac{1}{\tilde{A}}\int |H(x,\omega)|^{1+\rho}\mu(d\omega) \le V(x) + 1 \le \tilde{A}((\nabla V|h)(x) + 1). \end{cases}$$

Proof (sketch). Both settings rely on the following inequality, derived from the bounded increment formula

$$\forall\, x \in \mathbb{R}^d, \quad V\big(x - \varepsilon H(x,\omega)\big) \le V(x) - \varepsilon\big(\nabla V(x)|H(x,\omega)\big) + \varepsilon^{1+\rho}[\nabla V]_\rho |H(x,\omega)|^{1+\rho}.$$

Hence, $V(x - \varepsilon H(x,\omega)) \in L^1(\mu(d\omega))$, and integrating with respect to $\mu(d\omega)$ yields

$$\forall x \in \mathbb{R}^d, \; (P^\varepsilon V - V)(x) \le -\varepsilon\,(\nabla V|h)\,(x) + \varepsilon^{1+\rho}[\nabla V]_\rho \int |H(x,\omega)|^{1+\rho}\mu(d\omega).$$

A. Pakes–Has'minskii setting. One concludes by setting

$$\psi(x) := \frac{1}{A}\int |H(x,\omega)|^{1+\rho}\mu(d\omega) - (\nabla V|h)(x), \; \lambda(\varepsilon) = \varepsilon, \; \mu(\varepsilon) := 0,$$

$\varepsilon_0 < \frac{1}{(A[\nabla V]_\rho)^{\frac{1}{\rho}}}$, and $C := \sup_{x\in\mathbb{R}^d} \psi(x)$. ($\psi$ is bounded by (C^μ_ρ) (ii) and 1.2(b).)

B. Hajek setting. One concludes by setting

$$\alpha(\varepsilon) := \left(1 - \frac{1}{A}\varepsilon(1 - \varepsilon^\rho A[\nabla V]_\rho)\right), \; \beta(\varepsilon) = \varepsilon^{1+\rho}[\nabla V]_\rho A + (1 - \alpha(\varepsilon)) \sup_{(\nabla V|h)\ge\frac{1}{2A}} V(x),$$

and

$$\varepsilon_0 < \min\left(A, \frac{1}{(A[\nabla V]_\rho)^{\frac{1}{\rho}}}\right). \qquad \square$$

About other natural settings. A simpler setting that is commonly encountered in applications is the case of compact valued stochastic algorithms like the Kohonen algorithm (see, e.g., [5] or [6]) and the normalized Benzecri–Oja algorithm (see [26]). Thus, assume that for $\varepsilon \in (0, \varepsilon_0]$ and every x^0 lying in a compact set K, the algorithm defined by (1.1) a.s. lies in K. It is straightforward that, whenever the algorithm is Feller on K (with an obvious definition), every transition P^ε has at least one invariant distribution and $\{\nu^\varepsilon \,/\, \nu^\varepsilon P^\varepsilon = \nu^\varepsilon\}$ is tight. Furthermore, all the results below remain valid for such K-valued Feller algorithms with the straightforward adaptations: a function V is Lyapunov on K if $V \in \mathcal{C}^1(\overset{\circ}{K})$, ∇V continuously extends to K, $(\nabla V | h) \geq 0$, etc. Such an extension will work out for the Benzecri–Oja algorithm that lives in the unit d-dimensional sphere, or some special cases of the Kohonen algorithm (uniform stimuli and 0 neighbor setting).

When the algorithm is not Feller, another family of methods based on ergodicity and recurrence is available (see [22] or [8]) for the existence and tightness of the invariant distributions. Once this point is solved, similar studies to those carried out below can be done; thus, in [6], after proving the existence of the invariant distributions ν^ε by a Doeblin recurrence approach for the Kohonen algorithm with 0 neighbor, a localization of the support of the limiting distribution is carried out. This is done by specific methods since this algorithm is a stochastic gradient descent whose gradient function is not even continuous (see also [27]).

2. Asymptotics of ν^ε as $\varepsilon \downarrow 0$. This section is devoted to the location of the support of any weak limiting value of a tight family $(\nu^\varepsilon)_{\varepsilon \in (0, \varepsilon_0]}$ of P^ε-stationary probability measures. The results mentioned in the first subsection call upon some functional results on stochastic processes and need some more stringent hypothesis on the average function h through ODE_h $\dot{x} = -h(x)$. In the second subsection, we give some applications of the flow-invariance theorem. In the third subsection, we provide some results and examples requiring less regularity on h in the spirit of section 1.

2.1. A flow-invariance theorem for the ν^ε's. In this section we deal with a flow-invariance result for any tight family $(\nu^\varepsilon)_{\varepsilon \in (0, \varepsilon_0]}$ of P^ε-invariant distributions. This version for Markov stochastic algorithms can appear as the constant step version of the ODE method developed by Ljung, Kushner, and Clark. It looks like a variant of the old seminal averaging principle by Has'minskii (see [15] or [18]) for perturbed deterministic dynamical systems. Let us mention, too, for diffusion approximation, the work by Norman on limit theorems for stationary sequences (see [24]) and the synthetic result that can be found in ([9, p. 244]).

The following assumption will hold throughout this section:

$$H_{\text{ODE}} \equiv \begin{cases} \text{(a)} \quad h \text{ is continuous,} \\ \text{(b)} \quad \begin{cases} \text{for every } x_0 \in \mathbb{R}^d \text{ there exists a } \textit{unique } \Phi(x_0, .) \in \mathcal{C}^1(\mathbb{R}_+, \mathbb{R}^d) \\ \text{s.t.} \Phi(x_0, u) = x_0 - \displaystyle\int_0^u h(\Phi(x_0, v)) dv. \end{cases} \end{cases}$$

Uniqueness is generally provided by a local Lipschitz assumption on h.

Let $\varepsilon > 0$, $x_0 \in \mathbb{R}^d$. Set

$$X_0^{(\varepsilon, x_0)} := x_0 \quad \text{and} \quad X_u^{(\varepsilon, x_0)} := X^{(\varepsilon, x_0), t} \text{ if } u \in [t\varepsilon, (t+1)\varepsilon),$$

where $(X^{(\varepsilon, x_0), t})_{t \in \mathbb{N}}$ denotes the algorithm with constant step ε starting at x_0 as defined by (1.1). A straightforward computation shows that $X_u^{(\varepsilon,x_0)}$ satisfies

$$\forall\, u \in \mathbb{R}_+,\ X_u^{(\varepsilon,x_0)} = x_0 - \int_0^u h(X_v^{(\varepsilon,x_0)})dv + \varepsilon M^{\varepsilon,x_0,[\frac{u}{\varepsilon}]} + \int_{\varepsilon[\frac{u}{\varepsilon}]}^u h(X_v^{(\varepsilon,x_0)})dv, \tag{2.1}$$

$$\text{where } M^{\varepsilon,x_0,t} = \sum_{s=0}^{t-1} h(X^{(\varepsilon,x_0),s}) - H(X^{(\varepsilon,x_0),s},\omega^{s+1}), \quad M^{\varepsilon,x_0,0} = 0, \tag{2.2}$$

and u denotes the integral part of $u \in \mathbb{R}_+$. The sequence $(M^{\varepsilon,x_0,t})_{t\in\mathbb{N}}$ is a martingale with respect to its natural filtration under mild integrability assumptions.

Proposition 2.1. (a) *For every sequence $\varepsilon_p \downarrow 0$,*

$$\forall\, K > 0,\ \forall\, T \in \mathbb{R}_+,\ \sup_{u\in[0,T]} \sup_{|x_0|\le K} |X_u^{(\varepsilon_p,x_0)} - \Phi(x_0,u)| \xrightarrow[p\to\infty]{\mathbb{P}} 0.$$

(b) *In particular, for any function $f \in \mathcal{C}_b(\mathbb{R}^d,\mathbb{R}_+)$ s.t. $f(0) = 0$,*

$$\forall\, K > 0,\ \forall u \in \mathbb{R}_+, \qquad \sup_{|x_0|\le K} \mathbb{E}\left(f\left(X_u^{(\varepsilon,x_0)} - \Phi(x_0,u)\right)\right) \longrightarrow 0 \qquad \textit{as } \varepsilon \to 0.$$

Item (b) is an obvious corollary of item (a). Item (a) reads as follows in a sequential form: for any sequences $\varepsilon_p \downarrow 0$, $x_0^p \to x_0^\infty$,

$$\forall\, K > 0,\ \forall\, T \in \mathbb{R}_+,\ \sup_{u\in[0,T]} |X_u^{(\varepsilon_p,x_0^p)} - \Phi(x_0^\infty,u)| \xrightarrow[p\to\infty]{\mathbb{P}} 0.$$

One easily derives this convergence (in distribution) from some abstract weak functional convergence theorems for semimartingales (see, e.g., [17, Thm. 3.39, p. 510] for a comprehensive approach). A direct and self-contained proof is available in [10] (when the function h is not bounded, the proof involves some localization techniques relying on the Sk-regularity for stopped processes).

Theorem 2.2. *Assume that the family $(\nu^\varepsilon)_{\varepsilon\in(0,\varepsilon_0]}$ of P^ε-invariant distributions is tight and that both H_{ODE} and*

$$(UI)_{loc} \equiv \forall\, K \subset \mathbb{R}^d,\ \textit{compact set},\ (|H(x,\omega)|)_{x\in K} \textit{ is } \mu(d\omega)\textit{-uniformly integrable} \tag{2.3}$$

hold. Then, every limiting distribution ν^0 of ν^ε as $\varepsilon \downarrow 0$ satisfies

$$\forall u \in \mathbb{R}_+, \quad \Phi(\nu^0,u) = \nu^0. \tag{2.4}$$

Consequently, if the flow is uniquely ergodic, i.e., exactly one flow-invariant distribution ν^0 exists, then $\nu^\varepsilon \Rightarrow \nu^0$ as $\varepsilon \downarrow 0$.

Note that as a practical matter, $(UI)_{loc}$ always holds when $(\widetilde{Haj})_1$ does.

Proof of Theorem (2.2). Let $u \in \mathbb{R}_+$ and let f be a bounded Lipschitz function, with Lipschitz coefficient $[f]$. $\mathbb{E}(\,.\,)$ will denote the $\mathbb{P}$-expectation on the probability space $(\Omega,\mathcal{A},\mathbb{P})$ on which the innovations ω^t are defined. We have

$$\begin{aligned}
&\left|\int f(\Phi(x_0,u))\nu^\varepsilon(dx_0) - \int f(x_0)\nu^\varepsilon(dx_0)\right| \\
&\qquad = \left|\int \left(f(\Phi(x_0,u)) - P^{[\frac{u}{\varepsilon}]}f(x_0)\right)\nu^\varepsilon(dx_0)\right| \\
&\qquad = \left|\int \left(f(\Phi(x_0,u)) - \mathbb{E}\left(f(X_u^{(\varepsilon,x_0)})\right)\right)\nu^\varepsilon(dx_0)\right| \\
&\qquad \le \int \nu^\varepsilon(dx_0)\mathbb{E}\left(\left|f(X_u^{(\varepsilon,x_0)}) - f(\Phi(x_0,u))\right|\right).
\end{aligned}$$

Let $\varepsilon_p \to 0$ and $\eta > 0$. Since the sequence $(\nu^{\varepsilon_p})_{p\geq 0}$ is tight, there exists a compact set K_η such that

$$\nu^{\varepsilon_p}(K_\eta) \geq 1 - \frac{\eta}{2\|f\|_\infty}$$

for every $p \geq 1$. Hence, integrating separately on K_η and ${}^cK_\eta$ yields

$$\begin{aligned}&\left|\int f(\Phi(x_0,u))\nu^{\varepsilon_p}(dx_0) - \int f(x_0)\nu^{\varepsilon_p}(dx_0)\right| \\ &\qquad \leq \frac{2\|f\|_\infty}{2\|f\|_\infty}\eta + \sup_{x_0\in K_\eta} \mathbb{E}\left(2\|f\|_\infty \wedge ([f]|X_u^{(\varepsilon_p,x_0)} - \Phi(x_0,u)|)\right).\end{aligned}$$

Let $\varphi(x) := 2\|f\|_\infty \wedge ([f]|x|)$. $\varphi(0) = 0$ and $\varphi \in \mathcal{C}_b(\mathbb{R}^d, \mathbb{R}_+)$. Since the flow $\Phi(x_0,u)$ is continuous in x_0, Proposition 2.1(b) yields

$$\forall\, u \in \mathbb{R}_+, \quad \forall \eta > 0, \quad \overline{\lim_p}\left|\int f(x(x_0,u))\nu^{\varepsilon_p}(dx_0) - \int f(x_0)\nu^{\varepsilon_p}(dx_0)\right| \leq \eta,$$

hence

$$\forall\, u \in \mathbb{R}_+, \quad \int f(\Phi(x_0,u))\nu^0(dx_0) = \int f(x_0)\nu^0(dx_0)$$

as $\nu^\varepsilon \Rightarrow \nu^0$. This completes the proof. □

A similar flow-invariance theorem was obtained independently in [3], in the context of urn processes, which, roughly speaking, corresponds to the case where $H(x,\omega)$ is globally bounded.

The first consequence of the flow-invariance theorem above relies on the celebrated Poincaré recurrence theorem (see, e.g., [19, Thm. 3.1, p. 16]) for invariant distributions of a dynamical system. In our setting, it reads as follows.

PROPOSITION 2.3. *The* support *of any flow-invariant distribution ν^0 of ODE_h is contained in the Birkhoff's center $B(\Phi)$ of Φ (the set $B(\Phi)$ is defined as the closure of $\{x \in \mathbb{R}^d \,/\, \exists\, u_n \to +\infty$ with $\Phi(x,u_n) \to x\}$).*

2.2. Some first consequences and examples. The examples below are illustrations of Proposition 2.3.

PROPOSITION 2.4 (Discrete part of ν^0). *Let ν_p^0 denote the purely discrete part of ν^0. Then supp$(\nu_p^0) \subset \{h = 0\}$.*

Proof. Let $x^* \in \operatorname{supp}(\nu_p^0)$. The ν^0-invariance of the flow implies that, for every $u > 0$, $\nu^0(\{\Phi(x^*,u)\}) = \nu^0(\{x^*\}) > 0$. Hence $\{\Phi(x^*,u),\, u > 0\}$ is a path connected countable set containing x^*, that is, $\Phi(x^*,.) \equiv x^*$. Subsequently $h(x^*) = 0$. □

PROPOSITION 2.5 (pseudogradient case). *If V is differentiable on $\mathbb{R}^d$, and $(\nabla V|h) \geq 0$, then*

$$supp(\nu^0) \subset \{(\nabla V|h) = 0\}.$$

Proof. For every $x_0 \in \mathbb{R}^d$, the function $u \mapsto (\nabla V|h)(\Phi(x_0,u))$ converges to 0 as $u \to +\infty$. For every $p \in \mathbb{N}$, let $\varphi_p : \mathbb{R}_+ \to \mathbb{R}_+$ be a continuous nondecreasing function such that

$$\varphi_p(x) := \begin{cases} x, & \text{if } x \in [0,p], \\ p+1, & \text{if } x \geq p+1. \end{cases}$$

One readily checks using the flow-invariance theorem that

$$\forall\, p \in \mathbb{N}, \int \underbrace{\frac{\varphi_p(V(x_0)) - \varphi_p(V(\Phi(x_0,u)))}{u}}_{\geq 0} \nu^0(dx_0) = 0.$$

Letting u go to 0, it follows from Fatou's lemma that $\int \varphi_p'(V)(\nabla V|h)(x_0)\nu^0(dx_0) = 0$ for every $p \in \mathbb{N}$. Letting p go to infinity completes the proof since $\varphi_p' \to \mathbf{1}$. □

We will see below that this result admits an extension to the case where the existence of the flow fails. The next result is a straightforward application of Proposition 2.3.

PROPOSITION 2.6. (a) *Converging dynamics. If ODE_h has a converging flow, i.e., any solution of the ODE_h converges toward (a connected component of) $\{h = 0\}$, then*

$$supp(\nu^0) \subset \{h = 0\}.$$

(b) *Cascade converging dynamics. Assume that there exists a decreasing sequence $(A_n)_{n\in\mathbb{N}}$ of closed subset of $\mathbb{R}^d$ such that $A_1 := \mathbb{R}^d$ and that ODE_h satisfies*

$$\forall\, k \in \mathbb{N},\ \forall\, x_0 \in A_k, \quad \mathrm{dist}(\Phi(x_0,u), A_{k+1}) \to 0 \ \textit{as}\ u \to +\infty.$$

Then $\mathrm{supp}(\nu^0) \subset \cap_n A_n$.

Actually, we do not know any example of an *infinite* cascade of A_n in $\mathbb{R}^d$; for an example of a finite sequence see (2.9) and Proposition 3.5. Several criteria ensure the convergence of the flow without using a Lyapunov function, like the celebrated two-dimensional Poincaré–Bendixson theorem or the Bendixson–Dulac criterion for functions h having a never zero divergence. (See [1] or [11] for some applications to stochastic approximation.) One must mention, too, the cooperative irreducible differential systems introduced by Hirsch in [16]: when the equilibrium set $\{h = 0\}$ is reduced to a single point, the flow of such differential equations is converging. An application to the Kohonen algorithm is carried out in [4].

The Lemniscate example. The algorithm defined by (2.5) below is a typical example, where the decreasing and the constant step algorithms *seem* to behave differently. We will come back to that question further on.

Let $L : \mathbb{R}^2 \to \mathbb{R}$ be the Lemniscate function—$\{L = 0\}$ is a Lemniscate—defined by

$$\forall\, x := (x_1,x_2) \in \mathbb{R}^2, \qquad L(x) := \frac{(x_1^2 + x_2^2)^2}{16} - x_1x_2.$$

Then, set for every $x \in \mathbb{R}^2$

$$V := \frac{L^2}{2(1+L^2)^{\frac{3}{4}}} \text{ and } \theta := \mathcal{H}\left(\frac{L}{(1+L^2)^{\frac{3}{8}}}\right) \left(\mathcal{H}(f) := \begin{pmatrix} \frac{\partial f}{\partial y} \\ -\frac{\partial f}{\partial x} \end{pmatrix} \text{is for Hamiltonian}\right),$$

$$\nabla V(x) = \frac{(L^2+4)}{4(1+L^2)^{\frac{7}{4}}} L(x) \begin{pmatrix} x_1\frac{x_1^2+x_2^2}{4} - x_2 \\ x_2\frac{x_1^2+x_2^2}{4} - x_1 \end{pmatrix},$$

and

$$\theta(x) = \frac{L^2+4}{4(1+L^2)^{\frac{11}{8}}}(x) \begin{pmatrix} x_2\frac{x_1^2+x_2^2}{4} - x_1 \\ -x_1\frac{x_1^2+x_2^2}{4} + x_2 \end{pmatrix}.$$

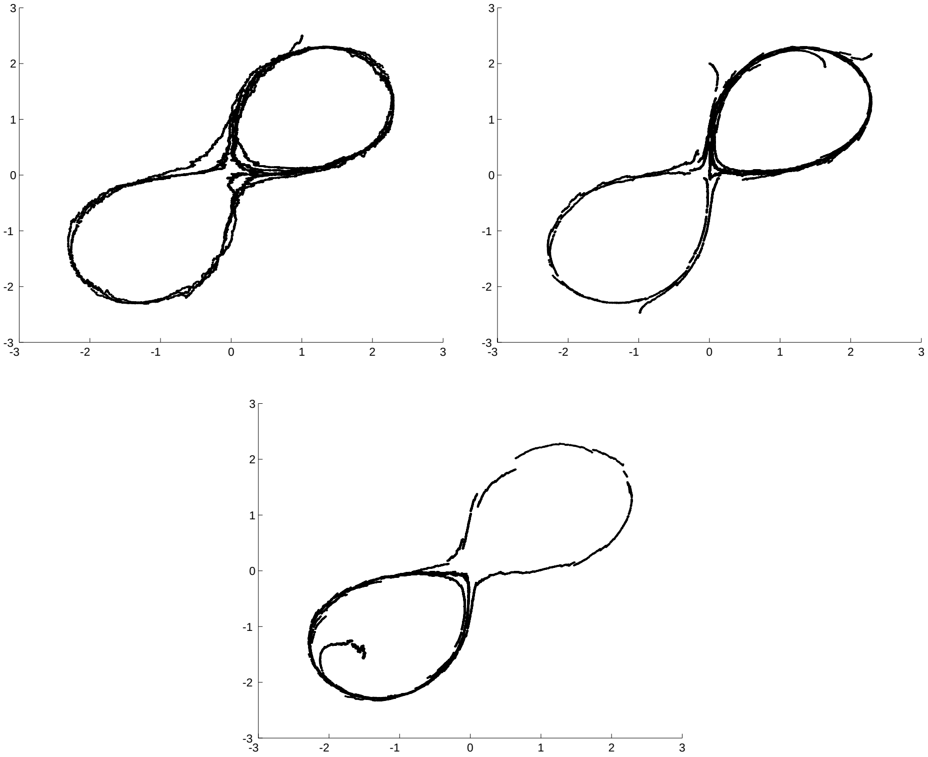


FIG. 1. *Trajectories of the stochastic algorithm* (2.5): $t = 1, \ldots, 10^4$ *and* $\epsilon := 1.5\ 10^{-3}$.

The algorithm (see a simulated path in Figure 1) is defined by

$$(2.5) \qquad h := \nabla V + \theta, \qquad \mu(d\omega) := \frac{d\omega}{(1+|\omega|)^3 \ln^2(2+|\omega|)},$$

$$X^{t+1} := X^t - \varepsilon\left(\underbrace{(I_2 + \varphi(\omega^{t+1}))\,h(X^t) + \mathrm{Diag}[1,\ 0.2]\,\omega^{t+1}}_{:=\,H(X^t,\,\omega^{t+1})}\right),\ \omega^t \overset{\mathcal{L}}{\sim} \mu,$$

where φ is a μ-square integrable and centered 2×2-matrix valued function. In Figure 1 are plotted trajectories of the stochastic algorithm for $t \in \{1, \ldots, 10^4\}$ and $\varepsilon := 1.5 \times 10^{-3}$.

One readily checks that assumption $(\widetilde{Haj})_1$ holds; hence, one derives from Proposition 1.2B the existence of a tight family of invariant distributions $(\nu^{\varepsilon})_{\varepsilon \in (0,\varepsilon_0]}$.

Now, since h is locally Lipschitz, (H_{ODE}) is fulfilled. So is $(UI)_{loc}$ due to $(\widetilde{Haj})_1$. Then one readily checks the following facts:

$$\bullet\ \{(\nabla V | h) = 0\} = \{\nabla V = 0\} = \{L = 0\} \cup \{\nabla L = 0\}$$
$$= \{\nabla V = 0\} = \{L = 0\} \cup \{(-\sqrt{2}; -\sqrt{2}),\ (\sqrt{2}; \sqrt{2})\};$$

• $\{L = 0\} = \{V = 0\}$ is flow-invariant since $V \geq 0$ and, for every $x \in \{L = 0\}$, $\Phi(x, u) \to (0, 0)$ as $u \to +\infty$ (set $x_1 := s$, $x_2 := ts$).

Now applying Proposition 2.6(b) yields the (partial) result

$$\mathrm{supp}(\nu^0) \subset \{(0; 0),\ (-\sqrt{2}; -\sqrt{2}),\ (\sqrt{2}; \sqrt{2})\}.$$

At this stage, a straightforward computation shows that these critical points are of different natures: $(0;0)$ is a saddle point while both $(-\sqrt{2};-\sqrt{2})$, $(\sqrt{2};\sqrt{2})$ are repulsive points as

$$\nabla h(0;0)=\begin{bmatrix}-1 & 0\\ 0 & 1\end{bmatrix},\ \nabla h(\sqrt{2};\sqrt{2})=\nabla h(-\sqrt{2};-\sqrt{2})=\frac{1}{2^{\frac{11}{4}}}\begin{bmatrix}-5 & 2^{-\frac{3}{8}}\\ -2^{-\frac{3}{8}} & -5\end{bmatrix}.$$

The aim of this paper is to investigate whether such repulsive or saddle points can be weighted by ν^0. The answer will be negative for repulsive critical points.

2.3. Location without the ODE. It is possible to get location results for ν^0 even if (H_{ODE}) fails, provided that a global Lyapunov function does exist. As in this paragraph the existence of a *tight* P^ε-invariant family of distributions $(\nu^\varepsilon)_{\varepsilon\in(0,\varepsilon_0]}$ is assumed, assumption (C^μ_ρ) can be relaxed into

$$(C^\mu_{0_+})\equiv\left\{\begin{array}{ll}\text{(i)} & \forall x\in\mathbb{R}^d,\ y\mapsto H(y,\omega)\text{ is }\mu(d\omega)\text{-a.s. continuous at }x,\\ \text{(ii)} & \forall K\subset\mathbb{R}^d,\text{ compact set, }(H(x,\omega))_{x\in K}\text{ is }\mu(d\omega)\text{-uniformly integrable.}\end{array}\right.$$

Note that

$$(\exists\,\rho\in(0,1],\ (C^\mu_\rho))\Longrightarrow(C^\mu_{0_+})\Longrightarrow\left\{\begin{array}{ll}\bullet & P^\varepsilon\text{ is Feller, }\varepsilon\in(0,\varepsilon_0],\\ \bullet & h\text{ is continuous.}\end{array}\right.$$

PROPOSITION 2.7 (Feller setting). *Let* $V:\mathbb{R}^d\xrightarrow{\mathcal{C}^1}\mathbb{R}_+$, *and let* ν^0 *be a weak limiting distribution of* ν^ε *as* $\varepsilon\downarrow 0$. *Assume that the function*

$$\text{(a)}\quad V\textit{ satisfies}\left\{\begin{array}{ll}\text{(i)} & (\nabla V|h)\geq 0,\\ \text{(ii)} & \nabla V\textit{ is bounded or }\lim\limits_{|x|\to+\infty}V(x)=+\infty,\end{array}\right.$$

$$\text{(b)}\quad H\textit{ satisfies }(C^\mu_{0_+})\textit{ and }\int|H(x,\omega)|\mu(d\omega)\textit{ is }(\nu^\varepsilon)_{\varepsilon\in(0,\varepsilon_0]}\textit{-uniformly integrable.}$$

(2.6)

Then, $\mathrm{supp}(\nu^0)\subset\{(\nabla V|h)=0\}$.

The proof of this result relies on a first-order Taylor expansion of the transition $P^\varepsilon(V)(x)$ (see, e.g., [10]). We provide here no further details since a similar method, but with a second-order Taylor expansion of $P^\varepsilon(V)$, will be used in section 3 to treat the case of unstable equilibrium points.

Proposition 2.7 yields the corollary below, which is more tractable for applications.

COROLLARY 2.8. A. PAKES–HAS'MINSKII SETTING. *If* H, h, V, *and the tight family of all the* P^ε*-invariant distributions* ν^ε, $\varepsilon\in(0,\varepsilon_0]$, *satisfy, for some* $\rho\in(0,1]$,

$$(2.7)\quad (Has)_\rho\equiv(\nabla V|h)\geq 0\ \textit{ and }\ (\nabla V|h)^{\frac{1}{1+\rho}}\ \textit{ is }\ (\nu^\varepsilon)_{\varepsilon\in(0,\varepsilon_0]}\textit{-uniformly integrable,}$$

then any weak limiting value ν^0 *of* ν^ε *as* $\varepsilon\downarrow 0$ *satisfies* $\mathrm{supp}(\nu^0)\subset\{(\nabla V|h)=0\}$.

B. HAJEK SETTING. *If* H, h, *and* V *satisfy for some* $\rho\in(0,1]$

$$(2.8)\qquad (\widetilde{Haj})_\rho\ \textit{ and }\ (\nabla V|h)\geq 0,$$

then the same conclusion as in A holds.

The periodic/quasi-cycle example. Consider the linearly perturbed algorithm $H(x,\omega):=h(x)+\omega$ defined by its vector field

$$(2.9)\ \forall\,x:=(x_1,x_2)\in\mathbb{R}^2,\ H(x,\omega):=h(x)+\omega,\ h(x):=\frac{|x|^2-1}{|x|^2+1}x+\varphi(x)\begin{pmatrix}-x_2\\ x_1\end{pmatrix},$$

where φ is a *bounded* continuous function and $(\omega^t)_{t\geq 1}$ is a $(1+\rho)$-integrable two-dimensional white noise. One readily checks that this algorithm is a pseudogradient related to the Lyapunov function

$$V(x) := \frac{1}{1+\rho}\left(|x|^{1+\rho} - 2\ln(1+|x|^{1+\rho})\right).$$

Furthermore, the Hajek criterion $(\widetilde{Haj})_\rho$ is fulfilled using V. Hence, it follows from Proposition 1.2B$'$ and Corollary 2.8 that the set $\{\nu^\varepsilon \,/\, \nu^\varepsilon P^\varepsilon = \nu^\varepsilon,\ \varepsilon\in(0,\varepsilon_0]\}$ is tight for some $\varepsilon_0>0$ and

$$\operatorname{supp}(\nu^0)\subset\{\nabla V|h)=0\}=\mathcal{C}(0;1)\cup\{(0,0)\},$$

where $\mathcal{C}(0;1)$ denotes the unit circle.

If no further assumption is made on the function φ, the existence of the flow of ODE_h may fail, which prevents the use of the flow-invariance theorem. The second feature to be noticed is that the origin $(0,0)$ is a repulsive point for the ODE_h as it is a local maximum of its Lyapunov function V. As in the Lemniscate example, one expects $(0,0)$ not to be weighted by a limiting distribution ν^0 since the random perturbation $H((0,0),\omega)=\omega$ is nonzero at this point.

The asymptotic behavior of this algorithm will be described in subsection 3.1.5 once the case of repulsive points and sets has been treated.

Proposition 2.7 admits a non-Feller version in which the regularity requirements on V are slightly strengthened. Namely, assumption (2.6) then becomes

$$(2.6')\qquad \exists\,\rho\in(0,1]\ \textit{such that}\ \left\{\begin{array}{ll} (\mathrm{a}') & \left\{\begin{array}{ll} (\mathrm{i}) & (\nabla V|h)\geq 0\ \textit{and is lower semi-}\\ & \textit{continuous,}\\ (\mathrm{ii}) & \nabla V \text{ is bounded or}\\ & \displaystyle\lim_{|x|\to+\infty} V(x)=+\infty,\\ (\mathrm{iii}) & \nabla V\ \textit{is}\ \rho'\textit{-Hölder on compact}\\ & \textit{sets for some}\ \rho'\geq\rho,\end{array}\right.\\ (\mathrm{b}') & \textit{the function}\ x\mapsto \displaystyle\int|H(x,\omega)|^{1+\rho}\mu(d\omega)\\ & \textit{is}\ L^1(\nu^\varepsilon)\textit{-bounded as}\ \varepsilon\downarrow 0.\end{array}\right.$$

3. About the support of ν^ε as $\varepsilon\downarrow 0$ near an unstable equilibrium point of the ODE. In the previous section the support of any limiting value ν^0 of ν^ε as $\varepsilon\downarrow 0$ was located under some reasonable assumptions. For instance, when ODE_h has a converging flow (resp., when there exists a global Lyapunov function V), it was shown that $\operatorname{supp}(\nu^0)\subset\{h=0\}$ (resp., $\operatorname{supp}(\nu^0)\subset\{(\nabla V|h)=0\}$). We know (see [7], [21]) that under suitable assumptions on the (decreasing) steps and on the variance function of $H(x,\omega)$, the corresponding algorithm a.s. cannot converge toward an unstable stationary point of ODE_h (i.e., to some maximum or saddle point of V).

The aim of this section is to try to obtain the same type of result. However, we will see further on that the global dependency of ν^ε with respect to the *global* behavior of the solutions of ODE_h does not lead to the same results.

The key of this section is a proposition based on a second-order Taylor expansion of the transition.

PROPOSITION 3.1. *Let $(\nu^\varepsilon)_{\varepsilon\in(0,\varepsilon_0]}$ be a tight family of $(P^\varepsilon)_{\varepsilon\in(0,\varepsilon_0]}$-stationary* *distributions and let ν^0 be one of its limiting distributions as $\varepsilon \downarrow 0$. Assume that*

$$\text{(a)}\ (C^{\mu}_{1_+}) \equiv \left\{ \begin{array}{ll} \text{(i)} & \forall\, x \in \operatorname{supp}(\nu^0),\ y \mapsto H(y,\omega) \textit{ is } \mu(d\omega)\textit{-a.s. continuous at } x, \\ \text{(ii)} & \left\{ \begin{array}{l} \forall\, K \subset \mathbb{R}^d,\ K \textit{ compact set, } (|H(x,\omega)|^2)_{x\in K} \textit{ is} \\ \mu(d\omega)\textit{-uniformly integrable,} \end{array} \right. \end{array} \right.$$

$$\text{(b)}\ x \mapsto \int |H(x,\omega)|^2 \mu(d\omega) \textit{ is } (\nu^{\varepsilon})_{\varepsilon\in(0,\varepsilon_0]} \textit{ uniformly integrable.}$$

Then, for every $f \in \mathcal{C}^2(\mathbb{R}^d,\mathbb{R})$ such that f and $\nabla^2 f$ are bounded and for every sequence $\nu^{\varepsilon_p} \Rightarrow \nu^0$, $\int(\nabla f|h)d\nu^0 = 0$ and

$$\lim_p \int (\nabla f|h)\frac{d\nu^{\varepsilon_p}}{\varepsilon_p} = \frac{1}{2}\int \nu^0(dx)\left(\int \mu(d\omega)\, {}^tH(x,\omega)\nabla^2 f(x) H(x,\omega)\right).$$

Remark. Assumption $(C^{\mu}_{1_+})$ is fulfilled whenever $C^{\mu}_{1+\rho}$(i) holds and

$$\sup_{x\in K}\int |H(x,\omega)|^{2+\rho}\mu(d\omega) < +\infty \text{ for every compact set } K \text{ of } \mathbb{R}^d.$$

Proof. Using a second-order Taylor–Lagrange expansion of the bounded $\mathcal{C}^2$ function f between x and $x - \varepsilon H(x,\omega)$ and successively integrating with respect to μ and ν^{ε} leads to

$$\begin{aligned} \Lambda_{\varepsilon} &:= \left| \varepsilon \int \nu^{\varepsilon}(dx)(\nabla f|h)(x) - \frac{\varepsilon^2}{2}\int \nu^{\varepsilon}(dx)\left(\int \mu(d\omega)\, {}^tH(x,\omega)\nabla^2 f(x)H(x,\omega)\right)\right| \\ &\le \frac{\varepsilon^2}{2}\int \nu^{\varepsilon}(dx)G^{(2)}_{\varepsilon}(x), \end{aligned}$$

where $G^{(2)}_{\varepsilon}(x) := \int \mu(d\omega)|H(x,\omega)|^2 \left(w(\nabla^2 f, x, \varepsilon|H(x,\omega)|) \wedge \|\nabla^2 f\|_{\infty}\right)$ and $w(g,x,.)$ denotes the continuity modulus of g at x.

We note that $G^{(2)}_{\varepsilon}(x)$ is nondecreasing as a function of ε; hence,

$$\forall\, \eta \ge \varepsilon > 0, \qquad \frac{\Lambda_{\varepsilon}}{\varepsilon^2} \le \int \nu^{\varepsilon}(dx)G^{(2)}_{\varepsilon}(x) \le \int \nu^{\varepsilon}(dx)G^{(2)}_{\eta}(x).$$

Assumption $(C^{\mu}_{1_+})$ ensures that $G^{(2)}_{\eta}$ is ν^{ε}-uniformly integrable. Assuming without loss of generality that $\nu^{\varepsilon} \Rightarrow \nu^0$, it follows that

$$\forall\, \eta > 0, \qquad \lim_{\varepsilon\to 0}\int \nu^{\varepsilon}(dx)G^{(2)}_{\eta}(x) = \int \nu^0(dx)G^{(2)}_{\eta}(x).$$

The Lebesgue-dominated convergence theorem yields $\lim_{\eta\to 0}\int \nu^0(dx)G^{(2)}_{\eta}(x) = 0$. This completes the proof of the first equality. The second straightforwardly follows. □

3.1. The case of repulsive equilibrium points of the ODE.

3.1.1. The pointwise result. The next theorem agrees with the intuition that generally speaking no excited enough repulsive point can be weighted when $\varepsilon \downarrow 0$.

THEOREM 3.2. *Let $(\nu^{\varepsilon})_{\varepsilon\in(0,\varepsilon_0]}$ be a tight family of $(P^{\varepsilon})_{\varepsilon\in(0,\varepsilon_0]}$-stationary distributions, and let ν^0 be one of its limiting distributions. Let $x^* \in \mathbb{R}^d$. Assume that the* *assumptions of Proposition* 3.1 *hold and that there exists an open neighborhood* $\mathcal{V}^*$ *of* x^* *and a local Lyapunov function* $V \in \mathcal{C}^2(\mathcal{V}^*, \mathbb{R})$ *such that*

(a) $\forall x \in \mathcal{V}^* \setminus \{x^*\}$, $(\nabla V|h)(x) > 0$ *and* $V(x) < V(x^*)$ (*hence* $\nabla V(x^*) = 0$),
(b) *there exists* $u \in (\mathrm{Ker}\nabla^2 V(x^*))^\perp$, *such that* $\int \mu(d\omega)(H(x^*,\omega)|u)^2 > 0$.

(Assertion (b) *means that the noise has a positive variance in the direction* u*). Then,*

$$\nu^0(\{x^*\}) = 0.$$

Remark. Assumption (b) is equivalent to $\int \mu(d\omega)|\pi(H(x^*,\omega))|^2 > 0$, where π denotes the orthogonal projection on $(\mathrm{Ker}\nabla^2 V(x^*))^\perp$. One may assume without loss of generality that $|u| = 1$. Then let $(e^1, \dots, e^d)$ be an orthonormal basis of $\mathbb{R}^d$ such that $e^1 = u$ and $\langle e^1, \dots, e^p\rangle = (\mathrm{Ker}\nabla^2 V(x^*))^\perp$. Then $\int \mu(d\omega)|\pi(H(x^*,\omega))|^2 = \sum_{i=1}^p \int \mu(d\omega)(H(x^*,\omega)|e^i)^2$.

Proof. The idea is to build a function f, whose support is contained in $\mathcal{V}^*$, that satisfies the assumptions of Proposition 3.1 and for which x^* is also a local maximum. Let $\alpha > 0$ be such that $\overline{B}(x^*, \alpha) \subset \mathcal{V}^*$ and set $K_\alpha := \sup\{V(x),\ \frac{\alpha}{2} \le \|x - x^*\| \le \alpha\}$. Assumption (a) implies $V(x^*) > K_\alpha$. Let $\varphi \in \mathcal{C}^\infty(\mathbb{R}^d, [0,1])$ such that

$$\forall x \in \overline{B}\left(x^*, \frac{\alpha}{2}\right), \quad \varphi(x) := 1 \ \text{ and } \ \forall x \in {}^c B(x^*, \alpha), \quad \varphi(x) := 0.$$

Set for every $x \in \mathbb{R}^d$, $f(x) := (V(x) - K_\alpha)_+^3 \varphi(x)$, where $y_+ := \max(y, 0)$. The function f is clearly $\mathcal{C}^2$ and nonnegative since $y \mapsto (y - K_\alpha)_+^3$ is. Furthermore

$$\nabla f(x) = 3(V(x) - K_\alpha)_+^2 \varphi(x) \nabla V(x) + \underbrace{(V(x) - K_\alpha)_+^3 \nabla \varphi(x)}_{\equiv 0}.$$

Consequently ∇f and $\nabla^2 f$ satisfy

$$(\nabla f|h)(x) = 3(V(x) - K_\alpha)_+^2 \varphi(x)(\nabla V|h)(x) \begin{cases} \ge 0 \text{ everywhere,} \\ > 0 \text{ on } B(x^*, \frac{\alpha}{2}) \setminus \{x^*\}. \end{cases}$$

$$\begin{aligned}\nabla^2 f(x) &= 6(V(x) - K_\alpha)_+ \varphi(x) \nabla V(x)\, {}^t\nabla V(x) + \underbrace{3(V(x) - K_\alpha)_+^2 \nabla V(x)\, {}^t\nabla\varphi(x)}_{\equiv 0} \\ &\quad +3(V(x) - K_\alpha)_+^2 \varphi(x) \nabla^2 V(x) \\ &= 6\big(V(x) - K_\alpha\big)_+ \varphi(x) \nabla V(x)^t \nabla V(x) + 3\big(V(x) - K_\alpha\big)_+^2 \varphi(x) \nabla^2 V(x).\end{aligned}$$

Applying Proposition 3.1 to the above function f first yields

$$\int \nu^0(dx)(V(x) - K_\alpha)_+^2 \varphi(x)(\nabla V|h)(x) = 0.$$

Hence, $\mathrm{supp}(\nu^0) \cap B(x^*, \frac{\alpha}{2}) \subset \{x^*\}$. Consequently,

$$\frac{1}{2}\nu^0(\{x^*\}) \int \mu(d\omega)^t H(x^*,\omega) \nabla^2 f(x^*) H(x^*,\omega) = \lim_p \frac{1}{\varepsilon_p} \int (\nabla f|h)(x) \nu^{\varepsilon_p}(dx) \ge 0.$$

Since $\nabla V(x^*) = 0$ and $V(x^*) > K_\alpha$, it follows that

$$\nu^0(\{x^*\}) \int \mu(d\omega)^t H(x^*,\omega) \nabla^2 V(x^*) H(x^*,\omega) \ge 0.$$

We may assume without loss of generality that the vector $u \in E_- := (\mathrm{Ker}\nabla^2 V(x^*))^\perp$ in assumption (b) satisfies $|u| = 1$. x^* being a strict local maximum of V, $\nabla^2 V(x^*)$ is at least nonpositive (i.e., ${}^tu\nabla^2 V(x^*)u \leq 0$). Set $|\lambda|_{\min} := \min\{|\lambda|,\ \lambda \neq 0,\ \lambda$ eigenvalue of $\nabla^2 V(x^*)\}$. One gets

$$\begin{aligned}{}^tH(x^*,\omega)\nabla^2 V(x^*)H(x^*,\omega) &= {}^t\pi^\perp_{E_-}(H(x^*,\omega))\nabla^2 V(x^*)\pi^\perp_{E_-}(H(x^*,\omega))\\ &\leq -|\lambda|_{\min}|\pi^\perp_{E_-}(H(x^*,\omega))|^2 \leq -|\lambda|_{\min}(H(x^*,\omega)|u)^2,\end{aligned}$$

where $\pi^\perp_{E_-}$ denotes the orthogonal projection on E_-. In turn these inequalities yield

$$\int \mu(d\omega)\,{}^tH(x^*,\omega)\nabla^2 V(x^*)H(x^*,\omega) \leq -|\lambda|_{\min}\int \mu(d\omega)(H(x^*,\omega)|u)^2 < 0,$$

implying that $\nu^0(\{x^*\}) = 0$. □

Remark. The assumptions of Theorem 3.2(a) do not imply that $h(x^*)$ is zero but only that $h(x^*)$ belongs to $\mathrm{Ker}\nabla^2 V(x^*)$. However, if $h(x^*) \neq 0$—and if the assumptions of the flow-invariance theorem (Theorem 2.2) are fulfilled by h—then $\nu^0(\{x^*\}) = 0$ *whatever the noise is*: the discrete part of ν^0 necessarily lies in $\{h = 0\}$ by Proposition 2.4.

A first application. If $x^* \in \{h = 0\}$, h is differentiable at x^*, and all eigenvalues of $\nabla h(x^*)$ have negative real part, then $\nu^0(\{x^*\}) = 0$.

This follows from the quite classical construction (see, e.g., [1]) of an appropriate local Lyapunov function defined by $V(x) := 1 - \int_0^{+\infty} |e^{s\nabla h(x^*)}(x - x^*)|^2 ds$.

3.1.2. The case of higher-order strict local maxima. Theorem 3.2 seems not to be powerful enough to treat the case where the noise exclusively lies in the direction of $(\mathrm{Ker}\nabla^2 V(x^*))^\perp$. This is the case, e.g., for a stochastic algorithm having, near $(0,0)$, $V(x,y) = 1 - (x^2 + y^4)$ as a (local) Lyapunov function. Actually, this question can be solved by simply changing the Lyapunov function.

DEFINITION 3.3. *Let $V : \mathbb{R}^d \to \mathbb{R}_+$ be a function having a (strict) local maximum at $x^* \in \mathbb{R}^d$. The function V has a polynomial local maximum at x^* of order $2p$, $p \in \mathbb{N}^*$, if V is $2p$-differentiable at x^* and*

$$\exists\, \alpha > 0,\ \forall\, u \in \overline{B}(0,\alpha) \setminus \{0\},\quad \sum_{k=2}^{2p} \frac{\nabla^k V(x^*)}{(k-1)!}.u^{(k)} < 0. \tag{3.1}$$

Remark. $\nabla V(x^*)$ is necessarily 0 and it is meaningless to stop at some odd order.

For such maxima, it turns out that $1 - |x - x^*|^2$ is locally a Lyapunov function around x^*.

PROPOSITION 3.4. *Assume that $V : \mathbb{R}^d \to \mathbb{R}_+$ is a local Lyapunov function for h near x^*. If*

(i). *x^* is a strict polynomial local maximum of V whose degree is $2p$,*

(ii). *$h(x) - \nabla V(x) = |x - x^*|^{2p-1}\eta(x)$, $\lim_{x\to x^*}\eta(x) = 0$ (this is fulfilled, e.g., when $h = \nabla V$ near x^*),*
then, $x \mapsto 1 - \frac{1}{2}|x - x^*|^2$ is a Lyapunov function near x^*.

Note that assumption (i) *with $p = 1$ amounts to $\nabla^2 V(x^*)$ being positive.*

Combining this result and Theorem 3.2 straightforwardly solves the following example.

Example. Set $H(x,\omega) := (-\frac{x}{1+|x|^{\frac{5}{2}}}, -y^3(1 + \sqrt{|x|})) + \omega$ and $\omega^t := (0, \omega_2^t)$, $\omega_2^t \sim \mathcal{N}(0,1)$. One readily observes that $V(x,y) := 1 - (x^2 + y^4)$ is a Lyapunov function locally around $(0,0)$, $(0,0)$ being a polynomial maximum of degree 4. Moreover, the noise has its support in $\mathrm{Ker}\nabla^2 V(0,0) := \mathbb{R}(0,1)$. So Theorem 3.2 fails when applied directly but succeeds when passing through Proposition 3.4.

Proof. The Taylor–Young formula applied at x^* to ∇V reads

$$\nabla V(x) = \sum_{k=1}^{2p-1} \frac{\nabla^{k+1}V(x^*)}{k!}.(x-x^*)^{(k)} + |x-x^*|^{2p-1}\varepsilon(x) \quad \textit{with} \quad \lim_{x\to x^*}\varepsilon(x) = 0,$$

hence

$$\begin{aligned} -(x-x^*|h(x)) &= -(x-x^*|\nabla V(x)) + o(|x-x^*|^{2p}) \\ &= \underbrace{-\sum_{k=1}^{2p-1} \frac{\nabla^{k+1}V(x^*)}{k!}.(x-x^*)^{(k+1)}}_{<0 \text{ if } |x-x^*|\le\alpha} + o(|x-x^*|^{2p}). \end{aligned}$$

Finally, it follows that $\exists\, \alpha^* \in (0,\alpha]$ and $\rho^* > 0$ s.t.

$$\forall\, x \in B(x^*,\alpha^*), \qquad -(x-x^*|h(x)) \ge \rho^*|x-x^*|^{2p}. \qquad \square$$

Further comments. One must keep in mind that the square Euclidean norm may be a local Lyapunov function near an obviously nonpolynomial strict local maximum of a function V. This is the case, e.g., when (in any dimension d) $h(x) = V(x) := 1 - \exp(-\frac{1}{|x|^2})$. One readily checks that all the derivatives of V at 0 are 0 and that $1 - \frac{1}{2}|x|^2$ is a strict Lyapunov function for h near 0.

Conversely, the square Euclidean norm is not always a Lyapunov function! Indeed, if

$$V(x) := -\left(\frac{1+\sqrt{2}}{2} + \sin\frac{1}{|x|^2}\right)\exp\left(-\frac{1}{|x|^2}\right) \qquad \text{and} \qquad h := \nabla V,$$

V is a C^∞ function whose all derivatives are bounded on $\mathbb{R}^d$, having a local/global maximum at 0. However,

$$(h(x)|x) = -\left(1+\sqrt{2}+2\sqrt{2}\cos\left(\frac{1}{|x|^2}-\frac{\pi}{4}\right)\right)|x|^{-2}\exp\left(-\frac{1}{|x|^2}\right)$$

whose sign is obviously not constant near 0. In fact, the only way to make up such an example is to choose functions V having infinitely many *undulations* arbitrarily close to its local maximum.

3.1.3. Extension to thin zero sets. A straightforward generalization of Theorem 3.2 is obtained by considering a compact connected set χ^*, instead of an isolated point x^*, provided that χ^* is *thin*, i.e., $\chi^* = \partial\chi^*$.

Assume that there exists a neighborhood $\mathcal{V}^*$ of χ^* and $V \in \mathcal{C}^2(\mathcal{V}^*,\mathbb{R}_+)$ such that

$$\forall\, x \in \chi^*,\ V(x) = v^* \ \text{ and } \ \forall\, x \in \mathcal{V}^* \setminus \chi^*,\ (\nabla V|h)(x) > 0 \text{ and } V(x) < v^*.$$

If, for every $x^* \in \chi^*$, there exists $u(x^*)$ such that

$$\text{(3.2)} \qquad u(x^*) \in \left(\mathrm{Ker}\nabla^2 V(x^*)\right)^\perp \quad \text{and} \quad \int (H(x^*,\omega)|u(x^*))^2\mu(d\omega) > 0,$$

then

$$\nu^0(\chi^*) = 0.$$

Example. Following the notations of subsection 2.2, let

$$h(x) = -\frac{|x|^2 - 1}{|x|^2 + 1}\psi(x) + \varphi(x)\begin{pmatrix} -x_2 \\ x_1 \end{pmatrix},$$

where $\psi : \mathbb{R}^2 \to \mathbb{R}^2$ is $\mathcal{C}^2$, $\psi(x) = x + o(|x| - 1)$ near the unit circle $\mathcal{C}(0;1)$, and $\psi(x) = x + o(|x|)$ when $|x| \to +\infty$. Then $\mathcal{C}(0;1)$ is a repulsive cycle and the above extension of Theorem 3 applies.

3.1.4. The Lemniscate example: Continuation. It was shown in the former Lemniscate example (section 2.2) that any weak limiting value ν^0 of the invariant distributions of the algorithm defined by (2.5) satisfies

$$\operatorname{supp}(\nu^0) \subset \{(0;0),\, (\sqrt{2};\sqrt{2}),\, (-\sqrt{2};-\sqrt{2})\},$$

where the last two points are repulsive stationary points for $ODE_h \equiv \dot{x} = -h(x)$. Following Theorem 3.2, namely assumption (b)(ii) in Proposition 3.1, one derives that

$$\varphi \in L^\infty(\mu) \implies \nu^0(\{(\sqrt{2};\sqrt{2}),\, (-\sqrt{2};-\sqrt{2})\}) = 0.$$

Consequently, $\operatorname{supp}(\nu^0) = \{(0;0)\}$, i.e.,

$$\nu^\varepsilon \overset{(\mathbb{R}^d)}{\Longrightarrow} \delta_{(0;0)} \quad \text{although } (0;0) \text{ is an unstable point.} \tag{3.3}$$

This result stresses the fact that the asymptotic behavior of ν^ε near an unstable point x^* as $\varepsilon \downarrow 0$ depends on the global behavior of the deterministic underlying flow of the dynamical system and on the local properties of the random perturbation term $H(x^*,\omega) - h(x^*) = H(x^*,\omega)$ at x^*.

For instance, in Figure 2 are plotted three histograms of $|X(t)|$ for $t \in \{1,\dots,2\times 10^5\}$, with $\varepsilon := 1.5 \times 10^{-4}$, $\varepsilon := 0.75 \times 10^{-4}$, and $\varepsilon := 0.3 \times 10^{-4}$. ν^ε clearly concentrates at $(0;0)$.

Comparison with the algorithm with decreasing step. The same algorithm, implemented with a regular decreasing step, will a.s. not converge. As a matter of fact, following, e.g., [11] or [2], one derives that, a.s., its limiting value set $\mathcal{X}^\infty$ makes up a flow-invariant compact connected subset of the Lemniscate $\{L = 0\}$ (hence containing $(0;0)$). Then calling upon [7] or [21] implies that it cannot converge to $(0;0)$ as it is a point at which the noise is not degenerate. So $\mathcal{X}^\infty$ is one of the two loops, or the whole Lemniscate $\{L = 0\}$ depending on the noise structure at $(0;0)$.

Actually, the constant step algorithm behaves the same way. What the convergence in (3.3) says is that the algorithm spends infinitely more time in any neighborhood of the origin than anywhere else.

3.1.5. The periodic/quasi-cycle example: Continuation. Let us come back to algorithm (2.9). We saw using the nonfunctional approach that whenever the function φ is continuous and bounded, if $\mathbb{E}(|\omega|^{1+\rho}) < +\infty$, then $\operatorname{supp}(\nu^0) \subset \mathcal{C}(0;1) \cup \{(0;0)\}$.

If, moreover, $0 < \mathbb{E}(|\omega|^2) < +\infty$, then Theorem 3.2 implies that $\nu^0((0;0)) = 0$ as $(0;0)$ is an isolated local maximum of the Lyapunov function V_2 whose Hessian is $\nabla^2 V_2(0;0) = I_2$.

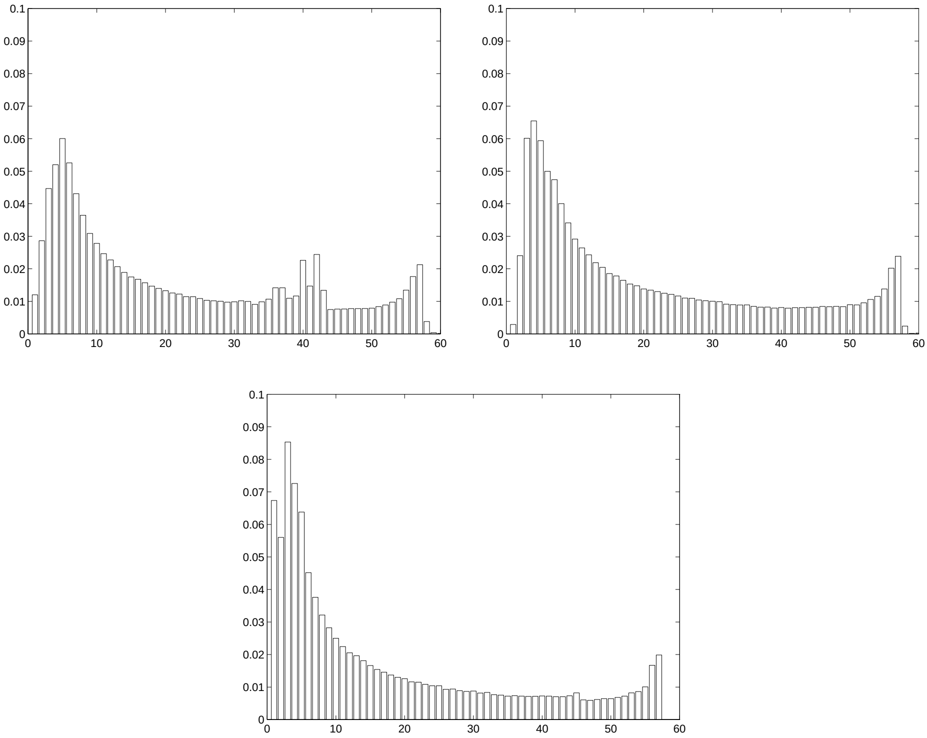


FIG. 2. *Histograms of time spent in* $\{5\times 10^{-2}k \le\mid x \mid\le 5\times 10^{-2}(k+1)\}, 0\le k\le 59$ *with time* $t\in\{1,\dots,2\times 10^5\}$*, and for* $\epsilon := 1.5\times 10^{-4}, \epsilon := 0.75\times 10^{-4}, \epsilon := 0.3\times 10^{-4}$.

PROPOSITION 3.5. *Assume that* φ *is continuous and bounded and if* $0<\mathbb{E}(|\omega|^2)<+\infty$*, then either*

(a) *if* φ *is never* 0 *on* $\mathcal{C}(0;1)$*, then*

$$\nu^\varepsilon \overset{(\mathbb{R}^d)}{\Longrightarrow} \nu^0 := \frac{\frac{1}{\varphi}(e^{i\theta_0})}{\int_0^{2\pi}\frac{1}{\varphi}(e^{i\theta})d\theta}\mathbf{1}_{[0,2\pi[}(\theta_0)d\theta_0, \quad \textit{or}$$

(b) *if* $\int_0^{2\pi}\frac{1}{\varphi}(e^{i\theta})d\theta=+\infty$ *then* $\operatorname{supp}(\nu^0)\subset\{\varphi_{|\mathcal{C}(0;1)}=0\}$.

Proof. (a) Assumption $(UI)_{loc}$ is obviously fulfilled by the algorithm. On the other hand, $ODE_h \equiv \dot{x}=-h(x)$ reads in polar coordinates $((x_1,x_2):=(r\cos\theta, r\sin\theta))$:

$$\dot{r}=-r\frac{r^2-1}{r^2+1} \qquad \text{and} \qquad \dot{\theta}=-\varphi(re^{i\theta}).$$

The unit circle $\mathcal{C}(0;1)$ is flow-invariant for ODE_h, on which the flow is well defined since φ is never 0 on $\mathcal{C}(0;1)$ and the differential equation $ODE_\theta\equiv\dot{\theta}=-\varphi(e^{i\theta})$ is scalar. A straightforward extension of Theorem 2.2 shows that ν^0 is still flow-invariant. Now, it is known that ODE_θ is uniquely ergodic with invariant distribution

$$\nu^0(d\theta_0)=\mathbf{1}_{[0,2\pi)}\frac{\frac{1}{\varphi}(e^{i\theta_0})d\theta_0}{\int_0^{2\pi}\frac{1}{\varphi}(e^{i\theta_0})d\theta_0}.$$

The proof is reproduced for the reader's convenience: let $T > 0$ and $v \in [0,T]$. The invariance property yields, for every $n \in \mathbb{N}^*$,

$$\begin{aligned}\int_0^{2\pi} e^{in\theta_0}\varphi(e^{i\theta_0})\nu^0(d\theta_0) &= \int_0^{2\pi} e^{in\theta(\theta_0,v)}\varphi(e^{i\theta(\theta_0,v)})\nu^0(d\theta_0)\\ &= \int_0^{2\pi}\frac{1}{T}\int_0^T e^{in\theta(\theta_0,v)}\varphi(e^{i\theta(\theta_0,v)})dv\,\nu^0(d\theta_0)\\ &= \int_0^{2\pi}\frac{1}{T}\frac{e^{in\theta(\theta_0,T)}-e^{in\theta(\theta_0,0)}}{-in}\nu^0(d\theta_0)\overset{T\to+\infty}{\longrightarrow} 0.\end{aligned}$$

The Fourier transform being one to one, $\varphi(e^{i\theta_0})\nu^0(d\theta_0) = c\lambda_{|[0,2\pi)}(d\theta_0)$, i.e.,

$$\nu^0(d\theta_0) = \mathbf{1}_{[0,2\pi)}\frac{\frac{1}{\varphi}(e^{i\theta_0})d\theta_0}{\int_0^{2\pi}\frac{1}{\varphi}(e^{i\theta_0})d\theta_0}.$$

Conversely, ν^0 actually is flow-invariant.

(b) Since φ does have zeros on $\mathcal{C}(0;1)$, the flow $\Phi(\theta,u)$ of ODE_θ is well defined and it converges to some zero of φ. Once again, it is straightforward to show that the flow invariance of ν^0 still holds in the following way: for every Borel set A and every $u \geq 0$, $\nu^0(A) = \Phi(\nu^0,u)(A)$. As any limiting distribution of $\Phi(\nu^0,u)$ is supported by $\{\varphi_{|\mathcal{C}(0;1)} = 0\}$, it follows that $\nu^0(A) = 0$, i.e., $\mathrm{supp}(\nu^0) \subset \{\varphi_{|\mathcal{C}(0;1)} = 0\}$. □

Remark. If $\varphi \geq 0$ does have zeros on $\mathcal{C}(0;1)$ while the integral

$$C_\varphi := \int_0^{2\pi}\frac{1}{\varphi}(e^{i\theta_0})d\theta_0 < +\infty, the probability measure \quad \frac{1}{C_\varphi}\frac{1}{\varphi(e^{i\theta_0})}\mathbf{1}_{[0,2\pi[}(\theta_0)\,d\theta_0$$

is still an invariant distribution of ODE_θ.

Some further investigations can be carried out by studying the nature of the critical points of $\{h_{|\mathcal{C}(0;1)} = 0\}$. Furthermore, if the distribution of the white noise is known and admits moments enough (Gaussian, etc.), one may restrict again the support (see, e.g., [23] in a continuous time-diffusion setting) using, e.g., some large deviation methods.

3.2. Saddle points of a stochastic pseudogradient. It was emphasized in section 3.1.4 that a saddle critical point x^* of the mean function h can be weighted by a limiting distribution ν^0 even if the random perturbation does not fade at x^*. The aim of this section is to show that, in many situations, this cannot occur—namely, when the algorithm is a *stochastic pseudogradient* with a $\mathcal{C}^2$ Lyapunov function V satisfying $\{(\nabla V|h) = 0\} = \{h = 0\}$ and x^* is a saddle point of V. Many learning algorithms issued from neural networks are stochastic pseudogradient descents. They are often performed with a (small) constant step (Boltzmann machine, back-propagation algorithm for the multilayer perceptron, and others) precisely to avoid false convergence phenomena to metastable critical points.

The fact that an algorithm is a stochastic pseudogradient provides the information about the *global behavior* of the flow of its ODE_h. For the sake of simplicity, we will assume that there is no spurious equilibrium (see section 3.3). The case of possible spurious points will be investigated later.

Before getting into the technicalities let us explain the method developed in this section. Basically, it consists once again in modifying the global Lyapunov function V near some critical points. Namely, we will flatten V at the local minima and nearly pull down to zero the positive part of the spectra of the Hessian of V near the saddle points. To this end we need to write V locally as a quadratic form up to some $\mathcal{C}^2$-diffeomorphic change of coordinates. The Morse lemma (see, e.g., [14]) will be called upon to ensure this is possible.

Proposition 3.6 (Morse). *Let $g \in \mathcal{C}^3(\mathcal{V}^*, \mathbb{R})$, where $\mathcal{V}^*$ is a neighborhood of x^*. Assume that $g(x^*) = 0$, $\nabla g(x^*) = 0$, and $\nabla^2 g(x^*)$ is invertible. Then there exist $\varepsilon_0 > 0$ and $u \in \mathcal{C}^2(B(x^*, \varepsilon_0), \mathbb{R}^d)$ such that*

$$u(x^*) = 0, \quad u'(x^*) = I_d, \qquad \textit{and} \qquad \forall x \in B(x^*, \varepsilon_0),\ g(x) = \frac{1}{2}\,{}^t u(x) \nabla^2 g(x^*) u(x).$$

Notation. From now on, $E_x^{\pm}$ will denote the vector spaces respectively spanned by eigenvectors of $\nabla^2 V(x)$ with positive and nonpositive eigenvalue.

Theorem 3.7. *Let $(\nu^\varepsilon)_{\varepsilon \in (0, \varepsilon_0]}$ be a tight family of $(P^\varepsilon)_{\varepsilon \in (0,\varepsilon_0]}$-stationary distributions, and let ν^0 be one of its limiting distributions as $\varepsilon \downarrow 0$.*

Assume that H, μ, and the ν^ε's fulfill assumptions (a) *and* (b) *of Proposition* 3.1 *and that there exists a global $\mathcal{C}^2$ Lyapunov function $V : \mathbb{R}^d \to \mathbb{R}_+$ satisfying*

$$(\mathcal{L}) \equiv \begin{cases} \text{(i)} & (\nabla V | h) \geq 0 \ \textit{ and } \ \{(\nabla V | h) = 0\} = \{\nabla V = 0\}, \\ \text{(ii)} & \forall\, K > 0,\ \sup_{\{V \leq K\}} |\nabla V|^2 + \|\nabla^2 V\|^2 + |h|^2 < +\infty. \end{cases}$$

(a) Local result. *Let $x_0^* \in \{\nabla V = 0\}$ and $v_0^* := V(x_0^*)$. Assume that there is some $\eta_0^* > 0$ such that*

(i) *the set $\mathcal{C}_0^* := \{\nabla V = 0\} \cap \{V \in (v_0^* - \eta_0^*, v_0^* + \eta_0^*)\}$ is finite;*

(ii) *for every $x^* \in \mathcal{C}_0^*$, x^* is*

$$\begin{cases} (\alpha) & \textit{a local maximum of } V, \\ (\beta) & \textit{or a local minimum of } V, \\ (\gamma) & \textit{or} \begin{cases} \textit{a } \mathcal{C}^3\textit{-Morse saddle point } (V \textit{ is locally } \mathcal{C}^3,\ \nabla^2 V(x^*) \textit{ is invertible}) \\ \textit{satisfying } (\nabla V | h)(x) \geq c_* |\nabla V|^2(x) \textit{ for some } c_* > 0 \\ \textit{and } \theta(x) := h(x) - \frac{(\nabla V|h)}{|\nabla V|^2}(x) \nabla V(x) = O(|x - x^*|^{1+a_*}) \textit{ for some } a_* > 0; \end{cases} \end{cases}$$

(iii) *there is some $v \in E^-_{x_0^*}$ such that $\int (H(x_0^*, \omega) | v)^2 \mu(d\omega) > 0$ (which implies that x_0^* is not a local minimum);*
then, $\nu^0(\{x_0^*\}) = 0$.

(b) Global result. *Assume that $\{\nabla V = 0\}$ is V-locally finite (i.e., for every $v > 0$, $\{\nabla V = 0\} \cap \{V \leq v\}$ is finite) and that, furthermore, it is made up exclusively of points of type (α), (β), and (γ). If every local maximum satisfies the assumption of Theorem* 3.2 *(possibly with a modified local Lyapunov function) and if the above assumption* (iii) *holds at every saddle point, then*

$$\operatorname{supp}(\nu^0) \subset \{\text{local minima of } V\}.$$

Remark. Actually, a fourth category of points could have been added in the above results (a) or (b) made up of all the points that can be "rejected" by any appropriate method: this category contains, e.g., the excited *thin* sets or the spurious points that satisfy the assumptions of Proposition 3.8 below.

Proof. Step 1 (*flattening of the local minima of $\mathcal{C}_0^*$*). Let $x_0^* \in \mathcal{C}_0^*$ be a local minimum. Being isolated, there is some $\delta_* > 0$ such that $V(x) > V(x^*)$ for every $x \in B(x^*, \delta_*) \setminus \{x^*\}$.

Let $m_* := \min\{V(x),\, x \in \partial B(x^*, \delta_*)\}$, $K_* \in (V(x^*), m_*)$, and let φ^* be a nonnegative increasing C^2 function defined on $[V(x^*), +\infty)$ satisfying

$$(\varphi^*)'(V(x^*)) = 0, \quad (\varphi^*)'(v) > 0 \text{ if } v > V(x^*), \qquad \text{and} \qquad \varphi^*_{|[K_*,+\infty)}(v) = v.$$

Then, we set

$$V^*(x) := \left\{ \begin{array}{ll} \varphi^*(V(x)) & \text{if } x \in B(x^*, \delta_*), \\ V(x) & \text{if } x \notin B(x^*, \delta_*). \end{array} \right.$$

One may assume without loss of generality that this modification is carried out around every local minimum contained in $\mathcal{C}_0^*$ so that the different (closed) balls $\overline{B}(x^*, \delta_*)$ remain pairwise distinct and meet no other critical point of V. The function V^* is then C^2, V^* and V have the same critical points and, for every one of them,

$$\left\{ \begin{array}{ll} \nabla^2 V^*(x^*) = (\varphi^*)'(V(x^*))\nabla^2 V(x^*) = 0 & \text{if } x^* \text{ is a local minimum in } \mathcal{C}_0^*, \\ \nabla^2 V^*(x^*) = \nabla^2 V(x^*) & \text{either.} \end{array} \right.$$

Step 2 (*lowering of the positive part of the spectrum of Morse saddle points of* $\mathcal{C}_0^*$). Let $x^* \in \mathcal{C}_0^*$ be a Morse saddle point. Set $A := \nabla^2 V(x^*)$, $A^{\pm} := \pm \pi_{E^{\pm}_{x^*}} A \, \pi_{E^{\pm}_{x^*}}$, $|\lambda|^* := \max\{|\lambda|,\, \lambda \text{ eigenvalue of } A\} = \|A\|$, $|\lambda|_* := \min\{|\lambda|,\, \lambda \text{ eigenvalue of } A\}$, and $|x|_*^2 := (x|(A^+ + A^-)x)$. $|\,.\,|_*^2$ defines an Euclidean norm.

It follows from the Morse lemma that there is some $\delta_* > 0$ and a C^2-diffeomorphism u: $B(x^*, \delta_*) \to \mathcal{N}_0$, $\mathcal{N}_0$ open neighborhood of 0, such that

$$\forall\, x \in B(x^*, \delta_*),\ V(x) = V(x^*) + \frac{1}{2}(Au(x)|u(x)),\ u(x^*) = 0 \ \text{ and } \ u'(x^*) = I_d. \tag{3.4}$$

Furthermore, by properly lowering δ_*, one may assume that there is some $K > 0$ satisfying

$$\begin{array}{l} u' \text{ is } K\text{-Lipschitz on } B(x^*, \delta_*) \text{ with } K\delta_* < 1/4, \\ \forall\, x \in B(x^*, \delta_*),\ |x - x^*|/2 \le |u(x)|_* \le 2|x - x^*|, \\ \forall\, x \in B(x^*, \delta_*),\ |\theta(x)| \le K|x - x^*|^{1+a_*} \text{ and } (\nabla V|h)(x) \ge c_* |\nabla V|^2(x), \\ \text{assuming without loss of generality that } a_* \in (0, 1]. \end{array} \tag{3.5}$$

Now, let

$$\gamma \in \left(0, \left(\frac{\delta_*}{4}\right)^{a_*} \times \frac{4(1 + K\delta_*)\left(2Kc_*(\frac{\delta_*}{4})^{1-a_*} + \frac{2^{1+a_*}}{|\lambda|_*}\right)}{c_*(\frac{1}{2} - K\delta_*)}\right]$$

and $\rho_* : \mathbb{R}_+ \to \mathbb{R}_+$ be a nonincreasing, nonnegative C^∞ function satisfying $\rho_*(0) := \frac{1-\gamma/2}{1-\gamma}$, $\rho^{(k)}(1) = 0$, $k \in \mathbb{N}$, and $\int_0^1 \rho_* = 1$. Then, the function φ_* defined by $\varphi_*(v) = \gamma + (1 - \gamma)\int_0^v \rho_*(u)du$ satisfies $\varphi_*(0) = \gamma$, $\varphi_*(1) = 1$, $\varphi_*^{(k)}(1) = 0$, $k \ge 1$, and $\|\varphi_*'\|_\infty = \frac{1-\gamma/2}{1-\gamma}$. Now, we are in the position to modify V^* around x^* by setting

$$V^*_\gamma(x) \\ = \begin{cases} V(x^*) + \frac{1}{2}\varphi_*\left(\frac{|u(x)|_*^2}{\alpha_*^2}\right)(A^+u|u)(x) - (A^-u|u)(x) & \text{if } |x - x^*| \le \delta_*,\ |u(x)|_* \le \alpha_*, \\ V^*(x) & \text{if } |u(x)|_* \ge \alpha_* \text{ or } |x - x^*| \ge \delta_*, \end{cases}$$

where

$$\alpha_* := \left(\gamma \frac{c_*(\frac{1}{2} - K\delta_*)}{4(1+K\delta_*)\left(2Kc_*(\frac{\delta_*}{4})^{1-a_*} + \frac{2^{1+a_*}}{|\lambda|_*}\right)}\right)^{\frac{1}{a_*}}.$$

As $\alpha_* \le \frac{\delta_*}{4}$ by construction, it follows that $\{x \,/\, |u(x)|_* \le \alpha_*\} \subset B(x^*, \delta_*/2)$, which ensures that the function V^*_γ is well defined as a C^2 function. Next point to be checked is the global Lyapunov property of V^*_γ.

Set for every $x \in B(x^*, \delta_*)$, $M(x) :=^t u'(x) - I_d$. u' being K-Lipschitz, one gets $\|M(x)\| \le K|x - x^*|$. Then, a little algebra on derivatives yields, still for every $x \in B(x^*, \delta_*)$,

$$\text{(3.6)} \qquad \left\{ \begin{array}{lcl} \nabla V(x) &=& (I_d + M(x))(A^+u(x) - A^-u(x)), \\ \nabla V^*_\gamma(x) &=& (I_d + M(x))(a_+(x)A^+u(x) - a_-(x)A^-u(x)), \end{array} \right.$$

with

$$\left\{ \begin{array}{ll} a_+(x) & := \varphi_*\left(\frac{|u(x)|^2_*}{\alpha^2_*}\right) + \frac{(A^+u|u)(x)}{\alpha^2_*}\varphi'_*\left(\frac{|u(x)|^2_*}{\alpha^2_*}\right), \\ a_-(x) & := 1 - \frac{(A^+u|u)(x)}{\alpha^2_*}\varphi'_*\left(\frac{|u(x)|^2_*}{\alpha^2_*}\right). \end{array} \right.$$

One readily sees that $a_+(x) + a_-(x) \le 2$, $a_+(x) \ge \gamma$, and $a_-(x) \ge 1 - (1-\gamma)\rho_*(0) = \frac{\gamma}{2}$, hence $\min(a_+, a_-) \ge \frac{\gamma}{2}$.

Then, it follows from formula (3.6) and the orthogonality relation $(A^+u(x)|A^-u(x)) = 0$ that, whenever $|u(x)|_* \le \alpha_*$,

$$\begin{aligned} (\nabla V^*_\gamma|\nabla V)(x) &\ge |Au(x)|^2\left(\frac{\gamma}{2}(1 - 2\|M(x)\|) - 2\|M(x)\|(1 + \|M(x)\|)\right) \\ &\ge |Au(x)|^2\left(\frac{\gamma}{2}(1 - 2K\delta_*) - 2K(1+K\delta_*)|x - x^*|\right) \\ &\ge |Au(x)|^2\left(\frac{\gamma}{2}(1 - 2K\delta_*) - 4K(1+K\delta_*)\alpha_*\right) \end{aligned}$$

and

$$\begin{aligned} |(\nabla V^*_\gamma|\theta)|(x) &\le |\nabla V^*_\gamma(x)||\theta(x)| \\ &\le (1 + \|M(x)\|)|a_+(x)A^+u(x) - a_-(x)A^-u(x)|\, K|x - x^*|^{1+a_*} \\ &\le 4(1+K\delta_*)|Au(x)|^2\frac{K}{|\lambda|_*}(2\alpha_*)^{a_*}. \end{aligned}$$

Finally, whenever $|u(x)|_* \le \alpha_*$,

$$\begin{aligned} &(\nabla V^*_\gamma|h)(x) \\ &\quad = \frac{(\nabla V|h)}{|\nabla V|^2}(x)(\nabla V^*_\gamma|\nabla V)(x) + (\nabla V^*_\gamma|\theta)(x) \\ &\quad \ge |Au(x)|^2\left(c_*\left(\frac{\gamma}{2}(1 - 2K\delta_*) - 4K(1+K\delta_*)\alpha_*\right) - 4(1+K\delta_*)\frac{(2\alpha_*)^{a_*}}{|\lambda|_*}\right) \\ &\quad \ge |Au(x)|^2\left(c_*\frac{\gamma}{2}(1 - 2K\delta_*) - 2(1+K\delta_*)\left(2Kc_*\left(\frac{\delta_*}{4}\right)^{1-a_*} + \frac{2^{1+a_*}}{|\lambda|_*}\right)\alpha_*^{a_*}\right) \\ &\quad \ge |Au(x)|^2 c_*\frac{\gamma}{4}(1 - 2K\delta_*) \ge 0. \end{aligned}$$

Hence $(\nabla V^*_\gamma | h) \geq 0$ everywhere since $V^* = V^*_\gamma$ when $|u(x)|_* > \alpha_*$. Furthermore

$$\nabla^2 V^*_\gamma(x^*) = \gamma A^+ - A^-.$$

As for the local minima, one may assume that this modification is made at every Morse saddle point of $\mathcal{C}^*_0$ and that all the closed balls $\overline{B}(x^*, \delta_*)$ provided by these first two steps have a pairwise empty intersection.

Step 3 (*localization*). Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be a $\mathcal{C}^\infty$ nondecreasing function whose derivative has exactly $[v^*_0 - \eta^*_0, v^*_0 + \eta^*_0]$ as its support. Set $W^*_\gamma := \psi \circ V^*_\gamma$. Then

$$\nabla W^*_\gamma = \psi'(V^*_\gamma)\nabla V^*_\gamma \quad \text{and} \quad \nabla^2 W^*_\gamma = \psi''(V^*_\gamma)(\nabla V^*_\gamma)\,{}^t(\nabla V^*_\gamma) + \psi'(V^*_\gamma)\nabla^2 V^*_\gamma.$$

It follows from assumption $(\mathcal{L})$ that W^*_γ meets the following properties:

$(\nabla W^*_\gamma | h)$ is nonnegative and bounded,
$\nabla^2 W^*_\gamma$ is bounded,
$\{(\nabla W^*_\gamma | h) = 0\} \cap \{V \in (v^*_0 - \eta^*_0, v^*_0 + \eta^*_0)\} = \mathcal{C}^*_0$ and $\nabla^2 W^*_\gamma = \underbrace{\psi'(V^*_\gamma)}_{>0} \nabla^2 V^*_\gamma$ on $\mathcal{C}^*_0$.

Step 4 (*fading*). Applying Proposition 3.1 to W^*_γ yields, as $\nu^{\varepsilon_p} \Rightarrow \nu^0$,

$$\int_{\mathbb{R}^d} \nu^0(dx) \left[\int {}^tH(x,u)\nabla^2 W^*_\gamma(x) H(x,u)\mu(du)\right] \geq 0.$$

Since $\nabla^2 W^*_\gamma = 0$ outside $\{V \in (v^*_0 - \eta^*_0, v^*_0 + \eta^*_0)\}$, one gets

$$\begin{aligned}
&\sum_{\substack{x^* \in \mathcal{C}^*_0 \\ x^* \text{ local max}}} \nu^0(x^*) \underbrace{\text{“nonpositive term”}}_{\text{as } \nabla^2 W^*_\gamma \leq 0} \\
&+ \sum_{\substack{x^* \in \mathcal{C}^*_0 \\ x^* \text{ local min}}} \nu^0(x^*) \int {}^tH(x^*,u) \underbrace{\nabla W^*_\gamma(x^*)}_{=0 \text{ by step } 1} H(x^*,u)\mu(du) \\
&+ \sum_{\substack{x^* \in \mathcal{C}^*_0 \\ x^* \text{ saddle point}}} \nu^0(x^*)\psi'(V(x^*)) \left(\gamma \int {}^tH(x^*,u) A^+_{x^*} H(x^*,u)\mu(du)\right. \\
&\qquad\qquad\left. - \int {}^tH(x^*,u) A^-_{x^*} H(x^*,u)\mu(du)\right) \geq 0.
\end{aligned}$$

Letting γ go to 0 eventually leads to

$$\sum_{\substack{x^* \in \mathcal{C}^*_0 \\ x^* \text{ saddle point}}} \nu^0(x^*) \underbrace{\psi'(V(x^*))}_{>0} \underbrace{\int {}^tH(x^*,u) A^-_{x^*} H(x^*,u)\mu(du)}_{\geq 0 \text{ and } > 0 \text{ if } x^* = x^*_0} \leq 0.$$

This completes the proof of item (a). Item (b) is obvious. □

3.3. Spurious zeros of a function *h*. One may consider the situation where the mean function h of a stochastic pseudogradient algorithm related to a global Lyapunov function V has some spurious equilibrium points x^* in the following sense:

$$(\nabla V | h)(x^*) = 0 \text{ with } \nabla V(x^*) \neq 0.$$

We will show that, if this spurious equilibrium is excited in the direction of $\nabla V(x^*)$, that is,

$$(\nabla V|h)(x^*) = 0 \text{ and } \int \mu(d\omega)|(\nabla V(x^*)|H(x,\omega))|^2 > 0,$$

then x^* is not in the support of ν^0. Let $\mathcal{P}^*$ be the set of spurious excited equilibrium.

PROPOSITION 3.8. *Assume that V is a C^2 function and that H satisfies the assumptions of Proposition* 3.1. *If x^* is an excited spurious equilibrium point at level v^* and if there is $\eta^* > 0$ such that $\{(\nabla V|h) = 0\}\cap\{V \in [v^*-\eta^*, v^*+\eta^*]\} \subset \{V = v^*\}$, then, for any limiting distribution ν^0 of $(\nu^\varepsilon)_{\varepsilon\in(0,\varepsilon_0]}$, $\nu^0(\{x^*\}) = 0$.*

Proof. Let L be a positive real number and let x^* be an excited spurious equilibrium of h. Let $\varphi_{_L}$ be a nondecreasing function satisfying $\varphi_{_L}(v) := v - L(v-v^*)^2$ if v lies in a small enough neighborhood $[v^*-\alpha^*(L), v^*+\alpha^*(L)]$ of v^*, $\varphi_{_L}(v^*\pm\alpha^*(L)) = v^*\pm\alpha^*(L)$, $\varphi_{_L}$ is constant on both sides of ${}^c[v^*-\eta^*, v^*+\eta^*]$, and $\varphi'_{_L}$ is nonnegative.

Set $W_{_L} := \varphi_{_L}(V)$. The inner product $(\nabla W_{_L}|\nabla V)$ is nonnegative and, for every $x \in \{V = v^*\}$, $\nabla^2 W_{_L}(x) := -2L\nabla V\, {}^t\nabla V(x) + \nabla^2 V(x)$. Subsequently Proposition 3.1 yields

$$-2L\int_{\mathcal{P}^*\cap\{V=v^*\}} \nu^0(dx^*)\int \mu(d\omega)|(\nabla V(x)|H(x,\omega))|^2 + C^* \geq 0,$$

where C^* is a fixed real number. Consequently, as L goes to infinity,

$$\int_{\mathcal{P}^*\cap\{V=v^*\}} \nu^0(dx)\int \mu(d\omega)|(\nabla V(x)|H(x,\omega))|^2$$

needs to be 0. A continuity argument yields that $\nu^0(x^*) = 0$. □

Application to inflection points in one dimension. This a typical example of an excited spurious equilibrium point. As a matter of fact, assume that the mean function h satisfies

$$h(x^*) = 0 \text{ and } h \text{ has constant sign on } [x^*-\alpha^*, x^*+\alpha^*]$$

for some $\alpha^* > 0$. Then, x^* is a spurious critical point for any C^2 nondecreasing function V satisfying

$$V(x) := x - x^* \text{ if } x \in [x^*-\alpha^*/2, x^*+\alpha^*/2], \quad V'(x) = 0 \text{ if } x \notin [x^*-\alpha^*, x^*+\alpha^*].$$

All the results of this section rely on a second-order Taylor expansion of the probability transition P^ε as $\varepsilon \downarrow 0$ (see the proof of Proposition 3.1). That is the main reason why they require a rather low integrability assumption on the noise $(2+\varepsilon)$. In [3], the noise $H(x,\omega) - h(x)$ is bounded in the original urn model that motivated the work. So, the assumption that the Legendre transform of the noise has a uniform quadratic bound is quite natural in that setting. An equivalence between the chain-transitive relationship and the large deviation zero-cost relationship is established under a dimensionality condition (chaining number of unstable attractors less than 2). Then, some large deviation estimates à la Freidlin–Wentzell of the invariant distribution yield some precise results on the possible supports for the limiting invariant distributions. Thus, the case of an isolated unstable (saddle) cycle can be excluded from the possible supports using such methods in a more general context than that, e.g., proposed in the *local result* item of Theorem 3.7.

4. Conclusion. We have seen that the asymptotic behavior of a constant step stochastic algorithm, when the step goes to 0, is almost the same as that of the corresponding decreasing step stochastic algorithm. They share many properties: concentration near the equilibrium points of ODE_h, convergence to the invariant measure in the case of an attracting limit cycle of ODE_h, avoiding of repulsive equilibrium points or cycles in general and of the saddle points in the case of stochastic pseudogradient with isolated singular points. Only one difference seems to occur: a saddle point can be asymptotically weighted when ODE_h is not a stochastic pseudogradient with isolated singular points (see the Lemniscate example) or, more generally, with some unstable equilibrium points (see the periodic/quasi-cycle example when h does have zeros on the cycle). This happens when the algorithm is trapped into a quasi-cycle of singular points. In fact, this dissimilarity is an illusion: the asymptotics of the invariant measures ν^ε essentially take into account the time spent by the algorithm near some points and there is no contradiction between spending *almost* all the time at some place and not converging to it. This is exactly what happens, e.g., in the Lemniscate example. By the way, this illuminates the result obtained by several authors ([7], [21], [28]) that states that a stochastic algorithm $(X^t)_{t\in\mathbb{N}}$ with decreasing step cannot converge toward an excited unstable equilibrium point x^*. Actually several behaviors are hidden under this nonconvergence: when x^* is a repeller or a saddle point of a global Lyapunov function, it means that $\underline{\lim}_t|X^t - x^*| > \varepsilon_0 > 0$. When x^* is a saddle point trapped in a quasi-cycle as in the Lemniscate example, it just means that $\overline{\lim}_t|X^t - x^*| > \varepsilon_0 > 0$. It is the same for the unstable equilibrium points in the quasi-cycle example (setting (b) of Proposition 3.5). In this latter example, if the step is decreasing at the right rate (fast enough indeed), it has been shown by using some shadowing techniques (see [2]) that the algorithm may still converge to such unstable equilibrium (which are not saddle points, which makes this result consistent with those mentioned above). Actually, these results obtained with constant step lead us to think that it would be more appropriate to describe the behavior of a stochastic algorithm with decreasing step in terms of time spent near the singular points, even when the step approaches 0 too slowly to get some a.s. convergence properties. This can be carried out by studying the weighted empirical measures of the algorithm itself (see [12]).

REFERENCES

[1] H. Amann, *Ordinary Differential Equations*, de Gruyter Stud. Math. 13, de Gruyter Berlin, New York, 1990.

[2] M. Benaïm, *A Dynamical system approach to stochastic approximations*, SIAM J. Control Optim., 34 (1996), pp. 437–472.

[3] M. Benaïm, *Recursive algorithms, urn process and chaining number of chain recurrent sets*, Ergodic Theory Dynam. Systems, 18 (1997), pp. 53–87.

[4] M. Benaïm, J.C. Fort, and G. Pagès, *Convergence of the one-dimensional Kohonen algorithms*, Adv. Appl. Probab., 30 (1998), pp. 850–869.

[5] C. Bouton and G. Pagès, *Convergence in distribution of the one-dimensional Kohonen algorithms when the stimuli are not uniform*, Adv. Appl. Probab., 26 (1994), pp. 80–103.

[6] C. Bouton and G. Pagès, *About the multidimensional competitive learning vector quantization algorithm with constant gain*, Ann. Appl. Probab., 7 (1997), pp. 679–710.

[7] O. Brandière, M. Duflo, *Les algorithmes stochastiques contournent-ils les pièges?*, Ann. Inst. H. Poincaré Probab. Statist., 32 (1996), pp. 395–427.

[8] M. Duflo, *Algorithmes Stochastiques*, Math. Appl. 23, Springer, Paris, 1996.

[9] S.N. Ethier and T.G. Kurtz, *Markov Processes, Characterization and Convergence*, John Wiley, New York, 1986.

[10] J.C. Fort and G. Pagès, *Asymptotics of the Invariant Distribution of a Constant Step*

Stochastic Algorithm, Technical report 324, Laboratoire de Probabilités, Université de Paris, Paris, 1996.

[11] J.C. Fort and G. Pagès, *Convergence of stochastic algorithms: From the Kushner and Clark theorem to the Lyapunov functional method*, Adv. Appl. Probab., 28 (1996), pp. 1072–1094.

[12] J.C. Fort and G. Pagès, *Stochastic algorithms with non constant step: a.s. weak convergence of weighted empirical measures*, SIAM J. Control Optim., submitted.

[13] M.I. Freidlin and A.D. Wentzell, *Random Perturbations of Dynamic Systems*, Springer-Verlag, Berlin, 1984.

[14] P. Hartman, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[15] R.Z. Ha'sminskii, *The average principle for parabolic and elliptic differential equations and Markov processes with small diffusions*, Theor. Probab. Appl., 8 (1963), pp. 1–21.

[16] M. Hirsch, *Stability and convergence in strongly monotone dynamical systems*, J. Reine Angew. Math, 383 (1988), pp. 1–53.

[17] J. Jacod and A.N. Shiryaev, *Limit Theorems for Stochastic Processes*, Springer-Verlag, Berlin, 1990.

[18] Y. Kifer, *Random Perturbations of Dynamical Systems*, Birkhauser, Boston, 1988.

[19] U. Krengel, *Ergodic Theorems*, de Gruyter Stud. Math. 6, Walter de Gruyter & Co., Berlin-New York, 1985.

[20] H.J. Kushner and H. Huang, *Asymptotic properties of stochastic approximations with constant coefficients*, SIAM J. Control Optim., 19 (1981), pp. 87–105.

[21] V.A. Lazarev, *Convergence of stochastic approximation procedures in the case of several roots of a regression equation*, Problemy Pederachi Informatsii, 28 (1992), pp. 75–88.

[22] S.P. Meyn and R.L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.

[23] M.B. Nevel'son, *On the behavior of the invariant measure of a diffusion process with small diffusion on a circle*, Theory Probab. Appl., 9 (1964), pp. 125–131.

[24] M. Norman, *Limit theorems for stationary distributions*, Adv. Appl. Probab., 7 (1975), pp. 561–575.

[25] E. Nummelin, *General Irreducible Markov Chains and Non-Negative Operators*, Cambridge University Press, New York, 1984.

[26] E. Oja, *Simplified neuron models as a principal component analyser*, J. Math. Biol., 15 (1982), pp. 267–273.

[27] G. Pagès, *A method for space quantization numerical integration*, J. Appl. Comput. Math., 89 (1997), pp. 1–38.

[28] R. Pemantle, *Nonconvergence to unstable points in urn models and stochastic approximations*, Ann. Probab., 18 (1990), pp. 698–712.

[29] G.Ch. Pflug, *Stochastic minimization with constant step size: Asymptotic laws*, SIAM J. Control Optim., 24 (1986), pp. 655–666.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 5, pp. 1483–1502

STABILITY OF PIECEWISE-DETERMINISTIC MARKOV PROCESSES*

FRANÇOIS DUFOUR† AND OSWALDO L. V. COSTA‡

Abstract. In this paper, we study a form of stability for a general family of nondiffusion Markov processes known in the literature as piecewise-deterministic Markov process (PDMP). By stability here we mean the existence of an invariant probability measure for the PDMP. It is shown that the existence of such an invariant probability measure is equivalent to the existence of a σ-finite invariant measure for a Markov kernel G linked to the resolvent operator U of the PDMP, satisfying a boundedness condition or, equivalently, a Radon–Nikodým derivative. Here we generalize existing results of the literature [O. Costa, *J. Appl. Prob.*, 27, (1990), pp. 60–73; M. Davis, *Markov Models and Optimization*, Chapman and Hall, 1993] since we do not require any additional assumptions to establish this equivalence. Moreover, we give sufficient conditions to ensure the existence of such a σ-finite measure satisfying the boundedness condition. They are mainly based on a modified Foster–Lyapunov criterion for the case in which the Markov chain generated by G is either recurrent or weak Feller. To emphasize the relevance of our results, we study three examples and in particular, we are able to generalize the results obtained by Costa and Davis on the capacity expansion model.

Key words. piecewise-deterministic Markov process, invariant measure, Markov chain

AMS subject classifications. 60J25, 60J10

PII. S0363012997330890

1. Introduction.

Piecewise-deterministic Markov processes (PDMPs) were introduced by Davis [3] as a general family of nondiffusion stochastic models suitable for formulating many optimization problems in several areas of applications. The motion of a PDMP $\{x_t\}$ in a state space E depends on three parameters, namely, the flow Φ, the jump rate λ, and the transition probability measure Q. Starting from a point x in E, the motion of the process follows the flow $\Phi(x,t)$ until the first jump time T_1. This jump time occurs either when the flow hits the boundary of the state space E or in a Poisson-like fashion with rate $\lambda(\Phi(x,t))$. The location Z_1 of the process at the jump time T_1 is selected by the measure $Q(\Phi(x,T_1),.)$ and the motion restarts from this new point Z_1 as before. This gives a piecewise-deterministic trajectory for the PDMP $\{x_t\}$ with jump times $\{T_1,T_2,\ldots\}$ and postjump location $\{Z_1,Z_2,\ldots\}$. By a suitable choice of the state space E and the parameters Φ, λ, and Q it is possible to virtually model all nondiffusion processes found in the international literature. The reader can find in [4] a complete and updated description of the theory of PDMPs as well as several important applications.

A problem of great importance in the theory of stochastic processes that has been

*Received by the editors December 1, 1997; accepted for publication (in revised form) September 29, 1998; published electronically July 27, 1999. This research was completed during the USP/COFECUB Franco-Brazilian cooperation mission UC 40/97.

http://www.siam.org/journals/sicon/37-5/33089.html

†Laboratoire des Signaux et Systèmes, C.N.R.S.–E.S.E., Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France (dufour@lss.supelec.fr).

‡Departamento de Engenharia Eletrônica, Escola Politécnica da Universidade de São Paulo, CEP: 05508 900-São Paulo, Brazil (oswaldo@lac.usp.br). The research of this author was supported by FAPESP (Research Council of the State of São Paulo) and CNPq (Brazilian National Research Council).

in evidence for over the last decades is to obtain conditions for the following problem.

$\mathbf{P}_1$. Existence of an invariant probability measure for a stochastic process.

The existence of such a measure corresponds to an important form of stochastic stability and it is the first step toward an ergodic analysis of the process. For a small sample of the huge theory available today on this subject, see [9, 8, 11, 12, 15, 13]. In general it is not a simple matter to determine whether a given process in a general state space has an invariant probability measure and, if so, whether it is unique. For discrete-time stochastic processes there is available in [14] a rather complete treatment of this subject.

The main goal of [2] (see also [4]) was to derive conditions for the existence of an invariant probability measure for a PDMP $\{x_t\}$. Associated with the PDMP $\{x_t\}$ there is a Markov chain given by the postjump location $\{Z_1, Z_2, \ldots\}$ with kernel $\overline{G}$ (see (8) for its definition). It is natural to ask then if it is possible to link the invariant probability measures associated with the Markov chain $\{Z_1, Z_2, \ldots\}$ and the invariant probability measures associated with the PDMP $\{x_t\}$. Since the conditions for existence and uniqueness of an invariant probability measure are more easily stated in the discrete-time setting, this connection provides a way of analyzing $\mathbf{P}_1$ for a PDMP $\{x_t\}$ without relying on the general continuous-time theory. It was shown in [2] that, under some technical assumptions, there is a one-to-one mapping between the invariant probability measures for the PDMP $\{x_t\}$ and the invariant probability measures of the Markov chain $\{Z_1, Z_2, \ldots\}$. However, this result is not entirely satisfactory since the technical assumptions in [2] require that, roughly speaking, the jump rate λ must be strictly positive and bounded. Important applications presented in [4] show that if these conditions are violated then there may exist an invariant probability measure for the PDMP $\{x_t\}$ but not for the associated Markov chain $\{Z_1, Z_2, \ldots\}$, and vice versa.

In the first part of this paper the problem $\mathbf{P}_1$ for a PDMP $\{x_t\}$ is revisited under a new approach that removes the assumptions made in [2]. Instead of using the kernel $\overline{G}$ associated with the Markov chain $\{Z_1, Z_2, \ldots\}$, we introduce a new kernel G (for its definition, see (4)), related to the resolvent kernel U of the PDMP $\{x_t\}$, and show that $\mathbf{P}_1$ for a PDMP $\{x_t\}$ is equivalent to the next problem.

$\mathbf{P}_2$. Existence of an invariant σ-finite measure for the Markov chain associated with the kernel G satisfying a boundedness condition.

It is important to stress that this equivalence requires no extra conditions than those usually imposed on the definition of the PDMP $\{x_t\}$, being therefore a significant improvement with respect to those results presented in [2]. Notice that $\mathbf{P}_2$ is related to an invariant σ-finite measure and therefore it is not necessarily finite. As will be seen in section 3, the boundedness condition in $\mathbf{P}_2$ can also be characterized as a condition in terms of the Radon–Nikodým derivative of the invariant σ-finite measure for G and a certain probability measure. As in [2], a one-to-one mapping between the invariant probability measures for the PDMP $\{x_t\}$ and the invariant σ-finite measures for the Markov chain associated with the kernel G is established. The examples in [4] are revisited under this new framework and, of course, the problems found in [4] are eliminated with our approach.

It is well known that $\mathbf{P}_1$ is equivalent to the problem of the existence of an invariant probability measure for the resolvent kernel associated to the transition function of the stochastic process (see [1]). It is important to point out that such a direct approach to our problem is in vain due to the fact that both the transition function of a general PDMP $\{x_t\}$ as well as its associated resolvent kernel cannot be explicitly calculated. Here our strategy is to use the special structure of the PDMP to provide a tractable answer to the problem $\mathbf{P}_1$.

In the second part of this paper we give weak sufficient conditions for ensuring that $\mathbf{P}_2$ admits a solution. These conditions are based on a modified Foster–Lyapunov criterion for the case in which the kernel G is either recurrent or weak Feller and, we believe, significantly improve previous results presented in [2, 4] for $\mathbf{P}_1$ for a PDMP $\{x_t\}$. Indeed, an interesting property of the PDMPs is that under weak conditions (see Proposition 4.9) they generate weak Feller kernels G and U although the PDMPs themselves are generally not weak Feller Markov processes. Our results are related to those in Meyn and Tweedie [15, 13], Lin [10], and Foguel [7]. Moreover, it is shown that an invariant probability measure always exists for the PDMP $\{x_t\}$ when G is weak Feller and the union of the state space of $\{x_t\}$ and the boundary Γ^+ (for its definition, see section 2) is a compact set. To illustrate the usefulness of our new conditions, the capacity expansion problem (see [4]) is revisited in a more general setup. A sufficient condition for the existence of an invariant measure is obtained, generalizing previous results in the literature.

The paper is organized in the following way. In section 2 we describe the model and introduce the notation. The equivalence between $\mathbf{P}_1$ for PDMP and $\mathbf{P}_2$ is established in section 3. In section 4 we derive sufficient conditions for checking $\mathbf{P}_2$. Illustrative examples are given in section 5. We conclude the paper in section 6 with some final remarks.

2. Model of piecewise deterministic Markov processes. Let E^0 be an open subset of $\mathbb{R}^n$ and ∂E^0 its boundary. As originally defined by Davis [3, 4], a PDMP is determined by its local characteristics $(\mathfrak{X}, \lambda, Q)$, where $\mathfrak{X}, \lambda, \text{and}\, Q$ are defined as follows.

(a) $\mathfrak{X}$ is a locally Lipschitz continuous vector field in E^0 with flow $\Phi(x,t)$.

Now define

$$\Gamma^+ \doteq \{x \in \partial E^0 : x = \Phi(y,t), y \in E^0, t > 0\}$$

and

$$\Gamma^- \doteq \{x \in \partial E^0 : x = \Phi(y,-t), y \in E^0, t > 0\}.$$

$\Gamma^+ \subset \partial E^0$ represents the boundary points at which the flow exits from E^0. $\Gamma^- \subset \partial E^0$ is characterized by the fact that the flow starting from a point in Γ^- will not leave E^0 immediately. Therefore it is natural to define the state space for the PDMP by

$$E \doteq E^0 \cup \Gamma^- - \Gamma^- \cap \Gamma^+.$$

$\mathcal{E}$ will denote the Borel σ-field of E.

Let us denote by $t_*(x) \doteq \inf\{t > 0 : \Phi(x,t) \in \partial E^0\}$ $\forall$ x in E. If this set is empty, then $t_*(x) = \infty$.

Assume that the explosion time of the flow Φ is equal to infinity when $t_*(x) = \infty$.

(b) $\lambda : E \to \mathbb{R}^+$ is the jump rate. It is assumed that this function is measurable and $(\forall x \in E)$ $(\exists \varepsilon > 0)$ such that $\int_0^\varepsilon \lambda(\Phi(x,s))ds$ exists.

(c) $Q : E \cup \Gamma^+ \times \mathcal{E} \to [0,1]$ is the transition measure satisfying the following property: $(\forall x \in E \cup \Gamma^+)$ $Q(x, E - \{x\}) = 1$.

It is shown in [3, 4] that there exists a probability space $(\Omega, \mathcal{F}_t, \mathcal{F}, P_x)$ on which the motion of the PDMP $\{x_t\}$ is defined as follows. Starting from x, the first jump time T_1 of the process is given by

$$P_x(T_1 > t) \doteq I_{\{t<t_*(x)\}} \exp\left\{-\int_0^t \lambda(\Phi(x,s))ds\right\}.$$

Then

$$(\forall t \in [0, T_1)) \qquad x_t = \Phi(x,t) \qquad \text{and} \qquad x_{T_1} = Z_1,$$

where Z_1 is the random variable having as distribution $Q(\Phi(x,T_1),.)$. Now the process follows the path $\Phi(x_{T_1}, t - T_1)$ up to another jump time T_2, where it jumps to the state $x_{T_2} = Z_2$. The interjump time $T_2 - T_1$ and the postjump random variable Z_2 are defined similarly.

It will be assumed that $(\forall (t,x) \in \mathbb{R}_+ \times E)$, $E_x[\sum_k I_{\{t \geq T_k\}}] < \infty$.

For a more general definition of PDMP the reader is referred to Davis [4, section 29]. However, our results are still valid in this setting.

Notation. Let us denote by μ_{Leb} the Lebesgue measure on $\mathbb{R}$ and $\overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$.

Let $B(E)$ be the space of real-valued bounded Borel-measurable functions on E.

$C_b(E)$ ($C_c(E)$, respectively) denotes the space of bounded real-valued continuous functions on E (with compact support, respectively).

$\forall A \in \mathcal{E}$, $A^c \doteq E - A$ and $\overline{A}$ is the closure of A.

Let ρ be a metric on E; then $\forall\ x \in E$ and $A \in \mathcal{E}$, $d(x, A) \doteq \inf y \in A \rho(x,y)$.

Let X be a metric space and $\mathcal{B}(X)$ its Borel σ-field.

$\mathcal{M}(X)$ ($\mathcal{M}_b(X)$, respectively) denotes the set of all σ-finite (finite, respectively) measures defined on $(X, \mathcal{B}(X))$.

Finally, we recall the following classical notations: let K_1 and K_2 be two kernels mapping $(E, \mathcal{E})$ into $\mathbb{R}_+$. Then

$$(\forall \mu \in \mathcal{M}(E)),\ (\forall A \in \mathcal{E}), \quad \mu K_1(A) \doteq \int_E K_1(x,A)\mu(dx),$$

$$(\forall f \in B(E)),\ (\forall x \in E), \quad K_1 f(x) \doteq \int_E f(y) K_1(x,dy),$$

$$(\forall x \in E),\ (\forall A \in \mathcal{E}), \quad K_1K_2(x,A) \doteq \int_E K_2(y,A)K_1(x,dy),$$

provided that the right members exist (see [17, Chapter 1, Definitions 1.2 and 1.5]).

3. Invariant probability measure. The main goal of this section is to show that $\mathbf{P}_1$ for a PDMP is equivalent to $\mathbf{P}_2$ (see Theorem 3.5). This new result represents a significant improvement with respect to those of Costa [2] and Davis [4]. Indeed, unlike [2, 4] this equivalence is established without any additional conditions. The key idea is to use a connection between the resolvent kernel U (see (6)) and the kernel G (see (4)) which comes from Lemma 3.2.

Moreover, in order to provide a complete comparison with the results obtained by Costa and Davis (see, for example, Proposition (34.36) in [4]), we derive transformations which link the invariant probability measures for $\{x_t\}$ and the stochastic kernel G (see Corollary 3.6).

For notational convenience, let us define for $x \in E$ and $t \in [0, t_*(x)[$

$$\Lambda(x,t) \doteq \int_0^t \lambda(\Phi(x,s))ds \tag{1}$$

and the following kernels mapping $(E, \mathcal{E})$ into $[0,1]$:

$$L(x,A) \doteq \int_0^{t_*(x)} \exp -\{s + \Lambda(x,s)\} I_A(\Phi(x,s))ds, \tag{2}$$

$$\begin{aligned} K(x,A) \doteq& \int_0^{t_*(x)} \lambda(\Phi(x,s)) \exp -\{s + \Lambda(x,s)\} Q(\Phi(x,s), A)ds \\ &+ \exp -\{t_*(x) + \Lambda(x, t_*(x))\} Q(\Phi(x, t_*(x)), A), \end{aligned} \tag{3}$$

$$G(x,A) \doteq K(x,A) + L(x,A), \tag{4}$$

$$L_n(x,A) \doteq \sum_{i=0}^{n-1} K^i L(x,A), \tag{5}$$

$$U(x,A) \doteq E_x \int_0^{\infty} I_A(x_t) \exp\{-t\} dt, \tag{6}$$

$$S(x,A) \doteq L(x,E) I_A(x), \tag{7}$$

$$\begin{aligned} \overline{G}(x,A) \doteq& \int_0^{t_*(x)} \lambda(\Phi(x,s)) \exp\{-\Lambda(x,s)\} Q(\Phi(x,s), A)ds \\ &+ \exp\{-\Lambda(x, t_*(x))\} Q(\Phi(x, t_*(x)), A). \end{aligned} \tag{8}$$

In the terminology used by Nummelin [16], the kernels G, U, $\overline{G}$ are stochastic kernels because $\forall\ x \in E$, $G(x,E) = U(x,E) = \overline{G}(x,E) = 1$. $\overline{G}$ is the stochastic kernel for the Markov chain $\{Z_1, Z_2, \ldots\}$. U is the resolvent kernel associated with $\{x_t\}$.

REMARK 3.1. *Note that* $\forall\ x \in E$, $0 < L(x,E) < 1$ *and that* G *is a stochastic kernel. Indeed,* $\forall\ x \in E$,

$$L(x,E) + K(x,E) = 1. \tag{9}$$

Therefore, we can define the kernel $R : E \times \mathcal{E} \longrightarrow \mathbb{R}_+$ such that

$$R(x,A) \doteq \frac{I_A(x)}{L(x,E)} \cdot \tag{10}$$

The following result establishes a connection between the kernels K, L and the resolvent U.

LEMMA 3.2.

$$\forall x \in E, \quad \forall A \in \mathcal{E}, \qquad KU(x,A) + L(x,A) = U(x,A). \tag{11}$$

Proof. From section 32.2 and Proposition 32.34 in Davis [4] and the definition of L_n (see (5)), it is easy to obtain

$$\forall x \in E, \quad \forall A \in \mathcal{E}, \qquad \lim_{n \to \infty} \uparrow L_n(x,A) = U(x,A) \tag{12}$$

and

$$\forall x \in E, \quad \forall A \in \mathcal{E}, \qquad KL_n(x,A) + L(x,A) = L_{n+1}(x,A).$$

Now using the monotone convergence theorem, the result follows. □

REMARK 3.3. *Let us introduce the kernel* $J : E \times \mathcal{E} \longrightarrow \mathbb{R}_+$ *such that*

$$J(x, A) \doteq \sum_{j=0}^{\infty} \int_A L(y, E) K^j(x, dy) \cdot \tag{13}$$

Using (12) *and the definition of* J *we have that* $\forall\ x \in E$, $J(x, E) = U(x, E) = 1$. *Therefore* J *is a stochastic kernel.*

Before presenting the first main result of this section, let us derive the following technical lemma.

LEMMA 3.4. *Let* α *be a* σ*-finite measure defined on* $(E, \mathcal{E})$. *Then assertions* (i) *and* (ii) *are equivalent.*

(i) α *is such that* $\alpha = c\nu R$, *where* ν *is a probability measure and* c *a constant in* $(0, \infty)$.

(ii) α *is such that* $\alpha S(E) < \infty$.

Proof. (i) ⇒ (ii). The proof is obvious.

(ii) ⇒ (i). Define the probability measure $\nu \doteq \frac{\alpha S}{\alpha S(E)}$. Then clearly $\alpha = c\nu R$ with $c = \alpha S(E)$. □

The next theorem establishes the connection between $\mathbf{P}_1$ and $\mathbf{P}_2$.

THEOREM 3.5. *The following assertions are equivalent.*

(i) *There exists an invariant probability measure* μ *for* $\{x_t\}$.

(ii) *There exists a probability measure* ν *such that the positive* σ*-finite measure* $\pi = \nu R$ *is invariant for the stochastic kernel* G.

(iii) *There exists a positive* σ*-finite measure* π *invariant for the stochastic kernel* G *such that* $\pi S(E) < \infty$.

Proof. (ii) ⇒ (i). Assume that there exists a probability measure ν such that $\pi \doteq \nu R$ is invariant for G and define the measure $\mu \doteq \pi L$. Note that $\forall\ A \in \mathcal{E}$, $\mu(A) = \int_E \frac{L(y,A)}{L(y,E)} \nu(dy)$. Therefore μ is a probability measure. Since π is a positive σ-finite measure, there exists a partition $\{E_i\}$ of E such that $\pi(E_i) < \infty$. Recalling that $L = G - K$ (see (4)) and $\pi = \pi G$ we have that $\forall\ A \in \mathcal{E}$,

$$\begin{aligned} \mu U(A) &= \sum_i \int_{E_i} U(x, A) \pi L(dx) \\ &= \sum_i \int_{E_i} U(x, A) \pi (I - K)(dx). \end{aligned}$$

Now, using Lemma 3.2, it follows that

$$\begin{aligned} \mu U(A) &= \sum_i \int_{E_i} \{U(x, A) - KU(x, A)\} \pi(dx) \\ &= \sum_i \int_{E_i} L(x, A) \pi(dx) \\ &= \mu(A). \end{aligned}$$

Therefore from Lemma 1 in [1], μ is an invariant measure for $\{x_t\}$.

(i) ⇒ (ii). Let μ be an invariant probability measure for the process $\{x_t\}$. Then $\mu U = \mu$ (see [1]). Let us define the measure ν by $\nu \doteq \mu J$ with J defined in (13). The measure so defined is a probability since $\mu J(E) = \mu U(E) = \mu(E) = 1$ (see Remark 3.3). Let us introduce $\pi \doteq \nu R$ (see (10)). Then, from (10) and (13) we obtain

$$\pi = \sum_{j=0}^{\infty} \mu K^j. \tag{14}$$

From (14) and (12) and Lemma 3.2, it is easy to obtain that $\pi L = \mu$ and $\pi K + \mu = \pi$.

Consequently $\pi G = \pi(K + L) = \pi K + \mu = \pi$.

(ii) $\Leftrightarrow$ (iii). Using Lemma 3.4 the equivalence is straightforward.

This completes the proof. □

As mentioned in the introduction, the condition on the σ-finite invariant measure π for the Markov chain associated with G can be written in terms of the boundedness of $\pi S(E)$ ((iii) in the theorem) or in terms of the Radon–Nikodým derivative $L(x, E)^{-1}$ of π with respect to a probability measure ν ((ii) in the theorem).

The next result gives explicitly the links between ν and μ using an approach that differs from Costa [2] and Davis [4]. In particular, we avoid introducing the so-called boundary measure associated with the invariant probability measure of $\{x_t\}$ (see, for example, [4, pp. 116–118]). Our approach directly uses the kernels J, L, and R. Moreover, it is shown that these mappings are one to one, providing a complete comparison with results known in the literature (see, for example, Proposition (34.36) in Davis, [4]).

COROLLARY 3.6. (i). *If μ is an invariant probability measure for $\{x_t\}$, then the σ-finite measure μJR is invariant for G and $\mu JRL = \mu$.*

(ii). *If ν is a probability measure such that νR is invariant for G, then the probability measure νRL is invariant for $\{x_t\}$ and $\nu RLJ = \nu$.*

Proof. (i) and the first part of (ii) are easy consequences of Theorem 3.5.

In order to show that if ν is a probability measure such that νR is invariant for G, then $\nu RLJ = \nu$, let us first note that

$$\forall x \in E, \quad \forall A \in \mathcal{E}, \qquad J(x, A) - KJ(x, A) = I_A(x)L(x, E). \tag{15}$$

Since νR is a positive σ-finite measure, there exists a partition $\{E_i\}$ of E such that $\nu R(E_i) < \infty$. Moreover, using the fact that $\nu RK + \nu RL = \nu R$ and (15), we have

$$\begin{aligned} \nu RLJ(A) &= \sum_i \int_{E_i} J(x, A)\nu RL(dx) \\ &= \sum_i \int_{E_i} I_A(x)L(x, E)\nu(dx) \\ &= \nu(A). \end{aligned}$$

This completes the proof. □

The following result shows that if λ is bounded, then $\mathbf{P}_1$ is equivalent to the existence of a finite invariant measure for G. In what follows, we set (see [4, sections 24 and 26])

$$\begin{aligned} N_t &\doteq \sum_{i=1}^{\infty} I_{\{t \geq T_i\}}, \\ p^*(t) &\doteq \sum_{i=1}^{\infty} I_{\{t \geq T_i\}} I_{\{x_{T_i-} \in \Gamma+\}}. \end{aligned}$$

Note that $\{N_t\}$ and $\{p^*(t)\}$ are counting processes, N_t being the number of jumps of the process $\{x_t\}$ in the time interval $[0,t]$ and $p^*(t)$ counting the number of jumps from the boundary up to time t.

PROPOSITION 3.7. *Let μ be an invariant probability measure for the process $\{x_t\}$ with $E_\mu[N_t] < \infty \ \forall \ t \in \mathbb{R}$, and define $\pi \doteq \mu JR$. Then the following assertions are equivalent:*

(i) $\pi(E) < \infty$.

(ii) $\int_E \lambda(x)\mu(dx) < \infty$.

Proof. By repeating the same arguments as in Proposition 34.13 and Theorem 34.15 in Davis [4, pp. 116–117] and using the fact that $E_\mu[N_t] < \infty \ \forall \ t \in \mathbb{R}$, we obtain that

$$E_\mu\left[\int_0^t \exp\{-s\}dp^*(s)\right] = (1-\exp\{-t\})\sigma(\Gamma^+),$$

where σ is a finite measure defined on $(\Gamma^+, \mathcal{B}(\Gamma^+))$ by Theorem 34.15 in Davis (see [4, p. 117]).

From the monotone convergence theorem, it follows that

$$E_\mu\left[\int_0^\infty \exp\{-s\}dp^*(s)\right] = \sigma(\Gamma^+). \tag{16}$$

Consequently, we can define:

$$\Psi_n \doteq E_\mu\left[\int_0^\infty \exp\{-s\}\lambda_n(x_s)ds + \int_0^\infty \exp\{-s\}dp^*(s)\right], \tag{17}$$

where

$$\lambda_n(x) \doteq \begin{cases} \lambda(x) & \text{if } \lambda(x) < n, \\ n & \text{otherwise.} \end{cases}$$

Using Fubini's theorem and the fact that μ is an invariant probability measure for $\{x_t\}$, we get that

$$E_\mu\left[\int_0^t \exp\{-s\}\lambda_n(x_s)ds\right] = (1-\exp\{-t\})\int_E \lambda_n(x)\mu(dx)$$

and again from the monotone convergence theorem,

$$E_\mu\left[\int_0^\infty \exp\{-s\}\lambda_n(x_s)ds\right] = \int_E \lambda_n(x)\mu(dx). \tag{18}$$

Therefore, combining (16), (17), and (18), we have

$$\Psi_n = \int_E \lambda_n(x)\mu(dx) + \sigma(\Gamma^+).$$

Now using the monotone convergence theorem, we obtain

$$\lim_{n\to\infty} \Psi_n = \int_E \lambda(x)\mu(dx) + \sigma(\Gamma^+). \tag{19}$$

On the other hand, it is easy to show from section 32.2 and Proposition 32.34 of Davis [4] that Ψ_n can be written as

$$\Psi_n = \mu \sum_{i=0}^{\infty} K^i \phi_n, \tag{20}$$

where

$$\phi_n = \int_0^{t_*(x)} \lambda_n(\Phi(x,s)) \exp -\{s + \Lambda(x,s)\} ds + \exp -\{t_*(x) + \Lambda(x, t_*(x))\}.$$

Recalling the definition of the stochastic kernel K (see (3)) and from the monotone convergence theorem, it follows that

$$\lim_{n\to\infty} \uparrow \phi_n = K(x, E).$$

Therefore, using (20), we have

$$\lim_{n\to\infty} \Psi_n = \mu \sum_{i=1}^{\infty} K^i(E).$$

Now, with (19) and the previous one, it follows that

$$\mu \sum_{i=0}^{\infty} K^i(E) = 1 + \mu \sum_{i=1}^{\infty} K^i(E) = 1 + \int_E \lambda(x)\mu(dx) + \sigma(\Gamma^+). \tag{21}$$

However, recalling the definitions of the stochastic kernels J and R (see (13) and (10)), we have

$$\mu JR(E) = \mu \sum_{i=0}^{\infty} K^i(E).$$

Thus, combining the previous equation with (21), we obtain the result. □

REMARK 3.8. *It is clear that if $\lambda(.)$ is bounded, then $\pi(E) < \infty$. Therefore, if $\lambda(.)$ is bounded, there exists an invariant probability measure for $\{x_t\}$ if and only if there exists an invariant probability measure for the stochastic kernel G. Assuming, moreover, that G is weak Feller, necessary and sufficient conditions can be given to ensure that $\{x_t\}$ has an invariant measure (see, for example,* [10, 7, 9, 8]). *However, these conditions are difficult to check in practice since they need the limit of the sum of iterations of G.*

4. Sufficient conditions. The aim of this section is to give simple sufficient conditions to check if $\mathbf{P}_2$ (and therefore $\mathbf{P}_1$) has a solution. Such test criteria are based on Foster–Lyapunov inequalities. It has been proven in Meyn and Tweedie [14] that this test function implies the existence of an invariant probability measure for a class of Markov chain. In our case the problem is different since we have to find a σ-finite measure π for G that satisfies the condition $\pi S(E) < \infty$. Following the line of Meyn and Tweedie [14], Lin [10], and Foguel [7], we are able to show that modified drift criteria ensure such a property for the case in which the Markov chain generated by G is either recurrent or weak Feller. Moreover, it is shown that an invariant probability measure always exists for the PDMP $\{x_t\}$ when G is weak Feller and the union of the state space of $\{x_t\}$ and the boundary Γ^+ (for its definition, see section 2) is a compact set.

First we recall some classical definitions related to Markov chains. For a complete exposition on the subject, the reader is referred to the book by Meyn and Tweedie [14].

Notation. We shall denote by $\{Y_n\}$ the Markov chain generated by the stochastic kernel G.

DEFINITION 4.1. *$\{Y_n\}$ is φ-irreducible if there exists a measure φ on $(E, \mathcal{E})$ such that, whenever $\varphi(A) > 0$, it follows that $P_x(\tau_A < \infty) > 0 \ \forall \ x \in E$, where $\tau_A = \inf\{n \geq 1 : Y_n \in A\}$.*

If $\{Y_n\}$ is φ-irreducible, then $\psi = \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^{n+1} \varphi G^n$ is a maximal irreducibility measure (see [14, Proposition 4.2.2, p. 88]). In the following, we will use ψ to denote a maximal irreducibility measure for $\{Y_n\}$. Associated with a maximal irreducibility measure for $\{Y_n\}$, we define

$$\mathcal{E}^+ \doteq \{A \in \mathcal{E} : \psi(A) > 0\}.$$

DEFINITION 4.2. *$\{Y_n\}$ is called recurrent if it is ψ-irreducible and*

$$\sum_{n=1}^{\infty} G^n(x, A) = \infty \ \forall \ x \in E \text{ and } A \in \mathcal{E}^+$$

.

DEFINITION 4.3. *A set $C \in \mathcal{E}$ is called ν_a-petite if there exists a nontrivial measure ν_a on $(E, \mathcal{E})$ such that*

$$(\forall x \in C), \ (\forall B \in \mathcal{E}) \quad \sum_{n=1}^{\infty} a(n) G^n(x, B) \geq \nu_a(B),$$

where the positive sequence $\{a(n)\}$, satisfies $\sum_{n=0}^{\infty} a(n) = 1$.

Following similar ideas as in Proposition 14.1.1 in the book by Meyn and Tweedie [14, p. 333], we can now establish a sufficient condition for the existence and uniqueness of an invariant probability measure for the PDMP $\{x_t\}$ based on the associated Markov chain $\{Y_n\}$. Note, however, that, unlike in Proposition 14.1.1 in [14], it cannot be claimed that the invariant measure for the Markov chain $\{Y_n\}$ is finite, since $L(Y_n, E) < 1$.

THEOREM 4.4. *Suppose that $\{Y_n\}$ is recurrent and for a petite set $C \in \mathcal{E}^+$ we have*

$$\sup_{x \in C} E_x \left[\sum_{n=1}^{\tau_C} L(Y_n, E) \right] < \infty, \tag{22}$$

where $\tau_C = \inf\{n \geq 1 : Y_n \in C\}$.

Then there exists a unique invariant probability measure for the PDMP $\{x_t\}$.

Proof. Since $\{Y_n\}$ is recurrent it admits a unique (up to constant multiples) σ-finite invariant measure π. From Proposition 10.1.2 in [14, p. 232], it follows that $\pi(C) < \infty$. Although here the function $L(., E) < 1$, the rest of the proof is the same as in Proposition 14.1.1 in [14, p. 232] and we obtain the result. □

The condition given by (22) can be replaced by the following one, which may be easier to check, and is inspired from condition $(V3)$ of Meyn and Tweedie [14, p. 337].

Condition $(\mathcal{D})$. For some set $C \in \mathcal{E}$, some constant $b < \infty$, and an extended real-valued function $V : E \to [0, \infty]$, we have

$$(\forall x \in E) \quad GV(x) - V(x) \leq -L(x, E) + bI_C(x).$$

This condition is of the same form as the drift or the Foster–Lyapunov criterion which is often used in the literature (see [13, 14] and references therein). However, we want to point out the fact that condition $(\mathcal{D})$ is weaker than conditions $(V2)$, $(V3)$ of the book by Meyn and Tweedie [14] due to the fact that $L(., E) < 1$. Therefore, direct application of their results are not straightforward. Note in particular that, unlike Theorem 11.3.4 in [14, p. 265], condition $(\mathcal{D})$ does not guarantee that $E_x[\tau_C] \leq V(x) + bI_C(x)$, since $L(Y_n, E) < 1$ (see (23) in the proof of Corollary 4.5). The next corollary gives the first way to ensure that there exists an invariant probability measure for $\{x_t\}$. It will be used in the next section to show that the pathologies found in the approach of Costa [2] and Davis (see section 34 in [4]) do not exist with our method.

COROLLARY 4.5. *Suppose that* $\{Y_n\}$ *is recurrent and that condition* $(\mathcal{D})$ *is satisfied for a function* V *bounded in a petite set* $C \in \mathcal{E}^+$. *Then there exists a unique invariant probability measure for the PDMP* $\{x_t\}$.

Proof. Applying Proposition 11.3.3 in [14, p. 264] with $Z_k = V(Y_k)$, $\varepsilon_k(x) = L(x, E)$, and $c = 1$, we obtain that

$$E_x\left[\sum_{n=1}^{\tau_C - 1} L(Y_n, E)\right] \leq \begin{cases} L(x, E) + GV(x), & x \in C, \\ V(x) \quad \text{otherwise.} \end{cases}$$

Recall that $\forall\ x \in C$, we have $GV(x) - V(x) \leq -L(x, E) + bI_C(x)$.

Therefore,

$$\begin{aligned} (\forall x \in E) \quad E_x\left[\sum_{n=1}^{\tau_C} L(Y_n, E)\right] &\leq E_x\left[\sum_{n=0}^{\tau_C - 1} L(Y_n, E)\right] + 1 \\ &\leq V(x) + 1 + bI_C(x). \end{aligned} \tag{23}$$

Then, $\sup_{x \in C} E_x[\sum_{n=1}^{\tau_C} L(Y_n, E)] < \infty$ and the result follows by using Theorem 4.4. ☐

In the case where the stochastic kernel G is weak Feller, the next result gives a sufficient condition to check if $\{x_t\}$ has an invariant measure. Before presenting the result, we need the following definition of norm-like functions (see [14, p. 214]).

DEFINITION 4.6. *A function* V *is a norm-like function if it is a positive real-valued function such that* $\lim_{n\to\infty} \inf_{x \notin K_n} V(x) = \infty$, *where* K_n *is an increasing sequence of compact sets satisfying* $\lim_{n\to\infty} \uparrow K_n = E$.

THEOREM 4.7. *Suppose that the Markov chain* $\{Y_n\}$ *is weak Feller,* $L(., E)$ *is a continuous function, and condition* $(\mathcal{D})$ *is satisfied with a compact set* C *and a positive norm-like function* V *which is finite at least at one* $x \in E$. *Then an invariant probability measure exists for the PDMP* $\{x_t\}$.

Proof. If there exists $x \in E$ such that $P_x(\tau_C < \infty) < 1$, then following the last part of the derivation of Theorem 12.3.3 in [14, p. 296] and using the fact that V is a norm-like function, there exists an invariant probability measure π for G. So, $\pi S(E) < \infty$, and using Theorem 3.5 the result follows.

Now if $P_x(\tau_C < \infty) = 1 \ \forall \ x \in E$, then using Proposition 9.1.1 in [14, p. 202], we have that

$$(\forall x \in E) \quad \sum_{n=0}^{\infty} G^n(x, C) = \infty.$$

Let us define the function $g(x) \in C_c(E)$ by

$$g(x) \doteq \frac{d(x, O_1^c)}{d(x, O_1^c) + d(x, \overline{O}_2)},$$

where O_1, O_2 are open sets with compact closure for which $C \subset O_2 \subset \overline{O}_2 \subset O_1$ and $d(.,.)$ is as defined in section 2. Then, we have

$$(\forall x \in E) \quad \sum_{n=0}^{\infty} G^n g(x) = \infty,$$

since $I_C(x) \leq g(x)$.

Therefore, applying Theorem 5.1 in [10], there exists an invariant σ-finite measure π for the stochastic kernel G defined by

$$(\forall f \in C_c(E)) \quad \int_E f(x)\pi(dx) = \underset{N}{\text{bLIM}} \frac{\sum_{n=0}^{N} \eta G^n f}{\sum_{n=0}^{N} \eta G^n g},$$

where η is a positive finite measure on $(E, \mathcal{E})$ and bLIM denotes the Banach limit [18].

Using the hypothesis, we have

$$L(x, E) + GV(x) \leq V(x) + bg(x).$$

Consequently,

$$\frac{\sum_{n=0}^{N} \eta G^n L(E)}{\sum_{n=0}^{N} \eta G^n g} \leq b + \frac{V(x)}{\sum_{n=0}^{N} \eta G^n g}. \tag{24}$$

By hypothesis, $L(., E) \in C_b(E)$, so there exists a sequence $\{f_i\}$ in $C_c(E)$ such that for all $x \in E$, $\lim_{n\to\infty} \uparrow f_i(x) = L(x, E)$. Therefore, with (24) we obtain

$$(\forall i \in \mathbb{N}) \quad \int_E f_i(y)\pi(dy) \leq b.$$

Now using the monotone convergence theorem, it follows that $\int_E L(y, E)\pi(dy) \leq b$.

So $\pi S(E) < \infty$ and with Theorem 3.5 the result follows. □

REMARK 4.8. *In the proof of Theorem* 4.7, *we need only that V is a norm-like function for the case in which $P_x(\tau_C < \infty) < 1$. Therefore, if the condition $P_x(\tau_C < \infty) = 1 \ \forall \ x \in E$ can be directly checked, then the norm-like function assumption can be removed from theorem* 4.7.

Let us introduce the following weak assumptions that will be used in Proposition 4.9 to show that the chain $\{Y_n\}$ is weak Feller.

(A1) The function $t_* : E \to (0, \infty]$ is continuous.

(A2) For any x in E, $\lambda(\Phi(y, t))I_{\{t<t_*(y)\}} \to \lambda(\Phi(x, t))I_{\{t<t_*(x)\}}$ when $y \to x$, μ_{Leb} a.s. on $(0, t_*(x))$.

(A3) For any x in E, there exists a neighborhood $\mathcal{N}(x)$ of x and a function $B_x(.) : \mathbb{R}_+ \to \overline{\mathbb{R}}_+$ such that $\forall\ y \in \mathcal{N}(x)$, $I_{\{t<t_*(y)\}}\lambda(\Phi(y,t)) \leq B_x(t)$, μ_{Leb} a.s. on $(0, t_*(x))$. If $\Lambda(x, t_*(x)) < \infty$, then $\int_0^{t_*(x)} B_x(s)ds < \infty$, otherwise $\int_0^t B_x(s)ds < \infty\ \forall\ t \in (0, t_*(x))$.

(A4) For any x in E, there exists a neighborhood $\mathcal{N}(x)$ of x and a constant $M_x > 0$ such that $\forall\ y \in \mathcal{N}(x)$, $\lambda(\Phi(y,t)) \exp\{-\Lambda(y,t)\} < M_x$.

(A5) For any open sets A in $\mathcal{E}$ and x in E,

$$Q(\Phi(y,t), A)I_{\{t<t_*(y)\}} \to Q(\Phi(x,t), A)I_{\{t<t_*(x)\}}$$

when $y \to x$, μ_{Leb} a.s. on $(0, t_*(x))$.

(A6) For any open sets A in $\mathcal{E}$ and x in E, $Q(\Phi(y, t_*(y)), A) \to Q(\Phi(x, t_*(x)), A)$, when $y \to x$.

Proposition 4.9. (i) *Suppose* (A1)–(A6)*. Then the stochastic kernel G and U are weak Feller and $L(., E) \in C_b(E)$.*

Proof. Let x and $t \in (0, t_*(x))$ be arbitrary fixed. Using assumptions (A2), (A3), and the bounded convergence theorem, we obtain that

$$\int_0^t \lambda(\Phi(y,s))I_{\{s<t_*(y)\}}ds \to \int_0^t \lambda(\Phi(x,s))I_{\{s<t_*(x)\}}ds \tag{25}$$

when $y \to x$. With (A5) it follows that for all open sets A in $\mathcal{E}$,

$$I_{\{t<t_*(y)\}}Q(\Phi(y,t), A)\lambda(\Phi(y,t)) \exp -\{t + \Lambda(y,t)\}$$

tends to

$$I_{\{t<t_*(x)\}}Q(\Phi(x,t), A)\lambda(\Phi(x,t)) \exp -\{t + \Lambda(x,t)\}$$

when $y \to x$, μ_{Leb} a.s. on $(0, t_*(x))$. Combining (A4) and the bounded convergence theorem, we have that

$$\int_0^{t_*(y)} Q(\Phi(y,t), A)\lambda(\Phi(y,t)) \exp -\{t + \Lambda(y,t)\}dt$$

tends to

$$\int_0^{t_*(x)} Q(\Phi(x,t), A)\lambda(\Phi(x,t)) \exp -\{t + \Lambda(x,t)\}dt$$

when $y \to x$.

If $\Lambda(x, t_*(x)) < \infty$, it is easy to deduce from (A3) that

$$\Lambda(y, t_*(y)) \to \Lambda(x, t_*(x))$$

when $y \to x$.

Now, if $\Lambda(x, t_*(x)) = \infty$, and $t_*(x) < \infty$ then

$$(\forall \varepsilon > 0), \quad (\exists \eta > 0) \quad \text{such that} \quad (\forall t \in [t_*(x) - \eta, t_*(x)]) \qquad \Lambda(x,t) > \frac{3\varepsilon}{2}. \tag{26}$$

Using (A1), there exists a neighborhood $\mathcal{N}_1(x)$ of x such that

$$(\forall y \in \mathcal{N}_1(x)) \qquad t_*(y) \in \left[t_*(x) - \frac{\eta}{2}, t_*(x) + \frac{\eta}{2}\right].$$

Therefore $t_*(x) - \eta < t_*(y)$.

So from (25), it follows that there exists a neighborhood $\mathcal{N}_2(x)$ of x such that

$$(\forall y \in \mathcal{N}_2(x)) \qquad |\Lambda(x, t_*(x) - \eta) - \Lambda(y, t_*(x) - \eta)| < \frac{\varepsilon}{2}. \tag{27}$$

Combining (26) and (27)

$$(\forall \varepsilon > 0), \quad (\forall y \in \mathcal{N}_1(x) \cap \mathcal{N}_2(x)), \qquad \Lambda(y, t_*(y)) > \varepsilon,$$

since $\forall\ x \in E$, $\Lambda(x, .)$ is increasing on $[0, t_*(x)]$.

The same result can be obtained for the case where $\Lambda(x, t_*(x)) = \infty$ and $t_*(x) = \infty$.

In conclusion,

$$\Lambda(y, t_*(y)) \to \Lambda(x, t_*(x))$$

when $y \to x$.

Using (A6), we obtain that for any open sets A in $\mathcal{E}$ and x in E, $K(y, A) \to K(x, A)$ when $y \to x$. Similarly, it can be shown that for any $f \in C_b(E)$, $Lf(y) \to Lf(x)$ when $y \to x$.

Consequently, it is straightforward with Proposition 6.1.1 in [14] to conclude that the stochastic kernel G is weak Feller.

Now using (12) and the result (12.7.7) in [6], it follows that for any positive function f in $C_b(E)$, Uf is a lower semicontinuous function. Since U is a stochastic kernel, Uf is continuous. Indeed, let $M_1 > 0$ be such that $f(x) < M_1$. Applying the previous result it follows that $U(M_1 - f)$ is lower semicontinuous. So Uf is upper semicontinuous and U is weak Feller. □

Therefore, the combination of Theorem 4.7 and Proposition 4.9 provides a second easy way of determining if there exists an invariant probability measure for $\{x_t\}$, as it will be shown in Example 1 of the next section.

REMARK 4.10. *An interesting property of the PDMPs is that under the weak conditions of Proposition* 4.9 *they generate weak Feller kernels G and U although the PDMPs themselves are generally not weak Feller Markov processes.*

Indeed, it was shown in [4, Example (27.5)] that the PDMP defined by $E = [0, 1[$, $\Phi(x, t) = x + t$, $\lambda(x) = 0$, and $Q(1, A) = I_A(0)$ is not a weak Feller process but it is easy to see that G and U are weak Feller kernels.

The results of this section have been developed considering the stochastic kernel G defined on $(E, \mathcal{E})$. It may be convenient in some applications to consider the extended state space $E \cup \Gamma^+$ instead of E, with the stochastic kernel G extended to include the points in Γ^+. Let us define the kernel H on $(E, \mathcal{B}(E \cup \Gamma^+))$ (corresponding to an extension of G) by $(\forall x \in E \cup \Gamma^+)$, $(\forall A \in \mathcal{B}(E \cup \Gamma^+))$:

$$H(x, A) \doteq I_E(x) G(x, A \cap E) + I_{\Gamma^+}(x) Q(x, A \cap E). \tag{28}$$

It is easy to see that H is a stochastic kernel on $(E, \mathcal{B}(E \cup \Gamma^+))$. Moreover, it has the following important properties:

(i) if π is a σ-finite invariant measure for H, then $\pi(\Gamma^+) = 0$.

(ii) a σ-finite measure is invariant for G if and only if it is invariant for H.

Therefore, all the results of the previous sections remain true if the kernel G is replaced by H.

It will be shown in the next section (see Example 1), that it may be useful to use H instead of G. Another possible illustration of the usefulness for introducing H is the following result.

COROLLARY 4.11. *Suppose* (A1)–(A6) and that $E \cup \Gamma^+$ is compact. Then there exists an invariant probability measure for the PDMP $\{x_t\}$.

Proof. As in Proposition 4.9, it can be shown that H is a weak Feller kernel. Now using the fact that $E \cup \Gamma^+$ is compact and Theorem 3.1 in [9], it follows that H has an invariant probability measure. Therefore, using the above properties of H and Theorem 3.5, we have the result. □

5. Examples. In this section we apply the results of sections 3 and 4 to three examples to emphasize the relevance of our approach. In the first one, we generalize Proposition (34.46) in [4] and Proposition 8 in [2]. The last two examples are taken from Davis [4]. They show that the equivalence between $\mathbf{P}_1$ and $\mathbf{P}_2$ with $\overline{G}$ replacing G may fail if some conditions are not satisfied. Due to its generality we are able to show that this kind of "counterexample" does not exist with our approach.

Example 1. This example is based on the capacity expansion model (see [2, section 7] and [4, Example (34.45)]. Capacity expansions are general processes of adding facilities to meet a rising demand. For a complete description of these models, the reader is referred to [5]. The demand for some utility is modeled as a random point process, i.e., it increases by one unit at random times. This demand is met by consecutive construction of expansion projects. Each project meets K_i units of demand when completed, where i corresponds the present level of excess demand. We assume that if there is an excess demand of at least p units, then the construction of a new project is started at a rate r_i per unit of time and it is completed after a lead time of $\mathcal{L}_i$ units of time. If the excess demand is less than p, then no construction takes place. New demand occurs with rate $\lambda_i(u)$ where u is the time spent by the current project. This problem can be modelled as a PDMP $\{x_t\}$ with state space

$$E \doteq \{\{p-K_p,\dots,p-1\}\times\{0\}\} \cup \bigcup_{n=p}^{\infty} \{n, [0,\mathcal{L}_n[\} , \tag{29}$$

where p and K_i are integers, $\mathcal{L}_i$ are strictly positive real numbers, and $\mathbb{N}_p$ denotes the set of integers greater than or equal to p. The local characteristics are given by

$$\lambda((i,u)) \doteq \lambda_i(u), \tag{30}$$

$$\Phi(t,(i,u)) = \begin{cases} (i,0) & \text{for } i \in \{p-K_p,\dots,p-1\}, \\ (i,u+r_i t) & \text{for } i \geq p, \end{cases} \tag{31}$$

$$Q((i,u),\{(i+1,u)\}) = 1 \quad \text{and} \quad Q((i,\mathcal{L}_i),\{(i-K_i,0)\}) = 1. \tag{32}$$

Unlike [2, 4], we allow here that the rate r_i, K_i, $\mathcal{L}_i$ can vary according to the excess demand i and that $\lambda_i(u)$ can vary according to the excess demand and to the time spent to complete the projects.

Using (30)–(32), it is easy to show that for a bounded real-valued function $W(.)$ defined on $E \cup \Gamma^+$ we have for $i \geq p$ and $u \in [0,\mathcal{L}_i]$

$$\begin{aligned} HW((i,u)) = &\int_0^{\frac{\mathcal{L}_i-u}{r_i}} \exp -\{s+\Lambda_i(u,s)\}\lambda_i(u+r_is)W((1+i,u+r_is))ds \\ &+\int_0^{\frac{\mathcal{L}_i-u}{r_i}} \exp -\{s+\Lambda_i(u,s)\}W((i,u+r_is))ds \\ &+\exp -\left\{\frac{\mathcal{L}_i-u}{r_i}+\Lambda_i\left(u,\frac{\mathcal{L}_i-u}{r_i}\right)\right\}W((i-K_i,0)), \end{aligned} \tag{33}$$

where $\Lambda_i(u,s) = \int_0^s \lambda_i(u + r_i v)dv$.

Let us introduce the following assumptions:

(H1) There exists $r > 0$ such that $\forall\ i$, $r_i \geq r$.

(H2) For $i \geq p$, $i - K_i \geq p - K_p > 0$ and $\lim_{i\to\infty} i - K_i = \infty$.

(H3) There exists an integer i_0 such that $\forall\ i \geq i_0$, $\frac{K_i}{\mathcal{L}_i} \leq \frac{K_{i+1}}{\mathcal{L}_{i+1}}$.

(H4) For $i \geq p$, $\lambda_i(u)$ is a continuous real-valued function on $[0, \mathcal{L}_i]$.

(H5) For $i \in \{p - K_p, \ldots, p-1\}$, $\lambda_i > 0$.

(H6) $\limsup_{i\to\infty} \frac{\mathcal{L}_i}{K_i r_i} \max_{u\in[0,\mathcal{L}_i]} \lambda_i(u) < 1$.

We have the following result.

Proposition 5.1. *Assume* (H1)–(H6). *Then there exists a unique invariant probability measure for the PDMP* $\{x_t\}$.

Proof. Let us introduce the real-valued function defined on $E \cup \Gamma^+$ by

$$V((i,u)) = i - \frac{K_i}{\mathcal{L}_i} u.$$

With (H2), it follows that the function V is positive norm-like function. Moreover, using (H1), (H3), and (H6), we have that there exists $\varepsilon > 0$ and $k_0 \geq \max\{p, i_0\}$ such that

$$(\forall i \geq k_0), \quad (\varepsilon + \lambda_i(u))\frac{\mathcal{L}_i}{K_i r_i} \leq 1. \tag{34}$$

Using (33) and the previous definition of $V(.)$, we obtain that

$$\begin{aligned} HV(i,u) - V(i,u) \leq & \int_0^{\frac{\mathcal{L}_i - u}{r_i}} \lambda_i(u + r_i s) \exp -\{s + \Lambda_i(u,s)\}ds \\ & -\frac{r_i K_i}{\mathcal{L}_i} \left\{ \int_0^{\frac{\mathcal{L}_i - u}{r_i}} s(1 + \lambda_i(u + r_i s)) \exp -\{s + \Lambda_i(u,s)\}ds \right. \\ & \left. + \frac{\mathcal{L}_i - u}{r_i} \exp - \left\{ \frac{\mathcal{L}_i - u}{r_i} + \Lambda_i\left(u, \frac{\mathcal{L}_i - u}{r_i}\right) \right\} \right\} \\ \leq & \int_0^{\frac{\mathcal{L}_i - u}{r_i}} \left(\lambda_i(u + r_i s) - \frac{r_i K_i}{\mathcal{L}_i} \right) \exp -\{s + \Lambda_i(u,s)\}ds. \end{aligned}$$

Using (34) and the fact that for (i,u) in $E \cup \Gamma^+$,

$$L((i,u),E) = \int_0^{\frac{\mathcal{L}_i - u}{r_i}} \exp -\{s + \Lambda_i(u,s)\}ds,$$

we obtain that there exists a constant b and a compact set

$$C \doteq \{\{p - K_p, \ldots, p-1\} \times \{0\}\} \cup \bigcup_{n=p}^{k_0} \{n, [0, \mathcal{L}_n]\}$$

such that

$$HV(i,u) - V(i,u) \leq -\varepsilon L((i,u),E)) + bI_C((i,u)).$$

It is easy to check that assumptions (A1)–(A6) are satisfied for this process. So, $L(.,E)$ is a continuous function and the stochastic kernel H is weak Feller. Applying Theorem 4.7, the result follows. □

Remark 5.2. *The previous proposition generalizes the result of Costa* [2] *and Davis* [4] *since here K_i, r_i, $\mathcal{L}_i$, and λ_i (as functions of i) may be unbounded, or $\lambda_i(u)$ may depend on the time u spent to complete the project.*

Example 2 (see [4, Example (34.28)]). This example was designed in [4] to point out the fact that the PDMP $\{x_t\}$ may have an invariant probability measure but the associated Markov chain $\{Z_1, Z_2, \ldots\}$ may not. Indeed, the approach developed by Costa [2] suffers from the fact that if μ is an invariant probability measure for $\{x_t\}$, then the associated Markov chain $\{Z_1, Z_2, \ldots\}$ will have an invariant probability measure if the following boundedness condition is satisfied: $\int_E \lambda(x)\mu(dx) < \infty$. Let us show, however, that in our case there exists a σ-finite invariant measure π for the stochastic kernel G as claimed in Theorem 3.5 such that $\pi S(E) < \infty$.

In this case $E = \mathbb{Z}$ and $\forall\ x \in E$, $\Phi(x,t) = x$, $Q(x, \{x+1\}) = Q(x, \{x-1\}) = \frac{1}{2}$ with $\lambda(x)$ satisfying the condition

$$\sum_{x=-\infty}^{\infty} \frac{1}{\lambda(x)} < \infty.$$

It was shown in [4] that $\{x_t\}$ has a unique invariant probability measure μ such that

$$(\forall x \in \mathbb{Z}) \quad \mu(\{x\}) = \frac{1}{\lambda(x)} \sum_{y=-\infty}^{\infty} \frac{1}{\lambda(y)}.$$

Now it is easy to obtain that

$$(\forall x \in \mathbb{Z}) \quad G(x, \{y\}) = \begin{cases} \frac{1}{2}\frac{\lambda(x)}{1+\lambda(x)} & \text{for } y = x+1 \text{ or } y = x-1, \\ \frac{1}{1+\lambda(x)} & \text{for } y = x, \\ 0 & \text{otherwise.} \end{cases} \tag{35}$$

It is easy to verify that $\forall\ x \in \mathbb{Z}$ the measure π defined by

$$(\forall x \in \mathbb{Z}) \quad \pi(\{x\}) = \frac{1+\lambda(x)}{\lambda(x)} \sum_{y=-\infty}^{\infty} \frac{1}{\lambda(y)}$$

is invariant for G.

Indeed,

$$\begin{aligned} \sum_{y=-\infty}^{\infty} G(y, \{x\})\pi(\{y\}) &= G(x-1, \{x\})\pi(\{x-1\}) + G(x, \{x\})\pi(\{x\}) \\ &\quad + G(x+1, \{x\})\pi(\{x+1\}) \\ &= \pi(\{x\}). \end{aligned}$$

Moreover, using the fact that $S(x, \mathbb{Z}) = \frac{1}{1+\lambda(x)}$, it is easy to prove that $\pi S(\mathbb{Z}) < \infty$. Therefore as claimed by Theorem 3.5, if there exists a solution to $\mathbf{P}_1$, then there exists one to $\mathbf{P}_2$.

For this example the fact that $\{x_t\}$ has an invariant probability measure had been established by direct calculation (see [4]). We want now to illustrate the use of the weak conditions of Theorem 4.5 for a suitable choice of function $V(.)$ and petite set C and recover this known result. Indeed, it is straightforward to prove that the Markov chain $\{Y_n\}$ associated with the kernel G is φ-irreducible for $\varphi(\{x\}) = \frac{1}{2^{|x|}}$. Then $\forall$ $x \in E$ we have $\psi(\{x\}) = \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^{n+1} \varphi G^n(\{x\}) > 0$. Therefore $\mathcal{E} = \mathcal{E}^+$.

Let us show that $\{x\}$ is a petite set for all $x \in \mathcal{E}$. Define

$$\nu(\{y\}) = \begin{cases} \frac{1}{1+\lambda(x)} & \text{for } y = x, \\ \frac{1}{2}\frac{\lambda(x)}{1+\lambda(x)} & \text{for } y = x+1 \text{ or } y = x-1, \\ 0 & \text{otherwise.} \end{cases}$$

So, we have $(\forall B \in \mathcal{E})$, $G(x,B) \geq \nu(B)$, and the claim follows.

Now let us prove that $\{Y_n\}$ is recurrent. Define the function $V_1(i)$ for $i \in \mathbb{Z}$ by

$$V_1(i) = 2\sum_{l=0}^{i}\sum_{j=l}^{\infty} \frac{1}{(j+1)^2} \quad \text{for } i \geq 0,$$

$$V_1(i) = 2\sum_{l=i}^{-1}\sum_{j=-\infty}^{l} \frac{1}{(|j|+1)^2} \quad \text{for } i \leq -1.$$

Then, for $i \geq 1$ and using (35), we have

$$\begin{aligned} GV_1(i) - V_1(i) &= \frac{\lambda(i)}{1+\lambda(i)}\left\{\frac{1}{2}V_1(i+1) + \frac{1}{2}V_1(i-1) - V_1(i)\right\} \\ &= \sum_{j=i+1}^{\infty}\frac{1}{(j+1)^2} - \sum_{j=i}^{\infty}\frac{1}{(j+1)^2} \quad = -\frac{1}{(i+1)^2}. \end{aligned}$$

Similarly, for $i \leq -2$, we obtain

$$GV_1(i) - V_1(i) = -\frac{1}{(|i|+1)^2}.$$

Therefore, the norm-like function $V_1(.)$ satisfies condition $(\mathcal{D})$ in [14, p. 190] for $C = \{-1, 0\}$. Moreover, it is clear that this function is unbounded off petite sets and that $C = \{-1, 0\}$ is a petite since it is a finite set. Consequently, using Theorem 8.4.3 in [14, p. 191], we obtain that $\{Y_n\}$ is recurrent.

In order to show that condition $(\mathcal{D})$ is satisfied for the petite set $C = \{-1, 0\}$, let us introduce the function $V_2(i)$ for $i \in \mathbb{Z}$ by

$$V_2(i) = 2\sum_{l=0}^{i}\sum_{j=l}^{\infty} \frac{1}{\lambda(j)} \quad \text{for } i \geq 0,$$

$$V_2(i) = 2\sum_{l=i}^{-1}\sum_{j=-\infty}^{l} \frac{1}{\lambda(j)} \quad \text{for } i \leq -1.$$

Using (35), we obtain

$$\begin{aligned} GV_2(i) - V_2(i) &= -\frac{1}{1+\lambda(i)} \\ &= -L(i, \mathbb{Z}) \end{aligned}$$

for $i \in \mathbb{Z} - C$.

Therefore, the function $V_2(.)$ satisfies condition $(\mathcal{D})$ for a suitable constant b. From Corollary 4.5, we conclude that there exists a unique invariant probability measure for the PDMP $\{x_t\}$.

Example 3 (see [4, Example (34.33)]). This example shows that the inverse problem as described in the previous example may happen. Indeed, if there exists an invariant probability measure (labeled π) for the stochastic kernel $\overline{G}$, then an invariant probability measure for the PDMP $\{x_t\}$ will exist if the following condition is satisfied: $\int_E \int_0^{t_*(x)} \exp\{-\Lambda(x,s)\} ds \pi(dx) < \infty$. Davis studied the following case: $E = \mathbb{N}$, $\Phi(x,t) = x$, $Q(x, \{x+1\}) = p$, $Q(x, \{x-1\}) = q$, $Q(0, \{1\}) = p$, $Q(0, \{0\}) = q$ with $q = 1 - p$, $p < \frac{1}{2}$. $\lambda(x)$ is bounded. He has shown that there exists an invariant probability measure for $\{x_t\}$ if and only if

$$\sum_{y=0}^{\infty} \frac{1}{\lambda(y)} \left(\frac{p}{q}\right)^y < \infty. \tag{36}$$

Therefore, for the choice of $\lambda(x) = \left(\frac{p}{q}\right)^x$, $\{x_t\}$ has no invariant probability measure although the birth and death process generated by $\overline{G}$ has one. However, we show that there exists an invariant probability measure for the stochastic kernel G if and only if there exists one for $\{x_t\}$, according to the fact that $\lambda(.)$ is bounded and Remark 3.8. By direct calculations, it is easy to show that $(\forall x \in \mathbb{N})$, $G(x, \{y\}) = 0$ if $y \notin \{x-1, x, x+1\}$, and for $x \geq 1$, $G(x, \{x\}) = \frac{1}{1+\lambda(x)}$, $G(x, \{x+1\}) = \frac{p\lambda(x)}{1+\lambda(x)}$, $G(x, \{x-1\}) = \frac{q\lambda(x)}{1+\lambda(x)}$, and $G(0, \{0\}) = \frac{1+q\lambda(0)}{1+\lambda(0)}$, $G(0, \{1\}) = \frac{p\lambda(0)}{1+\lambda(0)}$.

If π is an invariant measure for $\{x_t\}$, it must satisfy the following equations: $(\forall x \geq 1)$,

$$\pi(\{x\}) = G(x-1, \{x\})\pi(\{x-1\}) + G(x, \{x\})\pi(\{x\}) + G(x+1, \{x\})\pi(\{x+1\})$$

and

$$\pi(\{0\}) = G(0, \{0\})\pi(\{0\}) + G(1, \{0\})\pi(\{1\}).$$

Therefore,

$$(\forall x \geq 1), \quad pc(x-1) + qc(x+1) - c(x) = 0$$

and

$$-pc(0) + qc(1) = 0$$

with $c(x) \doteq \frac{\lambda(x)}{1+\lambda(x)}\pi(\{x\})$. The unique positive solution of the previous equations is

$$(\forall x \in \mathbb{N}), \quad c(x) = c(0)\left(\frac{p}{q}\right)^x.$$

So, there exists an invariant probability measure for the stochastic kernel G if and only if

$$\sum_{y=0}^{\infty} \pi(\{y\}) < \infty,$$

that is, if and only if

$$\sum_{y=0}^{\infty} \frac{1}{\lambda(y)} \left(\frac{p}{q}\right)^{y} < \infty,$$

which is the same necessary and sufficient condition for the PDMP (see (36)) in agreement with Theorem 3.5.

6. Conclusion. In this paper, we have shown that the existence of an invariant probability measure for a general PDMP is equivalent to the existence of a σ-finite invariant measure for the stochastic kernel G satisfying a boundedness condition or equivalently a Radon–Nikodým derivative. Here we generalize existing results of the literature [2, 4] since we do not require any additional assumptions to establish this equivalence. Moreover, we give sufficient conditions to ensure the existence of such a σ-finite measure satisfying the boundedness condition. They are based mainly on a modified Foster–Lyapunov criterion. To emphasize the relevance of our results, we studied three examples and in particular, we are able to generalize the results obtained by Costa and Davis on the capacity expansion model.

REFERENCES

[1] J. Azema, M. Kaplan-Duflo, and D. Revuz, *Mesure invariante sur les classes récurrentes des processus de markov*, Z. Wahrscheinlichkeitstheorie verw. Geb., 8 (1967), pp. 157–181.

[2] O. Costa, *Stationary distributions for piecewise-deterministic Markov processes*, J. Appl. Probab., 27 (1990), pp. 60–73.

[3] M. Davis, *Piecewise-deterministic markov processes: A general class of non-diffusion stochastic models*, J. Roy. Statist. Soc. Ser. B, 46 (1984), pp. 353–388.

[4] M. Davis, *Markov Models and Optimization*, Chapman and Hall, London, 1993.

[5] M. Davis, M. Dempster, S. Sethi, and D. Vermes, *Optimal capacity expansion under uncertainty*, Adv. in Appl. Probab., 19 (1987), pp. 156–176.

[6] J. Dieudonné, *Eléments d'Analyse,* Tome 2, Gauthier-Villard, Paris, 1974.

[7] S. Foguel, *The ergodic theory of positive operators on continuous functions*, Ann. Scuola Norm. Sup. Pisa, C. Sci. (4) 27, (1973), pp. 19–51.

[8] O. Hernandez-Lerma and J.-B. Lasserre, *Existence and uniqueness of fixed points for Markov operators and Markov processes*, Proc. London Math. Soc. (3), 76 (1998), pp. 711–736.

[9] J.-B. Lasserre, *Invariant probabilities for Markov chains on a metric space*, Statis. Probab. Lett., 34 (1997), pp. 259–266.

[10] M. Lin, *Conservative Markov processes on a topological space*, Israel J. Math., 8 (1970), pp. 165–186.

[11] S. P. Meyn, *Ergodic theorems for discrete time stochastic systems using a stochastic Lyapunov function*, SIAM J. Control Optim., 27 (1989), pp. 1409–1439.

[12] S. P. Meyn and P. E. Caines, *Asymptotic behavior of stochastic systems possessing Markovian realizations*, SIAM J. Control Optim., 29 (1991), pp. 535–561.

[13] S. Meyn and R. Tweedie, *Criteria for stability of markovian processes* III*: Foster–Lyapunov criteria for continuous time processes, with examples*, Adv. in Appl. Probab., 25 (1993), pp. 518–548.

[14] S. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, Berlin, 1993.

[15] S. Meyn and R. Tweedie, *Stability of markovian processes* II*: Continuous-time processes and sampled chains*, Adv. in Appl. Prob., 25 (1993), pp. 487–517.

[16] E. Nummelin, *General Irreducible Markov Chains and Non-negative Operators*, Cambridge University Press, Cambridge, 1984.

[17] D. Revuz, *Markov Chains*, North–Holland, Amsterdam, 1975.

[18] K. Yosida, *Functional Analysis*, 6th ed., Springer-Verlag, New York, 1980.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 5, pp. 1503–1537

A BACKSTEPPING CONTROLLER FOR A NONLINEAR PARTIAL DIFFERENTIAL EQUATION MODEL OF COMPRESSION SYSTEM INSTABILITIES*

ANDRZEJ BANASZUK†, HÖSKULDUR ARI HAUKSSON‡, AND IGOR MEZIĆ§

Abstract. We prove the existence and uniqueness of solutions in Sobolev spaces for the Moore–Greitzer nonlinear partial differential equation (PDE) model for compression system instabilities with mild conditions on the shape of the compressor characteristic and on the throttle control. To achieve this, the model is reformulated as an evolution equation on a Banach space. Using this new representation, we design a backstepping control of the model. Global stabilization of any axisymmetric equilibrium to the right of the peak of the compressor characteristic is achieved. We also prove that the dynamics can be restricted to the small neighborhood of the point on the left of the peak of the compressor characteristic. Thus, it is possible to restrict the magnitude of stall to arbitrary small values. In addition, finite-dimensional Galerkin projections of the partial differential equation model are studied. It is shown that truncated control laws stabilize truncated models. Numerical simulations of the model with and without control are presented.

Key words. control of partial differential equations, backstepping control of nonlinear differential equations, control of stall and surge in aeroengines, existence and uniqueness of solutions of partial differential equations

AMS subject classifications. 93C10, 93B29, 93B18, 53C65

PII. S0363012997317876

1. Introduction.

Surge and stall instabilities that occur in compression systems of jet engines are the topic of much research effort these days for two reasons: efficiency and safety. In particular, jet engines are currently forced to operate in nonoptimal conditions (relatively large mass flow) in order to stay clear of the aforementioned instabilities. Surge is an oscillatory instability of the mean mass flow: upon the onset of surge the air in the compression system of a jet engine starts oscillating back and forth, thus severely impairing its performance. Stall is characterized by the appearance of the so-called *stall cells*—regions of decreased pressure rise and reversed mass flow—at isolated locations around the rim of the compressor. A simplified model of these instabilities has been proposed by Moore and Greitzer [15], and it is this model (sometimes called the *full Moore–Greitzer model*) that is the topic of the present paper. The model consists of a linear PDE governing the behavior of disturbances in the inlet region of the compression system, with nonlocal and nonlinear boundary conditions which describe the coupling of disturbances with the mean flow behavior. Since such a system is hard to analyze, most of the research has been directed toward establishing properties of low-order Galerkin truncations (see [15], [11]). The simplest

*Received by the editors March 4, 1997; accepted for publication (in revised form) July 15, 1998; published electronically August 2, 1999. This research was supported in part by AFOSR under F49620-95-1-0409.

http://www.siam.org/journals/sicon/37-5/31787.html

†United Technologies Research Center, 411 Silver Lane, MS 129-15, East Hartford, CT 06108 (banasza@utrc.utc.com). The research of this author was mostly completed at the Department of Mathematics, University of California, Davis. Also, a significant part of the research presented in this paper was conducted during two visits to the Center for Control Engineering and Computation of University of California at Santa Barbara.

‡Department of Mathematics, University of California, Santa Barbara, CA 93106 (hauksson @math.ucsb.edu).

§Department of Mechanical Engineering, University of California, Santa Barbara, CA 93106 (mezic@ana.ucsb.edu).

approximate model used for bifurcation analysis (see [12]) and control (see [3], [6], [8], [9], [10], [16], [1]) called MG3 used only the first Fourier mode of the nonaxisymmetric flow disturbance (stall variable).

The Moore–Greitzer model deals with a simple compression system in which the air enters the compressor (with one or more rotor/stator stages), goes to a plenum, and exits through a throttle. Stationary operating points for the compressor correspond to a constant pressure rise across the compressor and a constant, circumferentially uniform mass flow through the compressor. The pressure rise versus mass flow curve representing the stationary operating points is called the compressor characteristic. For a given throttle opening the mass flow through the throttle is determined by the pressure drop across the throttle. The corresponding static relationship can be represented by the curve called the throttle characteristic. In a stationary condition the pressure rise across the compressor is balanced by the pressure drop across the throttle and the mass flow through the compressor and through the throttle are equal. Therefore, the intersection of the compressor characteristic and the throttle characteristic determines the operating point of the compressor. The operating point can be changed by adjusting the throttle opening.

The dynamic model for compression systems derived by Moore and Greitzer [15] describes the evolution of the mass flow and the pressure in plenum in a nonstationary condition. When the pressure rise across the compressor is not balanced by the pressure drop across the throttle, the resulting pressure difference is proportional to the rate of change of mass flow (i.e., mass acceleration). Similarly, if the mass flow through the compressor is not balanced by the mass flow through the throttle, the resulting difference is proportional to the rate of pressure rise in the plenum.

In terms of the dynamic model of a compression system, the stationary operating points are represented by the axisymmetric equilibria, surge corresponds to a limit cycle involving pressure rise and mass flow, while rotating stall is represented by a travelling wave of nonaxisymmetric mass flow around the compressor annulus with a constant low value of pressure rise.

From the point of view of efficiency, the desired operating points for the compressor should have high value of pressure rise and low mass flow that is uniform around the compressor annulus. However, the analysis of the dynamic model shows that the corresponding equilibria of the dynamic model have small domains of attraction that are shrinking as the pressure rise increases. A small disturbance is likely to force a transition of the compression system state to either rotating stall or surge.

In the present paper we study *stabilization* of a given axisymmetric equilibrium using the throttle opening as a control variable. The throttle opening is considered to be some function of the state of the system chosen so that the corresponding new dynamic model has a unique globally stable equilibrium at the prescribed location. While bounded disturbances would force the state of the system to evolve in some neighborhood of the desired equilibrium, after they disappear, the state would eventually return to an arbitrary small neighborhood of the equilibrium. Physically, the control would be implemented by varying the throttle opening in the way that prevents transition into rotating stall or surge.

After some manipulation, the dynamic model of compression system will be represented as an evolution equation in a Banach space. Stabilization of the desired equilibrium will be achieved by constructing a Lyapunov function for this equilibrium. A special pure feedback structure of the evolution equation will allow us to use *backstepping* for construction of a Lyapunov function. This technique has been introduced in [7] and applied to many systems described by nonlinear ordinary differential equations (ODEs) with pure feedback structure, including a reduced-order model of compressor dynamics, MG3, in [8] and [1].

Experimental observations of the nonaxisymmetric flow disturbance (stall) behavior indicate that its shape is often far from sinusoidal [2]. Our investigations of the full Moore–Greitzer model show that the resulting nonlinear behavior of the compressor is well represented by the model [14]. It is then natural to ask what can be said about the control of stall and surge using the PDE model. In this paper we show that a global stabilization of the full Moore–Greitzer model is possible. We present a conceptually simple but not necessarily optimal way of constructing a globally stabilizing controller using backstepping control design [7]. To our knowledge, the present paper presents the first successful attempt to globally stabilize a nonlinear PDE using backstepping. We concentrate on a specific model here, but the methods that we develop can be used more broadly. A variety of evolution equation problems with a pure feedback structure can be treated in a way similar to that presented here. The backstepping method presents a powerful tool even in the context of PDEs.

The paper is organized as follows. In section 2 we introduce some notation and represent the full Moore–Greitzer model as an evolution equation in a Banach space, using an operator that is an infinite-dimensional version of that studied by Mansoux, Gyrling, Statiawan, and Paduaro in [11]. In section 3 we prove global existence and uniqueness of solutions of this evolution equation by a simple application of the contraction mapping principle. We present some a priori estimates which, together with the existence of a unique local solution, guarantee the existence of a unique global solution. In section 4 we design a backstepping controller for the full Moore–Greitzer model. We show that the peak and any axisymmetric equilibrium to the right of the peak can be globally asymptotically stabilized.

In the case when the set-point parameter in the controller is such that there is no stable axisymmetric equilibrium we can still guarantee that the dynamics of the closed-loop system are confined to a ball, whose radius can be made arbitrarily small by choosing sufficiently high gains in the controller.

In section 5 we prove that the truncated feedback controller globally stabilizes the system of $2n+2$ ODEs consisting of the Galerkin projection of the PDE describing the stall dynamics onto its first n-modes and the two ODEs describing the surge dynamics. The results are valid for a general compressor characteristic.

2. Preliminaries.

2.1. The Moore–Greitzer model. The full Moore–Greitzer model is described by the following equations (cf. [15])

$$l_c \frac{d\Phi}{d\xi} = -\Psi(\xi) + \frac{1}{2\pi}\int_0^{2\pi} \Psi_c(\Phi + \phi'_\eta|_{\eta=0})d\theta, \tag{1}$$

$$l_c \frac{d\Psi}{d\xi} = \frac{1}{4B^2}(\Phi(\xi) - K_T(\Psi, u)), \tag{2}$$

where ϕ' solves Laplace's equation

$$\phi'_{\eta\eta} + \phi'_{\theta\theta} = 0 \tag{3}$$

for $(\eta, \theta) \in [0, 2\pi] \times (-\infty, 0)$. The boundary conditions are periodic in θ,

$$\frac{\partial}{\partial\xi}(m\phi' + \frac{1}{a}\phi'_\eta) - \left(\Psi_c(\Phi + \phi'_\eta) - \frac{1}{2\pi}\int_0^{2\pi} \Psi_c(\Phi + \phi'_\eta)d\theta - \frac{1}{2a}\phi'_{\theta\eta}\right) = 0 \tag{4}$$

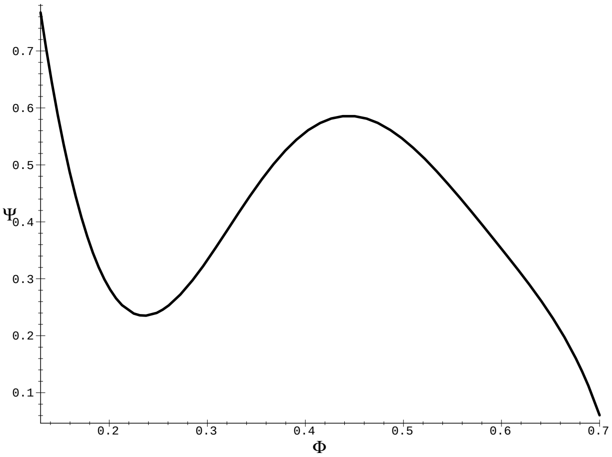


FIG. 1. C_3' *compressor characteristic.*

at $\eta = 0$. At $\eta = -\infty$ we have

$$\phi' = 0. \tag{5}$$

(Note that we try to keep our notation consistent with that of the original paper by Moore and Greitzer [15]. In particular, following [15] we use ξ to denote a nondimensional time variable and η to denote a nondimensional axial distance variable.) The state variables of this model are Φ, the nondimensionalized annulus averaged mass flow coefficient through the compressor; Ψ, the nondimensionalized annulus averaged pressure rise coefficient across the compressor; and ϕ', the disturbance velocity potential. (Note that the prime symbol in ϕ' *does not* refer to differentiation.) The function $\Psi_c(\phi)$ is called the compressor characteristic and is found empirically. It gives the local pressure rise when the local mass flow is ϕ. For most compressors it has an S shape as seen in Figure 1. The parameters a, m, l_c, and B are determined by the geometry of the compressor and the throttle parameter $K_T(\Psi, u)$ is the fraction of the throttle opening. Since the throttle parameter can be varied, it will be used as the control. We assume that we can modify $K_T(\Psi, u)$ at will by a choice of the control function u.

We assume that the compressor characteristic $\Psi_c(\Phi)$ is a general S-shaped curve. In particular, we assume that the following characteristics hold.

1. The characteristic $\Psi_c(\Phi)$ is twice continuously differentiable.

2. The characteristic has one peak (Φ_0, Ψ_0) and, to the left of the peak, one well. The characteristic is strictly decreasing to the right of the peak and to the left of the well; it is strictly increasing between the well and the peak.

3. The characteristic has exactly one inflection point $(\Phi_{infl}, \Psi_{infl})$ between the well and the peak. One has $\Psi_c''(\Phi) < 0$ for $\Phi > \Phi_{infl}$, and $\Psi_c''(\Phi)$ is bounded away from zero on any interval $[\Gamma, +\infty)$ for $\Gamma > \Phi_{infl}$.

Figure 1 shows a typical compressor characteristic $\Psi_c(\phi)$.

Let us give here a physical interpretation for the shape of the compressor characteristic. The desired possible stationary operating points of the compressor lie on the decreasing part of the characteristic to the right of the peak. The lower the value of the axial component of the mass flow entering compressor, the more flow turning is achieved by the blades. Consequently, more work is done on the air by the compressor blades and the pressure rise is higher. However, there is a limit to the value of pressure rise that can be achieved. When the axial component of the incoming air velocity is small relative to the rotor velocity, so that the air is approaching the blade at a high angle of attack, the flow separates on the suction side of a compressor blade; i.e., the blade *stalls*. When a blade stalls, the pressure rise at that blade drops significantly. The pressure rise of a stalled blade is represented by the part of the characteristic between the well and the peak. The peak represents the *stall inception point*: the maximum pressure rise is obtained by a blade that is just about to stall. When the flow is reversed, the air at the suction side of the blade is attached again and the pressure starts to rise. This is represented by the part of the characteristic to the left of the well, called the back-flow part of the characteristic.

The physical mechanism for rotating stall and surge inception can now be explained. When a blade stalls, the pressure in the plenum is usually greater than the local pressure rise produced by the compressor, so that the incoming air at the stalled blade faces a negative pressure gradient and hence has a negative acceleration. The mass flow at the stalled blade passage is locally reduced. There are several possible scenarios of how the situation will evolve. The extreme ones are a transition to a surge or rotating stall condition.

In surge the pressure in the plenum does not drop fast enough, and all the blades stall at the same time. The flow eventually reverses, as the only mechanism to balance the high pressure in the plenum at the stalled blades is for the mass flow to reach the back-flow part of the characteristic. This transition from the neighborhood of the stall inception point to the back-flow characteristic is very fast. The pressure in the plenum is now dropping, as the air escapes from the plenum both through the throttle and through the compressor. The pressure in the plenum eventually drops below the value of the pressure rise on the back-flow part of the characteristic and the pressure gradient becomes positive. The air accelerates slowly until zero mass flow is reached. The plenum pressure is now below the well value. The compressor starts to deliver more pressure rise while the pressure in plenum is about the well level, so the mass flow accelerates fast. Past the value corresponding to the peak the flow at the blades becomes attached. When the flow through compressor becomes bigger than the flow through the throttle, the pressure starts to rise and becomes bigger than the one produced by the compressor. The flow starts to decrease. When the flow reaches the value corresponding to the peak, pressure in the plenum is about the peak value, which is the condition of a stall inception. One full surge cycle is now completed and the next one is about to start.

When rotating stall occurs, one or several blades stall. Locally, the flow is redirected to the neighboring unstalled blades. On one side of the region of stalled blades, the angle of attack of the air flow will increase, causing more blades to stall. On the other side the angle of attack will decrease, making the blades on that side less susceptible to stalling. The air is coming to the blades in the direction of spinning rotor at high angle of attack. These blades are likely to stall. At the same time the blades neighboring the stalled ones in the direction opposite to the spinning rotor accept air at the lower angle of attack than the stalled ones so they are not likely to stall. More air coming through these blades may lower the angle of attack on the first stalled blade, which will result in more local pressure rise. If the pressure in plenum drops fast enough, the pressure gradient on some of the stalled blades will become positive and these blades will unstall. A stable rotating stall condition may develop when some of the blades operate in a stalled condition and the rest are not stalled. The cells of stalled air travel around the compressor so that each blade periodically becomes stalled and unstalled. Note that in a rotating stall condition the average pressure rise delivered by the compressor is low, as the stalled blades barely do any work on the air. Note also that in a stable rotating stall condition the air mass flow in the stalled blade passages is reversed, as the stalled blades cannot deliver enough pressure rise to balance the pressure in the plenum.

At this point the physical mechanism for the stabilization of the operating point close to the peak of the characteristic by varying the throttle opening can be explained. In both stall and surge inception the mechanism of instability is the same: a stalled blade cannot deliver enough pressure rise to balance the high pressure in the plenum and the resulting negative pressure gradient decelerates and eventually reverses the flow at some (stall) or all (surge) blade passages. The control action basically amounts to opening the throttle fast enough so that the pressure in the plenum drops faster than the pressure at a stalled blade. This produces the positive pressure gradient that accelerates the flow to the desired value. After this value is reached, the throttle is closed again.

2.2. Some function spaces. Let $\mathcal{L}^2$ be the space of square integrable functions on the circle $[0, 2\pi]$ and denote the norm by $\|\cdot\|_{\mathcal{L}^2}$. Let L^2, L^∞, and H^k, for $k = 1, 2, \dots$, denote the subspaces of $\mathcal{L}^2$ with zero average and norms $\|\cdot\|_{L^2}$, $\|\cdot\|_{L^\infty}$, $\|\cdot\|_{H^k}$. These norms are given by

$$\begin{aligned}
\|g\|_{L^2} &:= (\textstyle\int_0^{2\pi} g^2 d\theta)^{\frac{1}{2}} = (\pi \sum_{p=1}^{\infty} A_p(\xi)^2)^{\frac{1}{2}},\\
\|g\|_{H^k} &:= (\textstyle\int_0^{2\pi} (\frac{\partial^k g}{\partial \theta^k})^2 d\theta)^{\frac{1}{2}} = (\pi \sum_{p=1}^{\infty} (p^k A_p(\xi))^2)^{\frac{1}{2}},\\
\|g\|_{L^\infty} &:= \operatorname{esssup}_{\theta\in[0,2\pi]} g.
\end{aligned}$$

Here, the A_p's represent the magnitudes of the complex Fourier coefficients of g,

$$g = \sum_{p=1}^{\infty} A_p(\xi) \sin(p\theta + r_p(\xi)).$$

We denote by $\langle \cdot, \cdot \rangle$ the inner product of L^2:

$$\langle g_1, g_2 \rangle = \int_0^{2\pi} g_1 g_2 d\theta = \pi \sum_{p=1}^{\infty} A_p B_p.$$

Here, the A_p's and the B_p's represent the Fourier coefficients of the functions g_1 and g_2, respectively.

Let C^0 denote the space of continuous functions on the circle $[0, 2\pi]$ with zero average, with the norm

$$\|g\|_{C^0} := \max_{\theta\in[0,2\pi]} g.$$

Note that for $g \in C^0$ one has $\|g\|_{C^0} = \|g\|_{L^\infty}$. Thus, to avoid using too many symbols we will use $\|g\|_{L^\infty}$ to denote the norm of C^0 functions.

For future reference, we collect here some inequalities in the following lemma.

LEMMA 2.1. *One has* $H^1 \hookrightarrow C^0 \hookrightarrow L^\infty \hookrightarrow L^2$ *and*

$$\begin{aligned} \|g\|_{L^2} &\leq \sqrt{2\pi}\|g\|_{L^\infty}, \\ \|g\|_{L^\infty} &\leq \tfrac{\sqrt{\pi}}{\sqrt{6}}\|g\|_{H^1}, \\ \|g\|_{L^2} &\leq \|g\|_{H^1}. \end{aligned}$$

Proof. We first prove the inequalities. The first inequality is clear. The second one follows from $\|g\|_{L^\infty} \leq \sum_{p=1}^\infty |A_p(\xi)| \leq (\sum_{p=1}^\infty (\frac{1}{p^2}))^{\frac{1}{2}} (\sum_{p=1}^\infty (pA_p(\xi))^2)^{\frac{1}{2}} = \frac{\sqrt{\pi}}{\sqrt{6}}\|g\|_{H^1}$. The third one is the Poincaré inequality. $H^1 \hookrightarrow C^0$ is the Sobolev embedding of H^1 into C^0 in spatial dimension one. The other embeddings follow from the inequalities. □

Assume for now that ϕ' can be represented as

$$\phi' = \sum_{p=1}^\infty \mathrm{e}^{p\eta}\alpha_p(\xi)\sin(p\theta + r_p(\xi)), \tag{6}$$

where α_p and r_p are real functions.

Let $g := \phi'_\eta|_{\eta=0}$; then

$$g = \sum_{p=1}^\infty p\alpha_p(\xi)\sin(p\theta + r_p(\xi)) =: \sum_{p=1}^\infty A_p(\xi)\sin(p\theta + r_p(\xi)).$$

Equation (4) can therefore be written as

$$\frac{\partial}{\partial\xi}Kg = a\left(\Psi_c(\Phi + g) - \frac{1}{2\pi}\int_0^{2\pi} \Psi_c(\Phi + g)d\theta - \frac{1}{2}\frac{\partial g}{\partial\theta}\right), \tag{7}$$

where the operator K is defined as follows:

$$K\sum_{p=1}^\infty \alpha_p(\xi)\sin(p\theta + r_p(\xi)) = \sum_{p=1}^\infty \left(1 + \frac{am}{p}\right)\alpha_p(\xi)\sin(p\theta + r_p(\xi)). \tag{8}$$

The operator K is an infinite-dimensional analogue of the operator introduced in [11] for a study of finite-dimensional truncations of the full Moore–Greitzer model.

Remark 2.1. Suppose that we can show that the system (1), (2), and (7) has a unique solution such that $g \in H^1$. Then from the Fourier series representation of g we can calculate the corresponding potential ϕ'. Since $g \in H^1$, it follows that on the cylinder $[0, 2\pi] \times (-\infty, 0)$ the potential ϕ' is in the Sobolev space $H^2([0, 2\pi] \times (-\infty, 0))$. In particular, the partial derivatives $\phi'_{\eta\eta}$ and $\phi'_{\theta\theta}$ are in $L^2([0, 2\pi] \times (-\infty, 0))$. Note also that from (6) it follows that ϕ' satisfies the Laplace equation (3) and the boundary conditions (4) and (5). The existence of solutions of the full Moore–Greitzer model follows. The uniqueness follows from the uniqueness of g. It therefore suffices to prove the existence and uniqueness of solutions for the system (1), (2), and (7) to obtain the existence and uniqueness of solutions of the full Moore–Greitzer model.

The variable g represents the nonaxisymmetric mass flow disturbance, i.e., the stall. We shall refer to g as the *stall variable.*

PROPOSITION 2.1. *Let* $Z = L^2$ *or* H^k*, for* $k = 1, 2, \ldots$. $K : Z \mapsto Z$ *is a bounded, self-adjoint, positive definite operator with a bounded inverse. One has* $\|K\|_Z = 1{+}am$*, and* $\|K^{-1}\|_Z = 1$*. Moreover,* K*,* $\frac{\partial}{\partial\xi}$*, and* $\frac{\partial}{\partial\theta}$ *commute.*

Proof. We have $\|Kg\|_Z \leq (1 + \frac{am}{1})\|g\|_Z$ so K is bounded with $\|K\|_Z = 1 + am$. Similarly, K^{-1} is bounded with $\|K^{-1}\|_Z = 1$. Furthermore, $\langle g, Kg\rangle \geq \|g\|_Z$, so K is positive definite and bounded away from zero. It is easy to see that K is self-adjoint.

On their domains the operators $\frac{\partial}{\partial \xi}$, $\frac{\partial}{\partial \theta}$, and K all commute. This is clear by letting them operate termwise on the Fourier series. □

For future reference, we note that the inverse of K is

$$K^{-1}\left(\sum_{p=1}^{\infty} A_p(\xi)\sin(p\theta + r_p(\xi))\right) := \sum_{p=1}^{\infty}\left(\frac{p}{p+am}\right)A_p(\xi)(\sin(p\theta + r_p(\xi))).$$

Using K we can define weighted L^2 and H^k norms as follows:

$$\begin{aligned} \|g\|_{L^2_K} &:= \sqrt{\langle g, Kg\rangle}, \\ \|g\|_{H^k_K} &:= \sqrt{\langle \tfrac{\partial^k g}{\partial \theta^k}, K\tfrac{\partial^k g}{\partial \theta^k}\rangle}. \end{aligned}$$

Note that L^2 and H^k norms are equivalent with their weighted counterparts. In fact, one has Lemma 2.2.

LEMMA 2.2. *We have that*

$$\|g\|_{L^2} \leq \|g\|_{L^2_K} \leq \sqrt{1+am}\|g\|_{L^2},$$

$$\|g\|_{H^k} \leq \|g\|_{H^k_K} \leq \sqrt{1+am}\|g\|_{H^k}.$$

We define

$$\overline{\Psi_c} := \frac{1}{2\pi}\int_0^{2\pi} \Psi_c(\Phi + g(\xi,\theta))d\theta.$$

Then we can rewrite the model (4), (1), and (2) as

$$\frac{\partial}{\partial \xi}g(\xi,\theta) = K^{-1}\left(a(\Psi_c(\Phi(\xi) + g(\xi,\theta)) - \overline{\Psi_c}) - \frac{1}{2}\frac{\partial g(\xi,\theta)}{\partial \theta}\right), \tag{9}$$

$$\frac{d\Phi}{d\xi} = \frac{1}{l_c}(\overline{\Psi_c} - \Psi(\xi)), \tag{10}$$

$$\frac{d\Psi}{d\xi} = \frac{1}{4l_cB^2}(\Phi(\xi) - K_T(\Psi, u)). \tag{11}$$

We will frequently use a formula for a difference of values of a C^1 function at two points.

LEMMA 2.3. *Let f be a C^1 function. Then*

$$f(x+\Delta x) - f(x) = \left(\int_0^1 f'(x + s\Delta x)ds\right)\Delta x.$$

3. Existence and uniqueness of solutions.

3.1. Moore–Greitzer model as an evolution equation on a Banach space. To prove the existence and uniqueness of solutions of the full Moore–Greitzer model we represent the model as an evolution equation of the form

$$\frac{dx}{d\xi} = Ax + f(x), \tag{12}$$

where x belongs to a Banach space X, A is an unbounded operator in X, and f is a nonlinear operator.

Let X be a Banach space. We define the following two spaces:

$$C(0,T;X) := \{x(\cdot) : [0,T] \to X \text{ is strongly continuous in } \|\cdot\|_X \text{ norm}\},$$

$$C^1(0,T;X) := \{x(\cdot) : [0,T] \to X \text{ is continuously differentiable in } \|\cdot\|_X \text{ norm}\}$$

with corresponding norms

$$\|x\|_{C(0,T;X)} = \sup_{\xi\in[0,T]} \|x(\xi)\|_X,$$

$$\|x\|_{C^1(0,T;X)} = \sup_{\xi\in[0,T]} \|x(\xi)\|_X + \sup_{\xi\in[0,T]} \|\frac{d}{d\xi}x(\xi)\|_X.$$

We are going to use the following corollary from Kato's theorem [4], [5].

Theorem 3.1. *Let X be a Banach space, and let A be a generator of a strongly continuous semigroup on X. Let Y be the domain of A. Suppose that $f(\cdot)$ satisfies the conditions*

$$\|f(x)\|_Y \le C_{bdd}(\|x\|_Y) \tag{13}$$

and

$$\|f(x_1) - f(x_2)\|_X \le C_{Lip}(\|x_1\|_X, \|x_2\|_X, \|Ax_1\|_X, \|Ax_2\|_X)\|x_1 - x_2\|_X, \tag{14}$$

where functions C_{bdd} and C_{Lip} are bounded on bounded sets. Then for all $x_0 \in Y$ there exists a unique local strong solution of

$$\frac{dx}{d\xi} = Ax + f(x) \tag{15}$$

such that

$$x \in C(0,\delta;Y) \cap C^1(0,\delta;X), \quad x(0) = x_0 \tag{16}$$

for some $\delta > 0$.

Proof. See Theorem 10 of [4]. □

Define, for $k = 0, 1, \ldots$, the spaces

$$X^k := H_K^k \times \mathbf{R}^2$$

(with $H_K^0 := L_K^2$). The norms on X^k are defined by

$$\|(g, \Phi, \Psi)\|_{X^k}^2 = \|g\|_{H_K^k}^2 + |\Phi|^2 + |\Psi|^2.$$

In this paper we will apply Theorem 3.1 with $X = X^{k-1}$ and $Y = X^k$.

3.2. Local existence and uniqueness. With suitable conditions on the compressor characteristic, Ψ_c, the local existence of X^k solutions becomes rather elementary. In an attempt to appeal to a larger audience, here we will present a detailed proof of the local existence and uniqueness of X^1 solutions. We then state a theorem which gives local existence and uniqueness in X^k and outline the proof.

We are going to apply Theorem 3.1 with

$$X = X^0 = L^2_K \times \mathbf{R}^2, \quad Y = X^1 = H^1_K \times \mathbf{R}^2,$$

with the norms

$$\|(g, \Phi, \Psi)\|^2_{X^1} = \|g\|^2_{H^1_K} + |\Phi|^2 + |\Psi|^2,$$

$$\|(g, \Phi, \Psi)\|^2_{X^0} = \|g\|^2_{L^2_K} + |\Phi|^2 + |\Psi|^2.$$

These spaces are both Hilbert spaces, X^1 is continuously embedded in X^0, and X^1 is dense in X^0. We now define the operator $A : X^1 \to X^0$ as follows:

$$A(g, \Phi, \Psi) := \left(-\frac{1}{2} K^{-1} \frac{\partial g}{\partial \theta}, 0, 0 \right). \tag{17}$$

This operator is closed and, as we will show, it is an infinitesimal generator of a strongly continuous unitary semigroup on X^0. We define

$$f(g, \Phi, \Psi) := (aK^{-1}(\Psi_c(\Phi + g) - \bar{\Psi_c}), \frac{1}{l_c}(\bar{\Psi_c} - \Psi), \frac{1}{4 l_c B^2}(\Phi - K_T(\Psi, u)). \tag{18}$$

Remark 3.1. Using a square throttle characteristic and constant throttle control u (cf. [15]) will cause K_T not to be Lipschitz on the hyperplane defined by $\Psi = 0$. However, if we use a feedback control of the form $u = u(g, \Phi, \Psi)$, $K_T(\Psi, u)$ (and hence also $f(g, \Phi, \Psi)$) becomes a function of all the state variables. We assume that the feedback was chosen such that $K_T(\Psi, u)$ is Lipschitz on bounded subsets of $\mathbf{R}$.

Having defined the function spaces, the operator A, and the nonlinear operator f, we can prove the local existence and uniqueness of solutions of the full Moore–Greitzer model. For this, we will show in the following two lemmas that A given by (17) and $f(g, \Phi, \Psi)$ given by (18) satisfy the conditions of Theorem 3.1.

LEMMA 3.1. *The operator A given by* (17) *is a generator of a strongly continuous unitary semigroup on X^0.*

Proof. Using the fact that K is self-adjoint and integration by parts one can prove that $A^* = -A$; i.e., A is a skew-adjoint operator. Thus, A generates a strongly continuous unitary semigroup on X^0 (cf. Theorem 8 of [4]). □

LEMMA 3.2. *Suppose that $\Psi_c \in C^1(\mathbf{R})$. We also assume that $K_T(\Psi, u)$ is bounded from X^1 to $\mathbf{R}$ and X^0-Lipschitz on X^1-bounded sets, i.e., for all $x_i = (g_i, \Phi_i, \Psi_i) \in X^1$, $i = 1, 2$, $K_T(\Psi, u)$ satisfies*

$$|K_T(\Psi_1, u(x_1)) - K_T(\Psi_2, u(x_2))| \leq C_K \|x_1 - x_2\|_{X^0},$$

where C_K is a function of X^1 norms of x_i, $i = 1, 2$, which is bounded on bounded sets in X^1. Then the function $f(g, \Phi, \Psi)$ given by (18) *satisfies conditions* (13) *and* (14) *of Theorem* 3.1.

Proof. Let $\mathcal{M}$ be a bounded subset of X^1 and let $(g, \Phi, \Psi) \in \mathcal{M}$ be arbitrary. Then, because of embedding $H^1 \hookrightarrow C^0$, for every θ, $\Phi + g(\theta)$ belongs to a bounded interval $I_{\mathcal{M}} \subset \mathbf{R}$. (Note that $I_{\mathcal{M}}$ depends only on $\mathcal{M}$.) Thus, $|\Psi_c'(\Phi+g)| \leq \sup_{\phi\in I_{\mathcal{M}}} |\Psi_c'(\phi)|$. Therefore,

$$\begin{aligned}\|aK^{-1}(\Psi_c(\Phi+g)-\bar{\Psi}_c)\|^2_{H^1_K} &\leq a^2\|K^{-1}\| \left\| \frac{\partial}{\partial\theta}\Psi_c(\Phi+g)\right\|^2_{L^2} \\ &\leq a^2\|K^{-1}\| \sup_{\phi\in I_{\mathcal{M}}} |\Psi_c'(\phi)|^2 \|g\|^2_{H^1} \\ &\leq a^2 \sup_{\phi\in I_{\mathcal{M}}} |\Psi_c'(\phi)|^2\|g\|^2_{H^1_K}. \end{aligned} \tag{19}$$

We also have

$$|\bar{\Psi}_c| = \left|\int_0^{2\pi} \Psi_c(\Phi+g)d\theta\right| \leq \sup_{\phi\in I_{\mathcal{M}}} \Psi_c(\phi). \tag{20}$$

Using (19) and (20), we easily show that $f(g,\Phi,\Psi)$ satisfies (13).

To show that $f(g,\Phi,\Psi)$ satisfies (14), let $x_1 = (g_1,\Phi_1,\Psi_1) \in \mathcal{M}$ and $x_2 = (g_2,\Phi_2,\Psi_2) \in \mathcal{M}$ be arbitrary. To simplify notation, let us denote $F_i := \Psi_c(\Phi_i + g_i)$ for $i = 1, 2$. Recall that $\overline{F_i}$ denotes the average value of F_i. We have

$$\begin{aligned}&\|f(x_1)-f(x_2)\|^2_{X^0} \\ &= a^2\langle K^{-1}((F_1-F_2)-\overline{(F_1-F_2)}),(F_1-F_2)-\overline{(F_1-F_2)}\rangle \\ &\quad + \frac{1}{l_c^2}|\overline{(F_1-F_2)}-(\Psi_1-\Psi_2)|^2 \\ &\quad + \frac{1}{(4l_cB^2)^2}|(\Phi_1-\Phi_2)-(K_T(\Psi_1,u(x_1))-K_T(\Psi_2,u(x_2)))|^2 \\ &\leq a^2\|K^{-1}\|_{L^2}\|(F_1-F_2)-\overline{(F_1-F_2)}\|^2_{L^2} \\ &\quad + \frac{2}{l_c^2}|\overline{(F_1-F_2)}|^2 + \frac{2}{l_c^2}|(\Psi_1-\Psi_2)|^2 \\ &\quad + \frac{2}{(4l_cB^2)^2}|(\Phi_1-\Phi_2)|^2 + \frac{2C_K^2}{(4l_cB^2)^2}\|x_1-x_2\|^2_{X^0}.\end{aligned}$$

Note that $\|K^{-1}\|_{L^2} = 1$ and

$$\|F_1-F_2\|^2_{\mathcal{L}^2} = \|(F_1-F_2)-\overline{(F_1-F_2)}\|^2_{L^2} + 2\pi|\overline{(F_1-F_2)}|^2.$$

Hence,

$$\begin{aligned}&\|f(x_1)-f(x_2)\|^2_{X^0} \\ &\leq C_1\|F_1-F_2\|^2_{\mathcal{L}^2} + C_2\|x_1-x_2\|^2_{X^0},\end{aligned}$$

where $C_1 := a^2 + \frac{2}{l_c^2}$ and $C_2 := \frac{2}{l_c^2} + \frac{2}{(4l_cB^2)^2} + \frac{2C_K^2}{(4l_cB^2)^2}$. We will show that

$$\|(F_1-F_2)\|^2_{\mathcal{L}^2} \leq C_3\|x_1-x_2\|^2_{X^0}, \tag{21}$$

where C_3 is a function of X^1 norms of (g_i,Φ_i,Ψ_i), $i = 1, 2$, which is bounded on bounded sets in X^1. For this, note that by Lemma 2.3

$$\|F_1-F_2\|^2_{\mathcal{L}^2} = \left\|\left(\int_0^1 \Psi_c'(\Phi_2+g_2+s(\Phi_1+g_1-\Phi_2-g_2))ds\right)(\Phi_1-\Phi_2+g_1-g_2)\right\|^2_{\mathcal{L}^2}$$

$$\leq 2 \sup_{\phi \in I_{\mathcal{M}}} |\Psi_c'(\phi)|^2 (|\Phi_1 - \Phi_2|^2 + \|g_1 - g_2\|^2_{L^2_K})$$

$$\leq 2 \sup_{\phi \in I_{\mathcal{M}}} |\Psi_c'(\phi)|^2 \|x_1 - x_2\|^2_{X^0}.$$

Therefore, (21) holds with $C_3 := 2 \sup_{\phi \in I_{\mathcal{M}}} |\Psi_c'(\phi)|^2$. Note that C_3 is bounded on bounded sets in X^1. Thus,

$$\|f(x_1) - f(x_2)\|^2_{X^0} \leq C_4 \|x_1 - x_2\|^2_{X^0},$$

where $C_4 := C_1 C_3 + C_2$. Note that C_4 is bounded on bounded sets in X^1. Therefore, $f(g, \Phi, \Psi)$ satisfies (14). □

Therefore, we can state the following result.

Theorem 3.2. *Assume that Ψ_c is a C^1 function. Then the Cauchy problem*

$$\frac{dx}{d\xi} = Ax + f(x, u), \quad x(0) = x_0 \in X^1 \tag{22}$$

has a unique solution $x \in C(0, \delta; X^1) \cap C^1(0, \delta; X^0)$, such that $x(0) = x_0$, for sufficiently small δ (depending on x_0).

We now state a theorem which gives the local existence and uniqueness of X^k solutions for $k = 1, 2, \ldots$.

Theorem 3.3. *Suppose that $\Psi_c \in C^{k+1}(\mathbf{R})$. We assume that $K_T(\Psi, u)$ is bounded from X^k to $\mathbf{R}$ and X^{k-1}-Lipschitz on X^K-bounded sets, i.e., for all $x_i = (g_i, \Phi_i, \Psi_i) \in X^K$, $i = 1, 2$, $K_T(\Psi, u)$ satisfies*

$$|K_T(\Psi_1, u(x_1)) - K_T(\Psi_2, u(x_2))| \leq C_K \|x_1 - x_2\|_{X^{k-1}},$$

where C_K is a function of X^k norms of x_i, $i = 1, 2$, which is bounded on bounded sets in X^k.

Then the Cauchy problem

$$\frac{dx}{d\xi} = Ax + f(x, u), \quad x(0) = x_0 \in X^k$$

has a unique solution $x \in C(0, \delta; X^k) \cap C^1(0, \delta; X^{k-1})$, such that $x(0) = x_0$, for sufficiently small δ (depending on x_0).

Proof. We only outline the proof.

Since $\Psi_c \in C^{k+1}$ and the underlying space has only one dimension, it follows from the Sobolev embedding theorem that for $\Phi \in \mathbf{R}$ we have that the mapping

$$\Psi_c(\cdot) - \overline{\Psi_c(\cdot)} : \mathcal{H}^k \to H^k$$

is C^1 for $k > \frac{1}{2}$. (See McOwen, [13, p. 221].) (Here, $\mathcal{H}^k$ denotes the usual Sobolev space on the unit circle.) In particular, this mapping is locally X^{k-1}-Lipschitz X^k-bounded sets and because of

$$\begin{aligned} \|f(x)\|^2_{X^k} &\leq \|f(0)\|^2_{X^k} + \|f(x) - f(0)\|^2_{X^k}, \\ &\leq \|f(0)\|^2_{X^k} + C_L \|x\|^2_{X^k} \end{aligned}$$

it is also bounded on bounded sets. The result follows. □

3.3. A priori estimates for X^1 solutions. We have Proposition 3.1.

PROPOSITION 3.1. *Assume that Ψ_c is a C^2 function. Let X^1, X^0, f, and A be as in section* 3.2, *and let $x = (g, \Phi, \Psi) \in C(0, \delta; X^1) \cap C^1(0, \delta; X^0)$ be a solution to* (22) *for some $\delta > 0$. Then*

$$\frac{d}{d\xi}\frac{1}{2}\|g\|^2_{H^1_K} = a\int_0^{2\pi} \Psi_c'(\Phi + g)\left(\frac{\partial g}{\partial \theta}\right)^2 d\theta. \tag{23}$$

Proof. In the proof we will deal with the expression $\frac{\partial^2 g}{\partial \theta^2}$ which is not in L^2 for all X^1 functions. Therefore, we first need to prove that (23) holds on a dense subset of X^1 solutions of (22) for which $\frac{\partial^2 g}{\partial \theta^2}$ makes sense. For this subset we choose X^2 solutions of (22).

Assume that $x = (g, \Phi, \Psi) \in C(0, \delta; X^2) \cap C^1(0, \delta; X^1)$ is a solution to (22) for some $\delta > 0$. Then

$$\frac{d}{d\xi}\frac{1}{2}\|g\|^2_{H^1_K} = \frac{1}{2}\left(\left\langle \frac{d}{d\xi}\frac{\partial g}{\partial \theta}, K\frac{\partial g}{\partial \theta}\right\rangle + \left\langle \frac{\partial g}{\partial \theta}, \frac{d}{d\xi}K\frac{\partial g}{\partial \theta}\right\rangle\right).$$

Since K is self-adjoint and $\frac{d}{d\xi}$, K, and $\frac{\partial}{\partial \theta}$ commute, we have

$$\frac{d}{d\xi}\frac{1}{2}\|g\|^2_{H^1_K} = \left\langle \frac{\partial g}{\partial \theta}, \frac{\partial}{\partial \theta}K\frac{d}{d\xi}g\right\rangle.$$

Thus we have by (9),

$$\frac{d}{d\xi}\frac{1}{2}\|g\|^2_{H^1_K} = a\left\langle \frac{\partial \Psi_c(\Phi + g)}{\partial \theta}, \frac{\partial g}{\partial \theta}\right\rangle - a\left\langle \frac{\partial \overline{\Psi_c}}{\partial \theta}, \frac{\partial g}{\partial \theta}\right\rangle - \frac{1}{2}\left\langle \frac{\partial^2 g}{\partial \theta^2}, \frac{\partial g}{\partial \theta}\right\rangle.$$

One has $\frac{\partial \overline{\Psi_c}}{\partial \theta} = 0$. Moreover,

$$\frac{1}{2}\left\langle \frac{\partial^2 g}{\partial \theta^2}, \frac{\partial g}{\partial \theta}\right\rangle = \int_0^{2\pi} \frac{\partial}{\partial \theta}\left(\frac{\partial g}{\partial \theta}\right)^2 d\theta = 0.$$

Thus,

$$\begin{array}{rcl} \frac{d}{d\xi}\frac{1}{2}\|g\|^2_{H^1_K} & = & a\langle \frac{\partial \Psi_c(\Phi+g)}{\partial \theta}, \frac{\partial g}{\partial \theta}\rangle \\ & = & a\int_0^{2\pi} \Psi_c'(\Phi + g)(\frac{\partial g}{\partial \theta})^2 d\theta \end{array}$$

for $x \in X^2$. Since X^2 local solutions of (22) are dense in the set of X^1 local solutions of (22), if we can show that the right-hand side of (23) is X^1 continuous, then (23) will hold for all X^1 solutions of (22). Let $\mathcal{M}$ be a bounded subset of X^1 and let $x_1 = (g_1, \Phi_1, \Psi_1) \in \mathcal{M}$ and $x_2 = (g_2, \Phi_2, \Psi_2) \in \mathcal{M}$ be arbitrary. Then, by Lemma 2.1, for every θ, $\Phi_1 + g_1(\theta)$ and $\Phi_2 + g_2(\theta)$ belong to a bounded interval $I_{\mathcal{M}} \subset \mathbf{R}$. Therefore, using Lemma 2.3 one obtains

$$\begin{aligned} &|\Psi_c'(\Phi_1 + g_1) - \Psi_c'(\Phi_2 + g_2)| \\ &\qquad \leq |\textstyle\int_0^1 \Psi_c''(\Phi_2 + g_2 + s(\Phi_1 - \Phi_2 + g_1 - g_2))ds(\Phi_1 - \Phi_2 + g_1 - g_2)| \\ &\qquad \leq \sup_{\phi \in I_{\mathcal{M}}} |\Psi_c''(\phi)|(|\Phi_1 - \Phi_2| + \|g_1 - g_2\|_{L^\infty}) \\ &\qquad \leq \sup_{\phi \in I_{\mathcal{M}}} |\Psi_c''(\phi)|(|\Phi_1 - \Phi_2| + \tfrac{\sqrt{\pi}}{\sqrt{6}}\|g_1 - g_2\|_{H^1_K}) \\ &\qquad \leq \sup_{\phi \in I_{\mathcal{M}}} |\Psi_c''(\phi)|\|x_1 - x_2\|_{X^1}. \end{aligned}$$

We can now use this to show that the right-hand side of (23) is continuous in X^1:

$$\begin{aligned} &\left|\textstyle\int_0^{2\pi}\Psi_c'(\Phi_1+g_1)(\frac{\partial g_1}{\partial\theta})^2d\theta-\int_0^{2\pi}\Psi_c'(\Phi_2+g_2)(\frac{\partial g_2}{\partial\theta})^2d\theta\right|\\ \leq\;&\textstyle\int_0^{2\pi}\left|\Psi_c'(\Phi_1+g_1)\left((\frac{\partial g_1}{\partial\theta})^2-(\frac{\partial g_2}{\partial\theta})^2\right)\right|+|\Psi_c'(\Phi_1+g_1)-\Psi_c'(\Phi_2+g_2)|\,(\frac{\partial g_2}{\partial\theta})^2d\theta\\ \leq\;&\sup_{\phi\in I_{\mathcal{M}}}|\Psi_c'(\phi)|(\|g_1\|_{H^1}^2-\|g_2\|_{H^1}^2)+\sup_{\phi\in I_{\mathcal{M}}}|\Psi_c''(\phi)|\|x_1-x_2\|_{X^1}\|g_2\|_{H^1}\\ \leq\;&(\sup_{\phi\in I_{\mathcal{M}}}|\Psi_c'(\phi)|(\|g_1\|_{H^1_K}+\|g_2\|_{H^1_K})+\sup_{\phi\in I_{\mathcal{M}}}|\Psi_c''(\phi)|\|g_2\|_{H^1_K})\|x_1-x_2\|_{X^1}. \end{aligned}$$

Thus (23) holds for all X^1 solutions of (22). □

COROLLARY 3.1. *H^1_K solutions of* (9) *grow at most exponentially; i.e., there is no finite-time blow-up of H^1_K solutions of* (9). *In particular, one has*

$$\frac{d}{d\xi}\|g\|^2_{H^1_K}\leq a\sup_{\phi\in\mathbf{R}}\Psi_c'(\phi)\|g\|^2_{H^1_K} \tag{24}$$

and

$$\frac{d}{d\xi}\|g\|_{H^1_K}\leq a\sup_{\phi\in\mathbf{R}}\Psi_c'(\phi)\|g\|_{H^1_K}. \tag{25}$$

Proof. Observe that it follows from our assumptions that Ψ_c' is bounded from above. Now, using Proposition 3.1 and Lemma 2.2 one obtains

$$\frac{d}{d\xi}\|g\|^2_{H^1_K}\leq a\sup_{\phi\in\mathbf{R}}\Psi_c'(\phi)\|g\|^2_{H^1_K}.$$

Observe that $\frac{\partial}{\partial\xi}\frac{1}{2}\|g\|^2_{H^1_K}=\|g\|_{H^1_K}\frac{\partial}{\partial\xi}\|g\|_{H^1_K}$ so upon dividing (24) by $\|g\|_{H^1_K}$ we get (25).

By Grönwall's lemma we get that the solutions grow at most exponentially. □

3.4. Global existence and uniqueness of X^1 solutions. In section 4 we will construct a globally stabilizing feedback control u for the system (9), (10), and (11). As a consequence, the *global* existence of solutions of the system (9), (10), and (11) will be established. The main condition on characteristic Ψ_c is

$$\sup_{\phi\in\mathbf{R}}\Psi_c'(\phi)<+\infty;$$

i.e., the *positive* slopes of the characteristic are bounded. Note that this condition follows from our assumptions about the characteristic stated in the beginning of section 2.1.

THEOREM 3.4. *Assume that X^1, X^0, f, and A are as before. Assume that*

$$\sup_{\phi\in\mathbf{R}}\Psi_c'(\phi)<\infty$$

and there exist constants N_1, N_2 such that $|K_T(\Psi,u(x))|\leq N_1+N_2\|x\|_{X^1}$. Then for any $T>0$ the Cauchy problem

$$\frac{dx}{d\xi}=Ax+f(x),\quad x(0)=x_0\in X^1 \tag{26}$$

has a unique solution $x\in C(0,T;X^1)\cap C^1(0,T;X^0)$ such that $x(0)=x_0$.

Proof. Since the derivative of the characteristic is bounded from above, there exist positive constants L_1, L_2 such that $\Psi_c(\phi) > -L_1 - L_2|\phi|$ when $\phi < 0$ and $\Psi_c(\phi) < L_1 + L_2\phi$ for $\phi > 0$. We now get

$$\begin{array}{rcl}
\Phi\overline{\Psi_c} & = & \frac{1}{2\pi}\int_0^{2\pi} \Phi\Psi_c(\Phi+g)d\theta \\
& \leq & |\Phi|(L_1 + L_2(|\Phi| + \|g\|_{L^\infty})) \\
& \leq & L_1(1+|\Phi|^2) + L_2|\Phi|^2 + L_2|\Phi|\|g\|_{L^\infty} \\
& \leq & L_1(1+|\Phi|^2) + L_2|\Phi|^2 + L_2\frac{\sqrt{\pi}}{\sqrt{6}}\frac{1}{2}(\|g\|^2_{H^1} + |\Phi|^2).
\end{array}$$

Therefore,

$$\Phi\overline{\Psi_c} \leq L_1 + \left(L_1 + L_2\left(1 + \frac{\sqrt{\pi}}{\sqrt{6}}\frac{1}{2}\right)\right)\|x\|^2_{X^1}. \tag{27}$$

Now by Corollary 3.1 we have

$$\begin{array}{rcl}
\frac{d}{d\xi}\frac{1}{2}\|x\|^2_{X^1} & \leq & a\sup_{\mathbf{R}}\Psi'_c\|g\|^2_{H^1_K} + \Phi\frac{d}{d\xi}\Phi + \Psi\frac{d}{d\xi}\Psi \\
& = & a\sup_{\phi\in\mathbf{R}}\Psi'_c(\phi)\|g\|^2_{H^1_K} + \frac{1}{l_c}\Phi\overline{\Psi_c} - \frac{1}{l_c}\Phi\Psi + \frac{1}{4l_cB^2}\Phi\Psi \\
& - & \frac{1}{4l_cB^2}\Psi K_T(\Psi, u(x)) \\
& \leq & a\sup_{\phi\in\mathbf{R}}\Psi'_c(\phi)\|g\|^2_{H^1_K} + L_1 + \left(L_1 + L_2(1 + \frac{\sqrt{\pi}}{\sqrt{6}}\frac{1}{2})\right)\|x\|^2_{X^1} \\
& + & (\frac{1}{l_c} + \frac{1}{4l_cB^2})\frac{1}{2}(|\Phi|^2 + |\Psi|^2) + \frac{1}{4l_cB^2}|\Psi K_T(\Psi, u(x))| \\
& \leq & a\sup_{\phi\in\mathbf{R}}\Psi'_c(\phi)\|g\|^2_{H^1_K} + L_1 + \left(L_1 + L_2(1 + \frac{\sqrt{\pi}}{\sqrt{6}}\frac{1}{2})\right)\|x\|^2_{X^1} \\
& + & (\frac{1}{l_c} + \frac{1}{4l_cB^2})\frac{1}{2}(|\Phi|^2 + |\Psi|^2) + \frac{1}{4l_cB^2}(N_1 + (N_1+N_2)\|x\|^2_{X^1}).
\end{array}$$

Therefore, we obtain

$$\frac{d}{d\xi}\frac{1}{2}\|x\|^2_{X^1} \leq C_1 + C_2\|x\|^2_{X^1}. \tag{28}$$

Here

$$\begin{array}{rcl}
C_1 & = & \frac{1}{4l_cB^2}N_1 + L_1, \\
C_2 & = & a\sup_{\phi\in\mathbf{R}}\Psi'_c(\phi) + L_1 + L_2(1 + \frac{\sqrt{\pi}}{\sqrt{6}}\frac{1}{2}) + \left(\frac{1}{l_c} + \frac{1}{4l_cB^2}\right)\frac{1}{2} + \frac{N_1+N_2}{4l_cB^2}.
\end{array}$$

By Grönwall's lemma we now get

$$\|x\|^2_{X^1}(\xi) \leq \left(\|x\|^2_{X^1}(0) + \frac{C_1}{C_2}\right)e^{C_2\xi} - \frac{C_1}{C_2}. \tag{29}$$

We therefore see that solutions of (22) are bounded for all finite times and thus we have a global solution. ◻

Because of the embedding $H^1 \hookrightarrow L^\infty$, this means that L^∞ norms of H^1 solutions do not blow up in finite time either.

Global existence and uniqueness of H^1_K solutions of the full Moore–Greitzer model and the corresponding a priori estimates allow us to construct a controller that globally stabilizes the peak or any axisymmetric equilibrium to the right of the peak. The controller will have a similar form to a controller for an MG3 model with the H^1_K norm of the stall variable replacing the magnitude of the first Fourier mode of the stall cell.

4. H^1 backstepping. We are going to construct a feedback controller stabilizing the peak or any axisymmetric equilibrium to the right of the peak of the characteristic for the full Moore–Greitzer model. The feedback is constructed by the following backstepping procedure. In the first step we define a positive definite function $V_1(g)$ and construct a function $\widehat{\Phi}(\|g\|)$ such that for $\Phi = \widehat{\Phi}(\|g\|)$ $V_1(g)$ is a Lyapunov function for (9). $V_1(g)$ is called a *control Lyapunov function* and $\widehat{\Phi}(\|g\|)$ is called a *virtual control* for (9). In the second step we define a control Lyapunov function $V_2(\Phi, g)$ and a virtual control $\Psi = \widehat{\Psi}(\|g\|, \Phi)$ for (9) and (10). In the third (and last) step we construct the control Lyapunov function for the full system (9), (10), and (11) with the throttle function u being the control variable. We will refer to this procedure as H^1 backstepping. The obtained feedback control law u uses the H^1_K norm of a stall cell and resembles familiar control laws for MG3 (see [3], [8], [9], [10], [16], [1]) with A_1 replaced with the H^1_K norm of g. In terms of the Fourier coefficients A_i of g, this norm is $(\sum_{p=1}^{\infty}(1+\frac{am}{i})(iA_i)^2)^{\frac{1}{2}}$. To simplify notation, let us from now on denote the norm $\|\cdot\|_{H^1_K}$ by $\|\cdot\|$.

4.1. H^1 backstepping: Step 1. As a control Lyapunov function for (9) we will use the H^1_K norm of g. Let

$$V_1(g) := \frac{1}{2}\|g\|^2.$$

We will show that $\frac{d}{d\xi}V_1(g)$ can be made negative definite by a virtual control of the form

$$\Phi = \widehat{\Phi}(\|g\|) = \Gamma + \overline{c}_g\|g\| \tag{30}$$

for $\Gamma \geq \Phi_0$ and sufficiently large positive $\overline{c}_g$. (Φ_0 denotes the value of the mass flow coefficient at the peak.) In this paper we assume that $\overline{c}_g \geq 0$. We will need the following result.

LEMMA 4.1. *For every $\theta \in [0, 2\pi]$,*

$$\left(\overline{c}_g - \frac{\sqrt{\pi}}{\sqrt{6}}\right)\|g\| \leq \overline{c}_g\|g\| + g \leq \left(\overline{c}_g + \frac{\sqrt{\pi}}{\sqrt{6}}\right)\|g\|.$$

Proof. Note that $\overline{c}_g\|g\| + g \geq \overline{c}_g\|g\| - \|g\|_{L^\infty} \geq (\overline{c}_g - \frac{\sqrt{\pi}}{\sqrt{6}})\|g\|$ by Lemmas 2.1 and 2.2. □

Let

$$e_\Phi := \Phi - \widehat{\Phi}(\|g\|) = \Phi - \Gamma - \overline{c}_g\|g\|.$$

It follows from Proposition 3.1 and Lemma 2.3 that one can represent $\frac{d}{d\xi}V_1(g)$ as

$$\begin{aligned}
\tfrac{d}{d\xi}V_1(g) &= a\textstyle\int_0^{2\pi}\Psi_c'(\widehat{\Phi}(\|g\|) + g)(\frac{\partial g}{\partial\theta})^2 d\theta \\
&+ a(\textstyle\int_0^{2\pi}(\int_0^1 \Psi_c''(\widehat{\Phi}(\|g\|) + se_\Phi + g)ds)(\frac{\partial g}{\partial\theta})^2 d\theta)e_\Phi.
\end{aligned}$$

Using Lemma 2.3 again, one obtains

$$\begin{aligned}
\tfrac{d}{d\xi}V_1(g) &= a\textstyle\int_0^{2\pi}(\Psi_c'(\Gamma) + \int_0^1 \Psi_c''(\Gamma + s(\overline{c}_g\|g\| + g))ds(\overline{c}_g\|g\| + g))(\frac{\partial g}{\partial\theta})^2 d\theta \\
&+ a(\textstyle\int_0^{2\pi}(\int_0^1 \Psi_c''(\widehat{\Phi}(\|g\|) + se_\Phi + g)ds)(\frac{\partial g}{\partial\theta})^2 d\theta)e_\Phi \\
&= a\Psi_c'(\Gamma)\|g\|^2_{H^1} \\
&+ a\textstyle\int_0^{2\pi}(\int_0^1 \Psi_c''(\Gamma + s(\overline{c}_g\|g\| + g))ds(\overline{c}_g\|g\| + g))(\frac{\partial g}{\partial\theta})^2 d\theta \\
&+ a(\textstyle\int_0^{2\pi}(\int_0^1 \Psi_c''(\widehat{\Phi}(\|g\|) + se_\Phi + g)ds)(\frac{\partial g}{\partial\theta})^2 d\theta)e_\Phi.
\end{aligned}$$

Note that if $\Gamma > \Phi_{infl}$ then $\int_0^1 \Psi_c''(\Gamma + s(\overline{c}_g\|g\| + g))ds$ can be bounded from above by a negative constant that depends only on Γ. Namely,

$$\int_0^1 \Psi_c''(\Gamma + s(\overline{c}_g\|g\| + g))ds \le \sup_{\Gamma \le \phi} \Psi_c''(\phi) < 0. \tag{31}$$

Define

$$c_1 := \frac{a}{1+am}(\overline{c}_g - \frac{\sqrt{\pi}}{\sqrt{6}}) \sup_{\Gamma \le \phi} \Psi_c''(\phi),$$

$$c_2(\Phi, \|g\|) := a \sup_{\min(\widehat{\Phi}(\|g\|),\Phi) - \|g\|_{L^\infty} \le \phi \le \max(\widehat{\Phi}(\|g\|),\Phi) + \|g\|_{L^\infty}} |\Psi_c''(\phi)|.$$

PROPOSITION 4.1. *Assume that* $\Gamma \ge \Phi_0$ *and* $\overline{c}_g > \frac{\sqrt{\pi}}{\sqrt{6}}$. *Then* $\Psi_c'(\Gamma) \le 0$, $c_1 < 0$,

$$\begin{aligned} \tfrac{d}{d\xi}V_1(g) &\le (\tfrac{a}{1+am}\Psi_c'(\Gamma) + c_1\|g\|)\|g\|^2 \\ &+ c_2(\Phi, \|g\|)\|g\|^2|e_\Phi|, \end{aligned} \tag{32}$$

and

$$\begin{aligned} \tfrac{d}{d\xi}\|g\| &\le (\tfrac{a}{1+am}\Psi_c'(\Gamma) + c_1\|g\|)\|g\| \\ &+ c_2(\Phi, \|g\|)\|g\||e_\Phi|. \end{aligned} \tag{33}$$

Proof. Note that it follows from Lemma 4.1 that $\overline{c}_g\|g\| + g \ge (\overline{c}_g - \frac{\sqrt{\pi}}{\sqrt{6}})\|g\|$. Moreover, since $\Gamma > \Phi_{infl}$, (31) holds. Therefore,

$$\begin{aligned} \tfrac{d}{d\xi}V_1(g) &\le (a\Psi_c'(\Gamma) + a(\overline{c}_g - \tfrac{\sqrt{\pi}}{\sqrt{6}})\sup\nolimits_{\Gamma \le \phi} \Psi_c''(\phi)\|g\|)\|g\|_{H^1}^2 \\ &+ c_2(\Phi, \|g\|)\|g\|_{H^1}^2|e_\Phi|. \end{aligned}$$

The first term is nonpositive; the last one is positive. Therefore, the inequality (32) follows from Lemma 2.2. Now the inequality (33) follows from (32). □

Proposition 4.1 is the most important result. It allows us to carry out the backstepping procedure for an infinite-dimensional system (9), (10), and (11) without the necessity of working with its infinite-dimensional part (9). What we have done here is a *replacement of an infinite-dimensional evolution equation* (9) with two *finite-dimensional differential inequalities* (32) and (33). As we shall see this replacement makes the next two backsteps quite standard.

In particular, for $\Phi = \widehat{\Phi}(\|g\|)$ one obtains

$$\frac{d}{d\xi}V_1(g) \le \left(\frac{a}{1+am}\Psi_c'(\Gamma) + c_1\|g\|\right)\|g\|^2.$$

If $\Gamma \ge \Phi_0$ then $\Psi_c'(\Gamma) \le 0$, $c_1 < 0$, and hence for $\Phi = \widehat{\Phi}(\|g\|)$ $\frac{d}{d\xi}V_1(g)$ is negative definite. For $\|g\|$ small, if $\Gamma > \Phi_0$ then $\Psi_c'(\Gamma) < 0$ and $\frac{d}{d\xi}V_1(g)$ depends quadratically on $\|g\|$, whereas for $\Gamma = \Phi_0$ one has $\Psi_c'(\Gamma) = 0$ and therefore the dependence of $\frac{d}{d\xi}V_1(g)$ on $\|g\|$ is cubic.

Remark 4.1. Note that the essential property of the virtual control (30) that allows us to make the time derivative of the control Lyapunov function negative was the ability of moving the stall cell "over the top," so that the whole mass flow $\widehat{\Phi}(\|g\|)+g$ is to the right of the peak, where the slope of the characteristic is negative. The Sobolev embedding was used to guarantee that property. A natural question is why did we not use the L^∞ norm of the stall cell or its minimum value instead of the H^1_K norm. The reason is that in the next step of the backstepping procedure we will need a bound on the time derivative of whatever norm of the stall cell we use in the first step. We have such information about the time derivative of the H^1_K norm, but we do not yet have the information about the L^∞ norm of the stall cell. We are currently working on the design that uses L^∞ norm of the stall cell or its minimum in the first step of backstepping.

4.2. $\boldsymbol{H_1}$ backstepping: Step 2. As a control Lyapunov function for (9) and (10) we will use

$$V_2(\Phi, g) := \frac{1}{2}\|g\|^2 + \frac{1}{2}e_\Phi^2.$$

We will show that $\frac{d}{d\xi}V_2(\Phi, g)$ can be made negative definite by a virtual control of the form

$$\begin{aligned} \Psi &= \widehat{\Psi}(\|g\|, \Phi) \\ &= \Psi_c(\Gamma) + \overline{c}_\Phi(\|g\|, \Phi)e_\Phi \end{aligned}$$

for sufficiently large $\overline{c}_\Phi(\|g\|, \Phi)$. (For semiglobal stabilization $\overline{c}_\Phi(\|g\|, \Phi)$ can be chosen to be a constant depending on the desired region of operation.)

Let

$$\begin{aligned} e_\Psi &:= \Psi - \widehat{\Psi}(\|g\|, \Phi) \\ &= \Psi - \Psi_c(\Gamma) - \overline{c}_\Phi(\|g\|, \Phi)e_\Phi. \end{aligned}$$

To calculate $\frac{d}{d\xi}V_2(\Phi, g)$ we will need to express $\frac{d}{d\xi}e_\Phi$ in terms of e_Φ, e_Ψ, and g. For this, note that

$$\begin{aligned} \tfrac{d}{d\xi}e_\Phi &= \tfrac{d}{d\xi}\Phi - \tfrac{d}{d\xi}\widehat{\Phi}(\|g\|) \\ &= \tfrac{1}{l_c}(\overline{\Psi_c(\Phi+g)} - \Psi) - \overline{c}_g\tfrac{d}{d\xi}\|g\|. \end{aligned}$$

Applying Lemma 2.3 twice, one obtains

$$\begin{aligned} \tfrac{d}{d\xi}e_\Phi &= \tfrac{1}{l_c}(\Psi_c(\Gamma) + \textstyle\int_0^{2\pi}(\int_0^1 \Psi_c'(\Gamma + s(\overline{c}_g\|g\| + g))ds)(\overline{c}_g\|g\| + g)d\theta \\ &+ (\textstyle\int_0^{2\pi}(\int_0^1 \Psi_c'(\widehat{\Phi}(\|g\|) + se_\Phi + g)ds)d\theta)e_\Phi \\ &- \widehat{\Psi}(\|g\|, \Phi) - e_\Psi) - \overline{c}_g\tfrac{d}{d\xi}\|g\| \\ &= \tfrac{1}{l_c}(\textstyle\int_0^{2\pi}(\int_0^1 \Psi_c'(\Gamma + s(\overline{c}_g\|g\| + g))ds)(\overline{c}_g\|g\| + g)d\theta \\ &+ (\textstyle\int_0^{2\pi}(\int_0^1 \Psi_c'(\widehat{\Phi}(\|g\|) + se_\Phi + g)ds)d\theta)e_\Phi \\ &- \overline{c}_\Phi(\|g\|, \Phi)e_\Phi - e_\Psi) - \overline{c}_g\tfrac{d}{d\xi}\|g\|. \end{aligned}$$

To simplify calculations, we introduce the following notation:

$$\begin{aligned} &c_0 := \tfrac{a}{1+am}\Psi_c'(\Gamma), \\ &c_3(g) := |(\overline{c}_g + 1)\textstyle\int_0^{2\pi}\int_0^1 \Psi_c'(\Gamma + s(\overline{c}_g\|g\| + g))dsd\theta|, \\ &c_4(\Phi, g) := \textstyle\int_0^{2\pi}\int_0^1 \Psi_c'(\widehat{\Phi}(\|g\|) + se_\Phi + g)dsd\theta. \end{aligned}$$

Note that $c_3(g)$ and $c_4(\Phi, g)$ can be bounded by functions of $\|g\|$. For clarity, in the following calculations we use notation c_i instead of $c_i(g)$, etc.

Lemma 4.2. *Assume that* $\Gamma \geq \Phi_0$ *and* $\overline{c}_g > \frac{\sqrt{\pi}}{\sqrt{6}}$. *Then*

$$\begin{array}{rcl} |\frac{d}{d\xi}e_\Phi| & \leq & \frac{1}{l_c}(c_3\|g\| + (|c_4 - \overline{c}_\Phi|)|e_\Phi| + |e_\Psi|) \\ & + & \overline{c}_g(|(c_0 + c_1\|g\||)\|g\| + c_2\|g\||e_\Phi|). \end{array}$$

and

$$\begin{array}{rcl} e_\Phi\frac{d}{d\xi}e_\Phi & \leq & \frac{1}{l_c}(c_3\|g\||e_\Phi| + (c_4 - \overline{c}_\Phi)e_\Phi^2 + |e_\Phi e_\Psi|) \\ & + & \overline{c}_g(|(c_0 + c_1\|g\|)|\|g\||e_\Phi| + c_2\|g\|e_\Phi^2). \end{array}$$

Therefore, assuming $\Gamma \geq \Phi_0$ and $\overline{c}_g > \frac{\sqrt{\pi}}{\sqrt{6}}$ one has

$$\begin{array}{rcl} \frac{d}{d\xi}V_2(\Phi, g) & = & \frac{d}{d\xi}V_1(g) + e_\Phi\frac{d}{d\xi}e_\Phi \\ & \leq & (c_0 + c_1\|g\|)\|g\|^2 + c_2\|g\|^2|e_\Phi| \\ & + & \frac{1}{l_c}(c_3\|g\||e_\Phi| + (c_4 - \overline{c}_\Phi)e_\Phi^2 + |e_\Phi e_\Psi|) \\ & + & \overline{c}_g(|(c_0 + c_1\|g\|)|\|g\||e_\Phi| + c_2\|g\|e_\Phi^2) \\ & \leq & (c_0 + c_1\|g\|)\|g\|^2 \\ & + & (c_2\|g\| + \frac{1}{l_c}c_3 + \overline{c}_g|(c_0 + c_1\|g\|)|)\|g\||e_\Phi| \\ & + & (\frac{1}{l_c}(c_4 - \overline{c}_\Phi) + \overline{c}_g c_2\|g\|)e_\Phi^2 + \frac{1}{l_c}|e_\Phi e_\Psi|. \end{array}$$

Define

$$c_5 := (\overline{c}_g + 1)|\Psi_c'(\Gamma)|,$$
$$c_6(g) := (\overline{c}_g + 1)|\int_0^{2\pi}\int_0^1\int_0^1 \Psi_c''(\Gamma + s_1 s_2(\overline{c}_g\|g\| + g))ds_1 ds_2 d\theta|.$$

Note that $c_6(g)$ can be bounded by functions of $\|g\|$. It follows from Lemma 2.3 that

$$c_3 \leq c_5 + c_6\|g\|.$$

Therefore, one obtains

$$\begin{array}{rcl} \frac{d}{d\xi}V_2(\Phi, g) & \leq & (c_0 + c_1\|g\|)\|g\|^2 \\ & + & (c_2\|g\| + \frac{1}{l_c}(c_5 + c_6\|g\|) + \overline{c}_g|(c_0 + c_1\|g\|)|)\|g\||e_\Phi| \\ & + & (\frac{1}{l_c}c_4 + \overline{c}_g c_2\|g\| - \frac{1}{l_c}\overline{c}_\Phi)e_\Phi^2 + \frac{1}{l_c}|e_\Phi e_\Psi|. \end{array}$$

Hence, we have the following result.

Proposition 4.2. *Assume that* $\Gamma \geq \Phi_0$ *and* $\overline{c}_g > \frac{\sqrt{\pi}}{\sqrt{6}}$. *Then*

$$\begin{array}{rcl} \frac{d}{d\xi}V_2(\Phi, g) & \leq & (c_0 + c_1\|g\|)\|g\|^2 \\ & + & (\overline{c}_g|c_0| + \frac{1}{l_c}c_5 + (c_2 + \frac{1}{l_c}c_6 + \overline{c}_g|c_1|)\|g\|)\|g\||e_\Phi| \\ & + & (\frac{1}{l_c}(c_4 - \overline{c}_\Phi) + \overline{c}_g c_2\|g\|)e_\Phi^2 + \frac{1}{l_c}|e_\Phi e_\Psi|. \end{array}$$

Assume that $\Gamma \geq \Phi_0$ and $\overline{c}_g > \frac{\sqrt{\pi}}{\sqrt{6}}$. Observe that if $\Psi = \widehat{\Psi}(\|g\|, \Phi)$ then

$$\begin{array}{rcl} \frac{d}{d\xi}V_2(\Phi, g) & \leq & (c_0 + c_1\|g\|)\|g\|^2 \\ & + & (\overline{c}_g|c_0| + \frac{1}{l_c}c_5 + (c_2 + \frac{1}{l_c}c_6 + \overline{c}_g|c_1|)\|g\|)\|g\||e_\Phi| \\ & + & (\frac{1}{l_c}(c_4 - \overline{c}_\Phi) + \overline{c}_g c_2\|g\|)e_\Phi^2 \\ & = & c_{11}\|g\|^2 + 2c_{12}\|g\|e_\Phi + c_{22}e_\Phi^2, \end{array} \tag{34}$$

where

$$\begin{aligned} c_{11} &:= (c_0 + c_1\|g\|),\\ c_{12} &:= \tfrac{1}{2}(\bar{c}_g|c_0| + \tfrac{1}{l_c}c_5 + (c_2 + \tfrac{1}{l_c}c_6 + \bar{c}_g|c_1|)\|g\|),\\ c_{22} &:= (\tfrac{1}{l_c}(c_4 - \bar{c}_\Phi) + \bar{c}_g c_2\|g\|). \end{aligned}$$

Note that the right-hand side of (34) is a quadratic form in $\|g\|$ and e_Φ (with coefficients being functions of $\|g\|$ and e_Φ). This quadratic form can be made negative definite by choosing sufficiently large $\bar{c}_\Phi$. A sufficient condition for $\frac{d}{d\xi}V_2(\Phi, g)$ to be negative definite for $\Psi = \widehat{\Psi}(\|g\|, \Phi)$ is

$$\begin{aligned} \Delta_1 &:= c_{11} < 0,\\ \Delta_2 &:= c_{11}c_{22} - c_{12}^2 > 0, \end{aligned}$$

which is satisfied if

$$\bar{c}_\Phi \geq c_4 + l_c\bar{c}_g c_2\|g\| + l_c\frac{(\bar{c}_g|c_0| + \frac{1}{l_c}c_5 + (c_2 + \frac{1}{l_c}c_6 + \bar{c}_g|c_1|)\|g\|)^2}{4|c_0 + c_1\|g\||}. \tag{35}$$

Observe that $\Gamma = \Phi_0$ implies that $c_0 = 0$, so that it may seem that the gain function $\bar{c}_\Phi$ blows up when $g = 0$. However, this is not the case. Note that $\Gamma = \Phi_0$ also implies that $\bar{c}_g|c_0| + \frac{1}{l_c}c_5 = 0$. Therefore, for $\Gamma = \Phi_0$ the right-hand side of the inequality (35) becomes

$$c_4 + l_c\bar{c}_g c_2\|g\| + l_c\frac{(c_2 + \frac{1}{l_c}c_6 + \bar{c}_g|c_1|)^2\|g\|^2}{4|c_1|\|g\|}.$$

For $\|g\| = 0$ this quantity is not defined. However, it has a finite limit c_4 at $\|g\| = 0$. Hence, one can conclude that $V_2(\Phi, g)$ is a valid control Lyapunov function also for $\Gamma = \Phi_0$.

4.3. H^1 backstepping: Step 3.

As a control Lyapunov function for full model (9), (10), and (11) we will use

$$V_3(\Phi, \Psi, g) := \frac{1}{2}\|g\|^2 + \frac{1}{2}e_\Phi^2 + \frac{4l_cB^2}{2}e_\Psi^2.$$

We will show that $\frac{d}{d\xi}V_3(\Phi, \Psi, g)$ can be made negative definite by a throttle control of the form

$$K_T(\Psi, u) = \Phi + \bar{c}_\Psi(\|g\|, \Phi)e_\Psi \tag{36}$$

for sufficiently large $\bar{c}_\Psi(\|g\|, \Phi)$. (For semiglobal stabilization $\bar{c}_\Psi(\|g\|, \Phi)$ can be chosen to be a constant depending on the desired region of operation.)

To calculate $\frac{d}{d\xi}V_3(\Phi, \Psi, g)$ we will need to express $\frac{d}{d\xi}e_\Psi$ in terms of e_Φ, e_Ψ, and g. (Throughout the paper, for simplicity, we skip the arguments of functions.) One has

$$\begin{aligned} \tfrac{d}{d\xi}e_\Psi &= \tfrac{d}{d\xi}\Psi - \tfrac{d}{d\xi}\widehat{\Psi}\\ &= \tfrac{1}{4l_cB^2}(-\bar{c}_\Psi e_\Psi) - e_\Phi\tfrac{d}{d\xi}\bar{c}_\Phi - \bar{c}_\Phi\tfrac{d}{d\xi}e_\Phi\\ &= \tfrac{1}{4l_cB^2}(-\bar{c}_\Psi e_\Psi) - \tfrac{\partial\bar{c}_\Phi}{\partial\|g\|}\tfrac{d}{d\xi}\|g\| - \tfrac{\partial\bar{c}_\Phi}{\partial\Phi}\tfrac{d}{d\xi}\Phi - \bar{c}_\Phi\tfrac{d}{d\xi}e_\Phi. \end{aligned}$$

Thus, using Proposition 4.1 and Lemma 4.2 one obtains

$$\begin{aligned}
4l_cB^2e_\Psi\tfrac{d}{d\xi}e_\Psi &\leq -\overline{c}_\Psi e_\Psi^2\\
&+ 4l_cB^2(|\tfrac{\partial\overline{c}_\Phi}{\partial\|g\|}|(((|(c_0+c_1\|g\|)|)\|g\||e_\Psi|+c_2\|g\||e_\Phi e_\Psi|)\\
&+ |\tfrac{\partial\overline{c}_\Phi}{\partial\Phi}|(\tfrac{1}{l_c}(c_3\|g\||e_\Psi|+|(c_4-\overline{c}_\Phi)||e_\Phi e_\Psi|+e_\Psi^2)\\
&+ \overline{c}_\Phi(\tfrac{1}{l_c}(c_3\|g\||e_\Psi|+(|c_4-\overline{c}_\Phi|)|e_\Phi e_\Psi|+e_\Psi^2)\\
&+ \overline{c}_g(|(c_0+c_1\|g\|)|)\|g\||e_\Psi|+c_2\|g\||e_\Phi e_\Psi|))).
\end{aligned}\tag{37}$$

Hence, using Proposition 4.2 and (37) one obtains

$$\begin{aligned}
\tfrac{d}{d\xi}V_3(\Phi,\Psi,g) \leq{}& (c_0+c_1\|g\|)\|g\|^2\\
&+ (\overline{c}_g|c_0|+\tfrac{1}{l_c}c_5+(c_2+\tfrac{1}{l_c}c_6+\overline{c}_g|c_1|)\|g\|)\|g\||e_\Phi|\\
&+ (\tfrac{1}{l_c}(c_4-\overline{c}_\Phi)+\overline{c}_gc_2\|g\|)e_\Phi^2+\tfrac{1}{l_c}|e_\Phi e_\Psi|\\
&- \overline{c}_\Psi e_\Psi^2\\
&+ 4l_cB^2(|\tfrac{\partial\overline{c}_\Phi}{\partial\|g\|}|(((|(c_0+c_1\|g\|)|)\|g\||e_\Psi|+c_2\|g\||e_\Phi e_\Psi|)\\
&+ |\tfrac{\partial\overline{c}_\Phi}{\partial\Phi}|(\tfrac{1}{l_c}(c_3\|g\||e_\Psi|+|(c_4-\overline{c}_\Phi)||e_\Phi e_\Psi|+e_\Psi^2)\\
&+ \overline{c}_\Phi(\tfrac{1}{l_c}(c_3\|g\||e_\Psi|+(|c_4-\overline{c}_\Phi|)|e_\Phi e_\Psi|+e_\Psi^2)\\
&+ \overline{c}_g(|(c_0+c_1\|g\|)|)\|g\||e_\Psi|+c_2\|g\||e_\Phi e_\Psi|)))\\
&= c_{11}\|g\|^2+2c_{12}\|g\||e_\Phi|+2c_{13}\|g\||e_\Psi|+c_{22}e_\Phi^2+2c_{23}|e_\Phi||e_\Psi|+c_{33}e_\Psi^2,
\end{aligned}\tag{38}$$

where

$$\begin{aligned}
c_{11} &= (c_0+c_1\|g\|),\\
c_{12} &= \tfrac{1}{2}(\overline{c}_g|c_0|+\tfrac{1}{l_c}c_5+(c_2+\tfrac{1}{l_c}c_6+\overline{c}_g|c_1|)\|g\|),\\
c_{22} &= (\tfrac{1}{l_c}(c_4-\overline{c}_\Phi)+\overline{c}_gc_2\|g\|),\\
c_{13} &:= \tfrac{1}{2}(4l_cB^2|\tfrac{\partial\overline{c}_\Phi}{\partial\|g\|}|(((|(c_0+c_1\|g\|)|)\\
&\quad +4B^2|\tfrac{\partial\overline{c}_\Phi}{\partial\Phi}|c_3+4B^2\overline{c}_\Phi c_3+4l_cB^2\overline{c}_\Phi\overline{c}_g(|(c_0+c_1\|g\|)|),\\
c_{23} &:= \tfrac{1}{2}(\tfrac{1}{l_c}+4l_cB^2c_2\|g\|+4B^2|\tfrac{\partial\overline{c}_\Phi}{\partial\Phi}||(c_4-\overline{c}_\Phi)|\\
&\quad +4B^2\overline{c}_\Phi|c_4-\overline{c}_\Phi|+4l_cB^2\overline{c}_\Phi\overline{c}_gc_2\|g\|),\\
c_{33} &:= -\overline{c}_\Psi+4B^2|\tfrac{\partial\overline{c}_\Phi}{\partial\Phi}|+4B^2\overline{c}_\Phi.
\end{aligned}$$

Note that the right-hand side of (38) is a quadratic form in $\|g\|$, $|e_\Phi|$, and $|e_\Psi|$ (with coefficients being functions of g and e_Φ that can be bounded by functions of $\|g\|$ and e_Φ. Assuming that $\overline{c}_g > \frac{\sqrt{\pi}}{\sqrt{6}}$ and $\overline{c}_\Phi$ satisfies (35), we can make this quadratic form negative definite by choosing sufficiently large $\overline{c}_\Psi$. Sufficient conditions for $\frac{d}{d\xi}V_3(\Phi,\Psi,g)$ to be negative definite everywhere are

$$\begin{aligned}
\Delta_1 &= c_{11}<0,\\
\Delta_2 &= c_{11}c_{22}-c_{12}^2>0,\\
\Delta_3 &:= c_{33}\Delta_2+2c_{12}c_{13}c_{23}-c_{22}c_{13}^2-c_{11}c_{23}^2<0.
\end{aligned}$$

The condition $\Delta_1 < 0$ is obviously satisfied (see Step 1). To enforce the condition $\Delta_2 > 0$ one should choose $\overline{c}_\Phi$ that satisfies (35) (see Step 2). Finally, once $\overline{c}_\Phi$ satisfies the inequality (35), to assure that $\Delta_3 < 0$, at each point, the gain $\overline{c}_\Psi$ should satisfy the inequality

$$\overline{c}_\Psi > 4B^2 \left|\frac{\partial \overline{c}_\Phi}{\partial \Phi}\right| + 4B^2 \overline{c}_\Phi + \frac{2c_{12}c_{13}c_{23} - c_{22}c_{13}^2 - c_{11}c_{23}^2}{\Delta_2}. \tag{39}$$

If $\Gamma > \Phi_0$ then $\Delta_2 > 0$ holds everywhere and hence the right-hand side of the inequality (39) is defined for everywhere (see Step 2).

However, $\Gamma = \Phi_0$ implies that $c_0 = 0$, and thus Δ_2 vanishes if $g = 0$. Therefore, it may seem that the gain function $\overline{c}_\Psi$ blows up when $g = 0$. However, this is not the case. One can show that the quantity $2c_{12}c_{13}c_{23} - c_{22}c_{13}^2 - c_{11}c_{23}^2$ also vanishes if $g = 0$ and $\frac{2c_{12}c_{13}c_{23} - c_{22}c_{13}^2 - c_{11}c_{23}^2}{\Delta_2}$ has a finite limit as $\|g\|$ goes to zero. Hence, one can conclude that $V_3(\Phi, \Psi, g)$ is a valid control Lyapunov function also for $\Gamma = \Phi_0$. (See similar remarks at the end of section 4.2.)

4.4. The case $\Gamma < \Phi_0$. If the position of the peak is unknown or if the characteristic shifts from its nominal position (because of disturbance, etc.), it may happen that $\Gamma < \Phi_0$. In that case it follows from Proposition 3.1 that $(\Phi, \Psi, g) = (\Gamma, \Psi_c(\Gamma), 0)$ is an unstable equilibrium that cannot be stabilized by the virtual control (30). However, one can prove that the controller of the form (36) will guarantee that the dynamics of the closed-loop system are confined to a ball containing $(\Gamma, \Psi_c(\Gamma), 0)$. The radius of the ball can be made arbitrarily small if one can use arbitrarily high gains in the controller. This modification of the gains in the controller in comparison with the case $\Gamma \geq \Phi_0$ is to be expected, as the controller gains proposed for the case $\Gamma \geq \Phi_0$ were not designed to work also in the case $\Gamma < \Phi_0$.

What we present below is a simple, but not necessarily optimal, way of constructing a controller that confines the dynamics to a ball. Our goal was to provide a simple proof that this is possible, not to actually design a controller that is optimal in any sense.

Assume that $\Phi_{infl} < \Gamma < \Phi_0$. We are going to use notation of the previous sections. We need to introduce two new symbols:

$$\overline{c}_0 := a\Psi_c'(\Gamma),$$

$$\overline{c}_{11} := \overline{c}_0 + c_1\|g\|.$$

These quantities will replace c_0 and c_{11}, respectively. One can show that

$$\frac{d}{d\xi}V_3(\Phi, \Psi, g) \leq \overline{c}_{11}\|g\|^2 + 2c_{12}\|g\||e_\Phi| + 2c_{13}\|g\||e_\Psi| + c_{22}e_\Phi^2 + 2c_{23}|e_\Phi||e_\Psi| + c_{33}e_\Psi^2.$$

Observe that we had to replace c_0 and c_{11} with $\overline{c}_0$ and $\overline{c}_{11}$ since now $\Psi_c'(\Gamma) > 0$.

It will be also useful to introduce the following notation. Let

$$\overline{DV}_2(\Phi, g) := \overline{c}_{11}\|g\|^2 + 2c_{12}\|g\||e_\Phi| + c_{22}e_\Phi^2.$$

Then

$$\frac{d}{d\xi}V_3(\Phi, \Psi, g) \leq \overline{DV}_2(\Phi, g) + 2c_{13}\|g\||e_\Psi| + 2c_{23}|e_\Phi||e_\Psi| + c_{33}e_\Psi^2.$$

Note that $\overline{c}_0 > 0$ and $c_1 < 0$. Therefore, the upper bound on $\frac{d}{d\xi}V_3(\Phi, \Psi, g)$ cannot be made negative everywhere; as for $e_\Phi = e_\Psi = 0$ and $0 < \|g\| < \frac{-\overline{c}_0}{c_1}$ one has

$\overline{c}_{11} > 0$. However, we will show that one can arbitrarily reduce the size of the set where $\frac{d}{d\xi}V_3(\Phi,\Psi,g) > 0$ by using high gains in the controller. This can be accomplished as follows. First, one can arbitrarily reduce the interval on which $\overline{c}_{11} > 0$ by using a high gain $\overline{c}_g$, which makes c_1 big negative. Second, one can use high gains $\overline{c}_\Phi$ and $\overline{c}_\Psi$, which make c_{22} and c_{33} big negative.

Let ϵ be an arbitrary positive number. We are going to show that by using sufficiently high gains $\overline{c}_g$, $\overline{c}_\Phi$, and $\overline{c}_\Psi$ one can guarantee that $\frac{d}{d\xi}V_3(\Phi,\Psi,g) < 0$ outside the set $\mathcal{M}_\epsilon := \{(\Phi,\Psi,g) : \|g\| < \epsilon,\ |e_\Phi| < \epsilon,\ |e_\Psi| < \epsilon\}$.

Step 1. Choose $\overline{c}_g$ such that for $\|g\| \geq \epsilon$ one has $\overline{c}_{11} \leq -3$.

Step 2. Choose $\overline{c}_\Phi$ such that the following conditions (2a) and (2b) are satisfied:

(2a) For $\|g\| \geq \epsilon$ one has $c_{22} \leq \frac{c_{12}^2}{\overline{c}_{11}+2}$.

Note that for a fixed $\|g\|$, $\overline{DV}_2(\Phi,g)$ can be viewed as a quadratic function of $|e_\Phi|$. One can show using some elementary algebra that our choice of $\overline{c}_\Phi$ guarantees that for $\|g\| \geq \epsilon$ one has $\overline{DV}_2(\Phi,g) \leq -2\epsilon^2$.

(2b) For $\|g\| < \epsilon$ one has $\frac{-2c_{12}}{c_{22}} + \sqrt{|\frac{c_{11}}{c_{22}}|} + \sqrt{|\frac{2}{c_{22}}|} < 1$.

One can show using some elementary algebra that this choice guarantees that for $\|g\| < \epsilon$ and $|e_\Phi| \geq \epsilon$ one has $\overline{DV}_2(\Phi,g) \leq -2\epsilon^2$.

Step 3. Choose $\overline{c}_\Psi$ so that the following conditions (3a) and (3b) are satisfied:

(3a) For $\|g\| \geq \epsilon$ or $|e_\Phi| \geq \epsilon$ one has $c_{33} \leq \frac{(c_{13}\|g\|+c_{23}|e_\Phi|)^2}{\overline{DV}_2(\Phi,g)+\epsilon^2}$.

Note that c_{33} is bounded, as the choice of $\overline{c}_\Phi$ in Step 2 guarantees that for $\|g\| \geq \epsilon$ or $|e_\Phi| \geq \epsilon$ one has $\overline{DV}_2(\Phi,g) + \epsilon^2 \leq -\epsilon^2$.

One can show that this choice of $\overline{c}_\Psi$ guarantees that for $\|g\| \geq \epsilon$ or $|e_\Phi| \geq \epsilon$ one has $\frac{d}{d\xi}V_3(\Phi,\Psi,g) \leq -\epsilon^2$.

(3b) For $\|g\| < \epsilon$ and $|e_\Phi| < \epsilon$ one has $\frac{-2c_{13}-2c_{23}}{c_{33}} + \sqrt{\frac{|c_{11}|+|c_{22}|+2c_{12}+1}{|c_{33}|}} < 1$.

One can show that this choice guarantees that for $\|g\| < \epsilon$ and $|e_\Phi| < \epsilon$, but for $|e_\Psi| \geq \epsilon$ one has $\frac{d}{d\xi}V_3(\Phi,\Psi,g) \leq -\epsilon^2$.

Therefore, we have the following result.

PROPOSITION 4.3. *Let the gains $\overline{c}_g$, $\overline{c}_\Phi$, and $\overline{c}_\Psi$ satisfy the conditions stated in Steps 1–3 above. Then outside the set $\mathcal{M}_\epsilon = \{(\Phi,\Psi,g) : \|g\| < \epsilon,\ |e_\Phi| < \epsilon,\ |e_\Psi| < \epsilon\}$ one has*

$$\frac{d}{d\xi}V_3(\Phi,\Psi,g) < -\epsilon^2.$$

Therefore, the state of the closed-loop system enters in a finite time the set

$$\mathcal{N}_\epsilon := \left\{(\Phi,\Psi,g) : V_3(\Phi,\Psi,g) < \frac{4l_cB^2}{2(1+4l_cB^2)}\epsilon^2\right\}.$$

Note that the high gains of the controller presented in this section are required if one wants to reduce the size of the dynamics and in particular of the stall cell, not to stabilize a small stall cell. If one just wants to confine the dynamics to a ball, high gains are not required.

It is not clear at the moment what are the dynamics of the closed-loop system inside the absorbing set $\mathcal{N}_\epsilon$. This issue is currently under investigation.

5. Controllers for Galerkin projections of the full model. In section 4 we constructed a feedback controller stabilizing a peak or any axisymmetric equilibrium to the right of the peak for the full Moore–Greitzer PDE model. The feedback law is given by (36) and has a general form

$$K_T(\Psi, u) = K_T(\|g\|_{H^1_K}, \Phi, \Psi).$$

In terms of the magnitudes A_p of the Fourier modes of a stall cell g the control law looks like familiar backstepping control laws for MG3 with A_1 replaced with $(\sum_{p=1}^{\infty}(1+\frac{am}{p})(pA_p)^2)^{\frac{1}{2}}$. An implementation of this control law would require access to an infinite number of modes of a stall cell g, which is practically impossible.

Remark 5.1. The following was communicated to the authors by Richard Murray from Caltech.

The number of accessible modes depends on the number of pressure sensors used to detect a nonaxisymmetric pressure distribution and their distance from the compressor face. Since it requires $2n+1$ sensors to instantaneously detect the first n modes (by fitting a linear combination of spatial sinusoids), the number of sensors gets somewhat large for higher modes. In addition, with the minimal number of sensors, the last mode is pretty noisy.

The Caltech compressor rig has six sensors so that a measurement of up to second mode magnitude is possible. The rig has enough ports to use 16 sensors, which would make it possible to measure the magnitudes of up to the seventh mode.

Another factor is the distance back from the compressor face. Recall that the magnitudes of the Fourier modes of a stall cell fall off by $e^{-\eta n}$ where η is the nondimensional distance from the compressor face. For the Caltech rig, η is about 0.5, so beyond the third or fourth mode one would not be able to pick out the signal from the noise.

Remark 5.2. An alternative to an instantaneous detection of the first n modes by using $2n+1$ sensors would be using fewer sensors and an observer to reconstruct the modes. If the speed of rotation of the stall cell is known, then one can easily verify that the first few modes are observable (even from a single sensor). The fast decay of the higher modes because of the distance of the sensors from the compressor face makes the observability of these modes poor and therefore is still a limiting factor in the number of detectable modes.

If n modes are accessible, a practical implementation of the controller could use $(\sum_{p=1}^{n}(1+\frac{am}{p})(pA_p)^2)^{\frac{1}{2}}$, i.e., a truncation of the infinite series $(\sum_{p=1}^{\infty}(1+\frac{am}{p})(pA_p)^2)^{\frac{1}{2}}$ representing the H^1_K norm of the stall cell g. A natural question arises: what can we say about the controller of the form (36) that uses first n modes of the stall cell g? In this section we will show that this controller *stabilizes the Galerkin projection of the full Moore–Greitzer PDE model onto its first* n *modes.*

It is easy to show that the ODEs describing evolution of the pth mode of g are

$$\dot{a}_p = \frac{p}{p+am}\left(\frac{a}{\pi}\int_0^{2\pi} \Psi_c(\Phi+g)\sin(p\theta)d\theta + \frac{1}{2}pb_p\right),$$

$$\dot{b}_p = \frac{p}{p+am}\left(\frac{a}{\pi}\int_0^{2\pi} \Psi_c(\Phi+g)\cos(p\theta)d\theta - \frac{1}{2}pa_p\right)$$

or, equivalently,

$$\dot{A}_p = \frac{p}{p+am}\left(\frac{a}{\pi}\int_0^{2\pi} \Psi_c(\Phi+g)\sin(p\theta+r_p)d\theta\right),$$

$$\dot{r}_p = \frac{p}{p+am}\left(-\frac{1}{2}p + \frac{a}{\pi A_p}\int_0^{2\pi}\Psi_c(\Phi+g)\cos(p\theta+r_p)d\theta\right),$$

where g is represented as

$$g = \sum_{p=1}^{\infty}(a_p\sin(p\theta)+b_p\cos(p\theta)) = \sum_{p=1}^{\infty}A_p\sin(p\theta+r_p).$$

A Galerkin projection of the stall PDE with n modes would be the set of $2n$ ODEs as above with g replaced with

$$g_n := \sum_{p=1}^{n}A_p\sin(p\theta+r_p).$$

Hence, an approximate model would consist of equations

$$\dot{A}_p = \frac{p}{p+am}\left(\frac{a}{\pi}\int_0^{2\pi}\Psi_c(\Phi+g_n)\sin(p\theta+r_p)d\theta\right), \tag{40}$$

$$\dot{r}_p = \frac{p}{p+am}\left(-\frac{1}{2}p + \frac{a}{\pi A_p}\int_0^{2\pi}\Psi_c(\Phi+g_n)\cos(p\theta+r_p)d\theta\right) \tag{41}$$

for $p = 1,\dots,n$ and (10) and (11).

One can prove the following result.

THEOREM 5.1. *The controller of the form* (36) *that uses first* n *modes of the stall stabilizes the system of* $2n+2$ *ODEs consisting of the Galerkin projection of* (9) *onto first* n*-modes of* g *and* (10) *and* (11).

Proof. We are going to use a backstepping controller design, almost identical to the one used for the full model.

Step 1. As a control Lyapunov function for (40) and (41) one uses

$$V_1^n(\Phi,\Psi,g_n) := \frac{1}{2}\|g_n\|_{H_K^1}^2 = \frac{1}{2}\sum_{p=1}^{n}\left(1+\frac{am}{p}\right)(pA_p)^2.$$

One has

$$\frac{d}{d\xi}V_1^n(\Phi,\Psi,g_n) = \sum_{p=1}^{n}\left(1+\frac{am}{p}\right)\left(p^2A_p\frac{d}{d\xi}A_p\right),$$
$$\frac{a}{\pi}\int_0^{2\pi}\left(\Psi_c(\Phi+g_n)\sum_{p=1}^{n}p^2A_p\sin(p\theta+r_p)d\theta\right).$$

Integrating by parts, one gets

$$\begin{aligned}\frac{d}{d\xi}V_1^n(\Phi,\Psi,g_n) &= \frac{a}{\pi}\int_0^{2\pi}\left(\Psi_c'(\Phi+g_n)\frac{dg_n}{d\theta}\sum_{p=1}^{n}pA_p\cos(p\theta+r_p)d\theta\right)\\ &= \frac{a}{\pi}\int_0^{2\pi}\Psi_c'(\Phi+g_n)\left(\frac{dg_n}{d\theta}\right)^2 d\theta.\end{aligned}$$

Steps 2 and 3 of the backstepping procedure are exactly the same as in the case of the full model (with g replaced with g_n). $\square$

Simulations. In this section we illustrate the action of a truncated H^1-controller with some simulations. The full Moore–Greitzer model has been simulated using 64 Fourier modes (128 states) to represent the stall cell dynamics. The compressor characteristics are assumed to be a cubic function [15]

$$\Psi_c(\phi) := \psi_0 + H\left(1 + \frac{3}{2}\left(\frac{\phi}{W} - 1\right) - \frac{1}{2}\left(\frac{\phi}{W} - 1\right)^3\right).$$

The coefficients ψ_0, H, and W represent, respectively, the shut-off pressure rise, semiheight, and semiwidth of the characteristic. The parameters ψ_0, H, W, a, m, l_c, and B determine the compressor model. The Greitzer B parameter determines if the compressor is likely to stall or surge. Stalling compressors are characterized by a low value of the B parameter, while surging compresssors are characterized by a high value of the B parameter. We simulated a low B compressor and a high B compressor. We initialized both models with the initial condition for the surge dynamics near the peak of the compressor characteristic. The initial shape of the stall cell was a pure first mode.

Figures 2 and 3 show that, as expected, the state of the uncontrolled low B compressor settled at a rotating stall condition with a significant pressure drop.

The simulations of uncontrolled dynamics are followed by simulations of the dynamics controlled with a truncated H^1-controller using first four Fourier modes of the stall variable and constant gains. The control function was saturated at 0 to avoid using negative values of the throttle coefficient. We see the state of the low B compressor after a transient period of growing the stall variable and a drop in pressure settled at the desired axisymmetric equilibria near the peak of the compressor characteristic. Figures 4 and 5 show the state evolution.

Figures 6 and 7 show the evolution of the state of the uncontrolled high B compressor. The stall cell initially grows fast, but after the mean mass flow reaches the reverse flow part of the characteristic, it decays. The state of the system undergoes a deep surge cycle.

Figures 8 and 9 show the evolution of the state of the controlled high B compressor. The controller opens the throttle and prevents a transition into a deep surge cycle. Note that the mean flow spends more time in the interval between the well and the peak than in the uncontrolled case. This causes the stall variable to grow and stay large. The stall variable starts to decay only after the mean mass flow becomes bigger than the value corresponding to the peak of the compressor characteristic. This explains why the stall variable decays faster in the uncontrolled case than in the controlled one.

Conclusion. We have constructed a feedback controller stabilizing a peak or any axisymmetric equilibrium to the right of the peak of the compressor characteristic for the full Moore–Greitzer model. The control law resembles control laws for a one-mode truncation of the full model. In the case when the set-point parameter in the controller is such that there is no stable axisymmetric equilibrium we can still guarantee that the dynamics of the closed-loop system are confined to a ball, whose radius can be made arbitrarily small by choosing sufficiently high gains in the controller.

A practical implementation of the H^1_K-controller would use a finite sum $(\sum_{p=1}^{n}(1+\frac{am}{p})(pA_p)^2)^{\frac{1}{2}}$. We proved that this truncated feedback controller actually globally stabilizes the system of $2n+2$ ODEs consisting of the Galerkin projection of the PDE describing the stall onto its first n-modes and two ODEs describing the surge dynamics.

Simulations: stall and surge dynamics
Low B, uncontrolled.

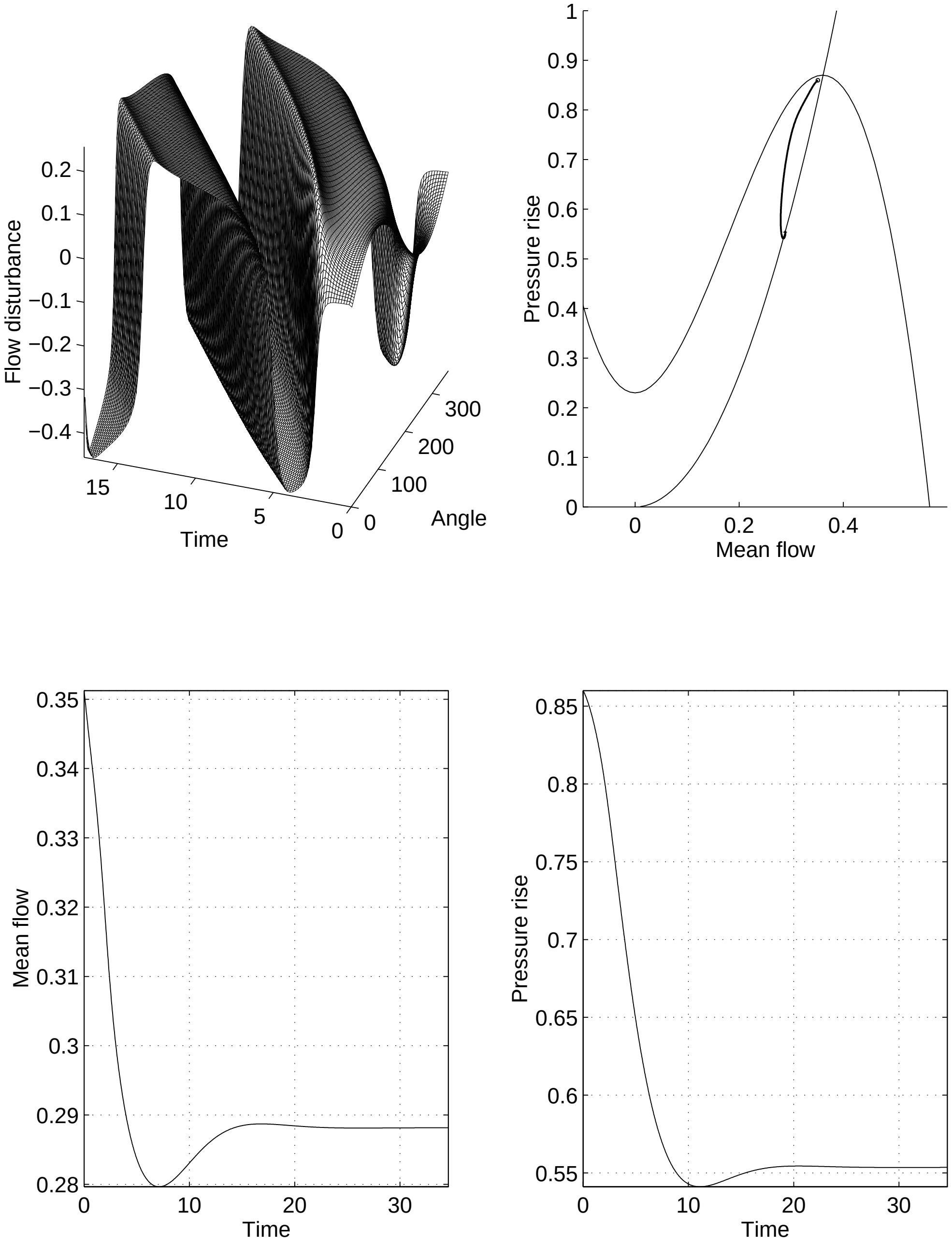


Fig. 2.

Stall cell evolution
Low B, uncontrolled

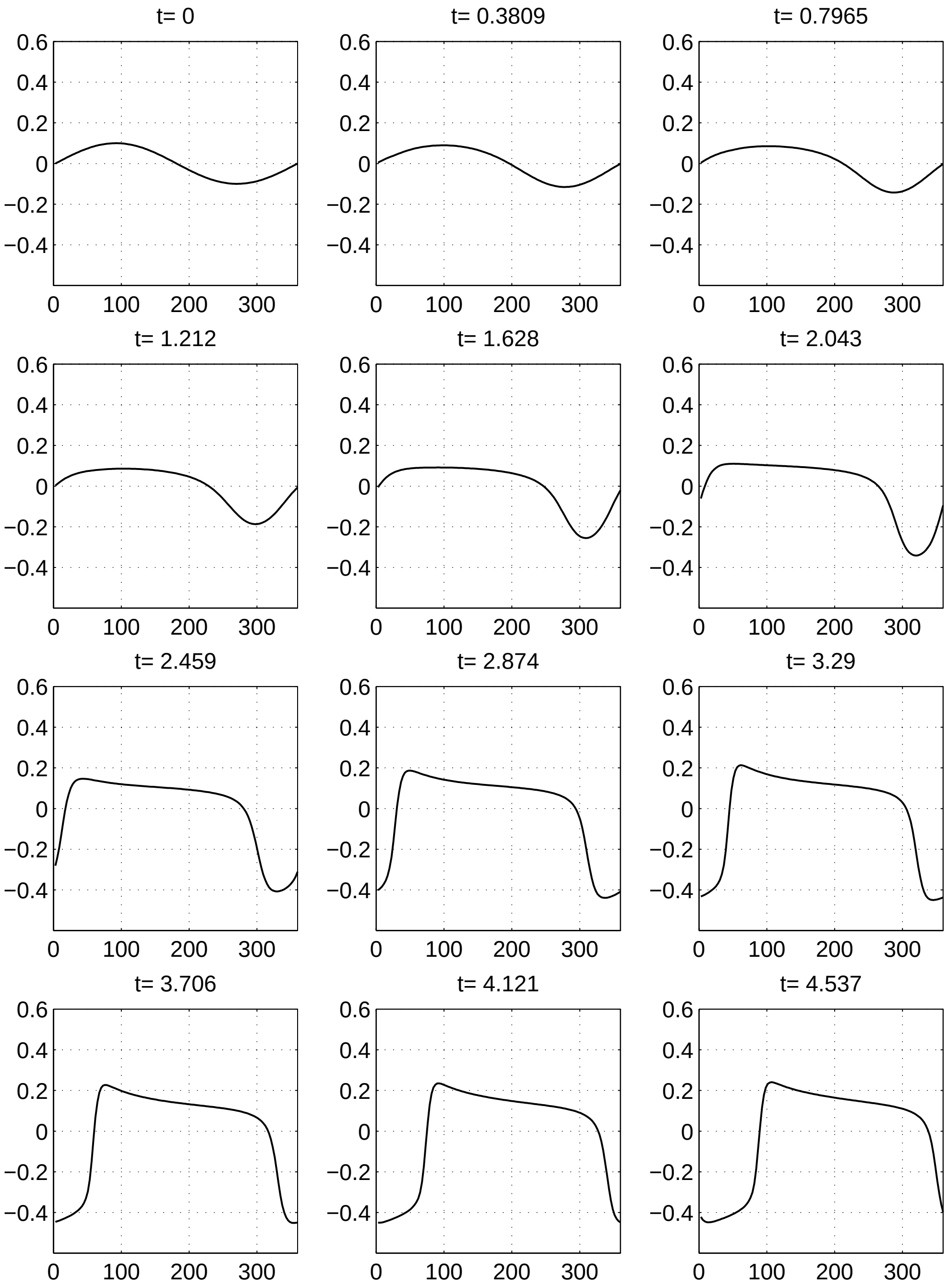


Fig. 3.

Simulations: stall and surge dynamics
Low B, controlled

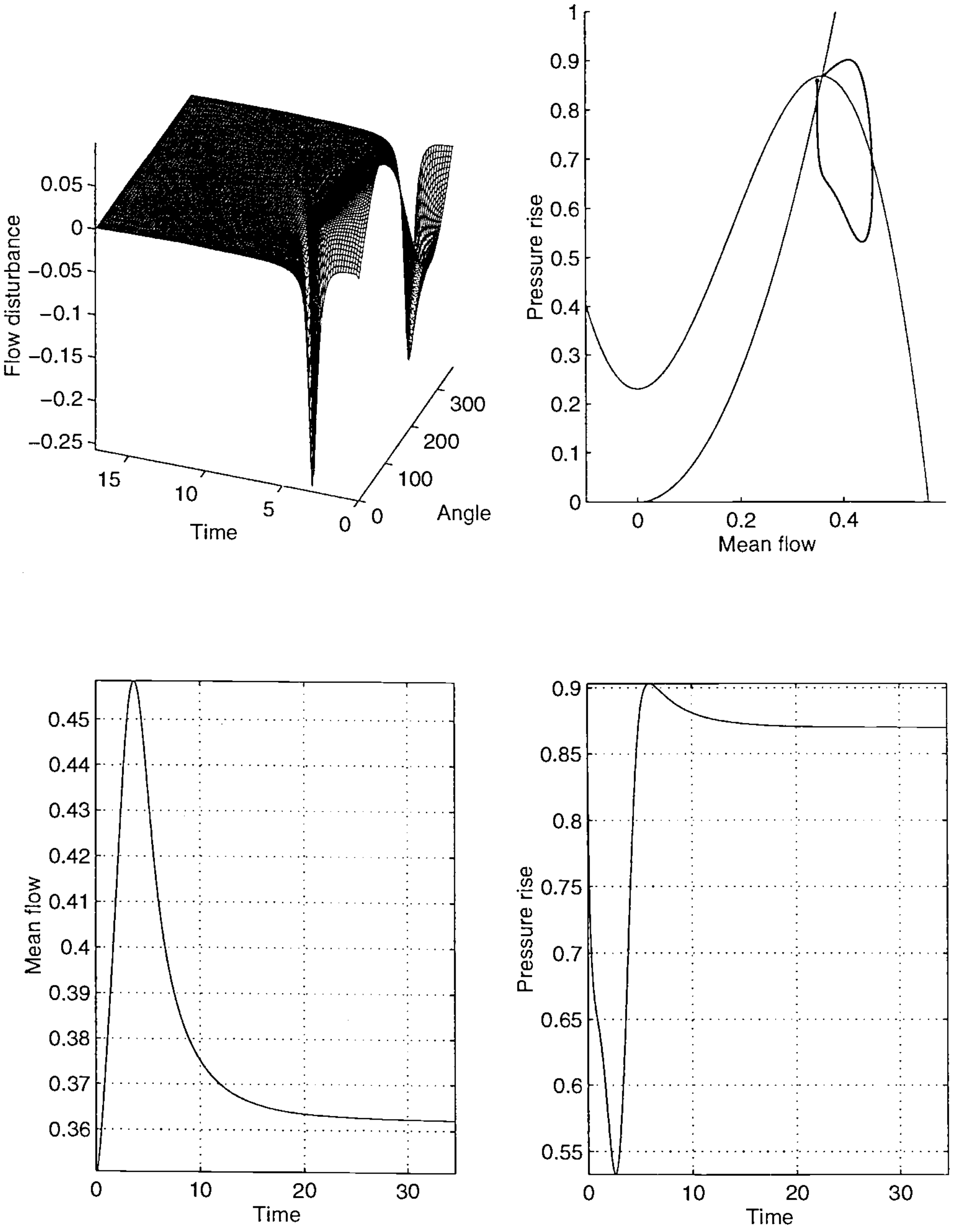


FIG. 4.

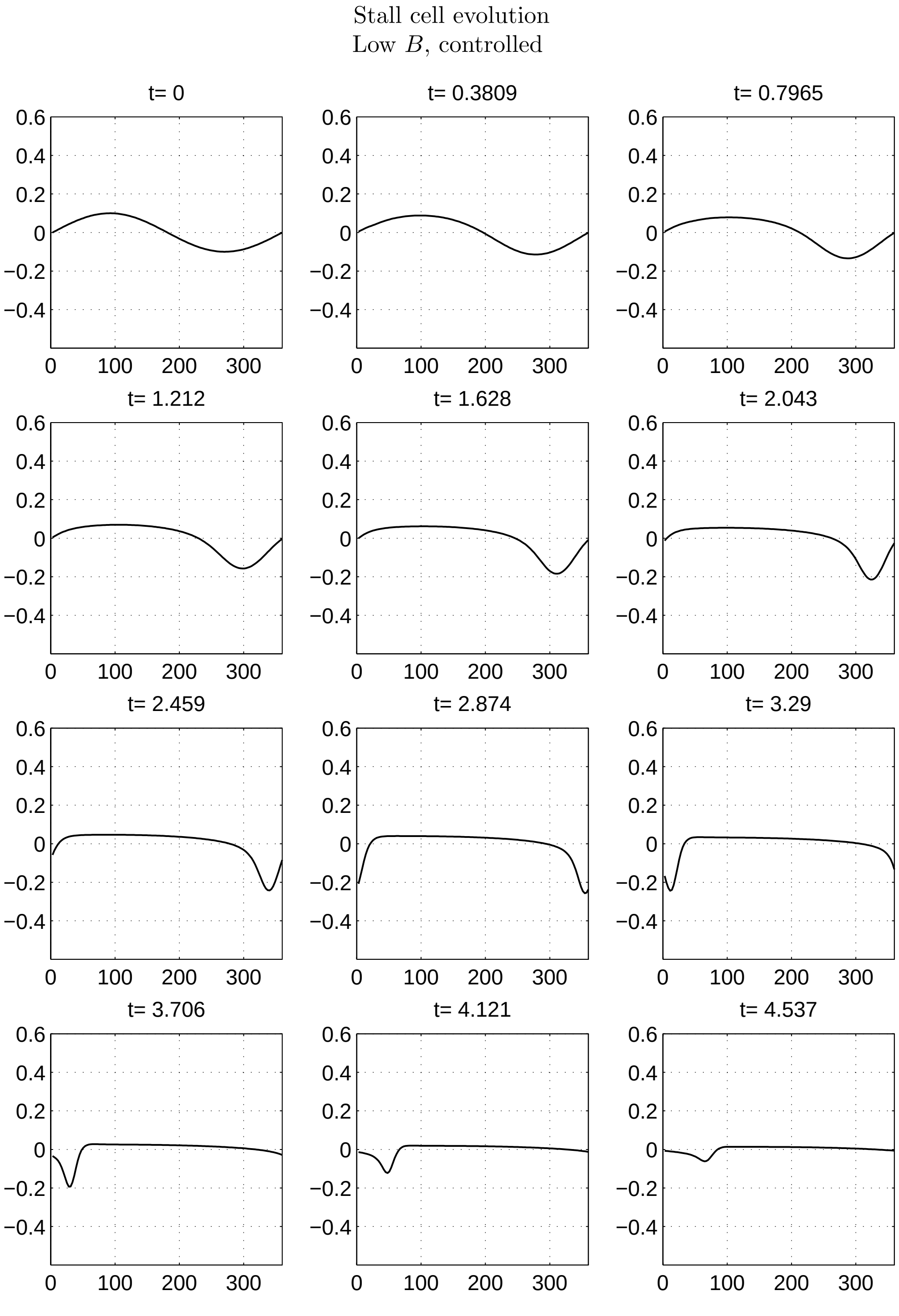

Stall cell evolution
Low B, controlled
t= 0
t= 0.3809
t= 0.7965
t= 1.212
t= 1.628
t= 2.043
t= 2.459
t= 2.874
t= 3.29
t= 3.706
t= 4.121
t= 4.537
0.6
0.4
0.2
0
−0.2
−0.4
0
100
200
300

FIG. 5.

Simulations: stall and surge dynamics
High B, uncontrolled

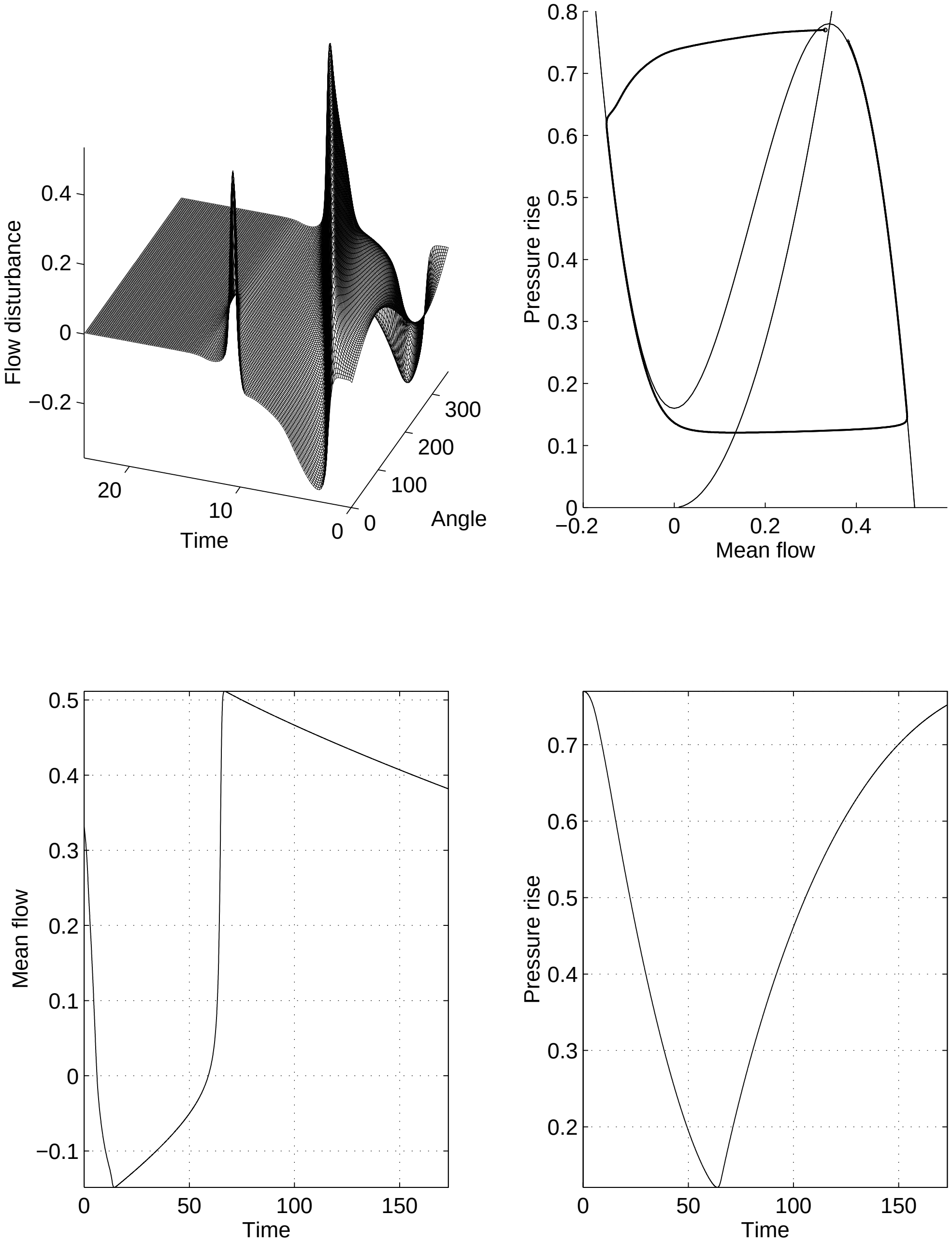


Fig. 6.

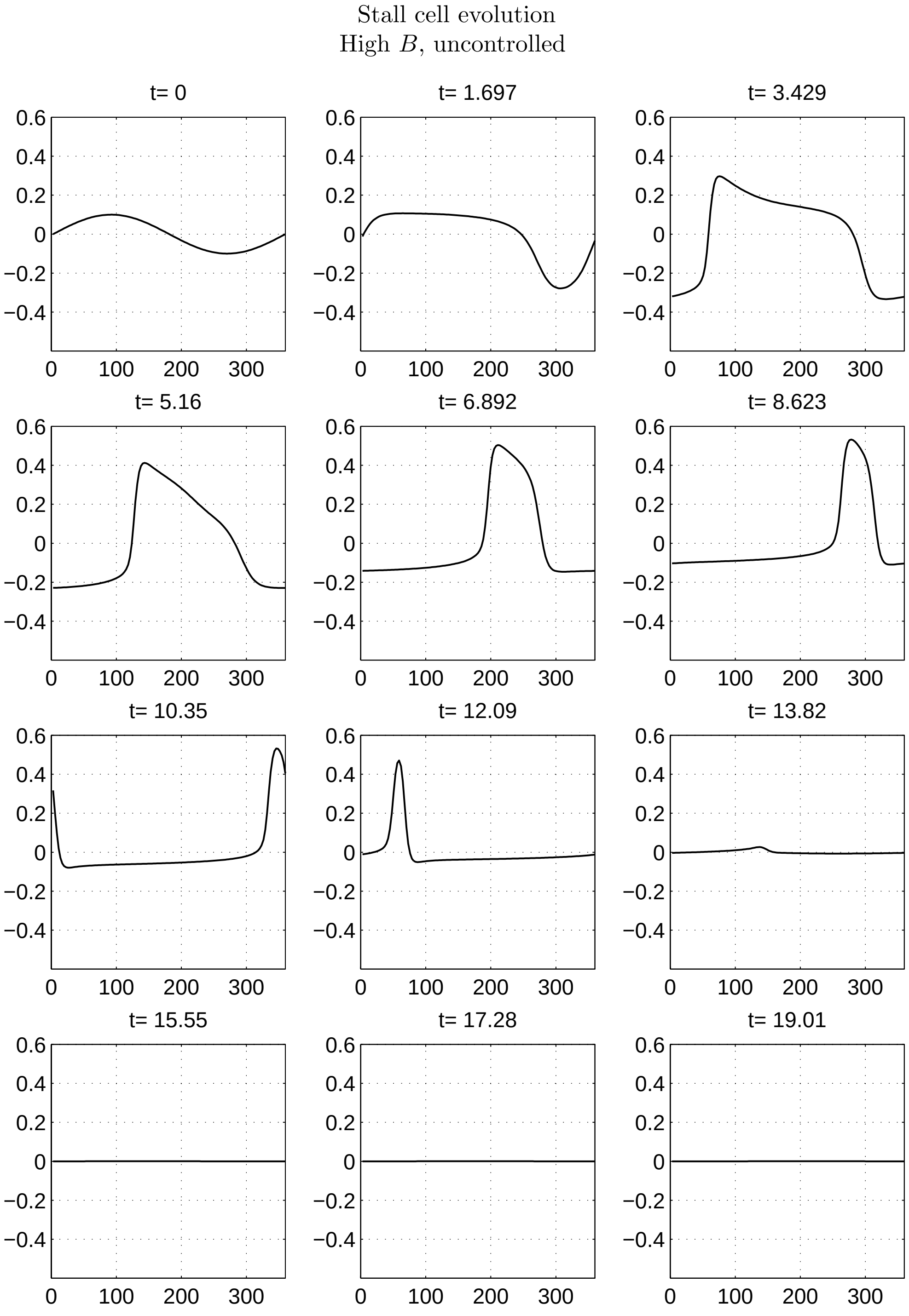
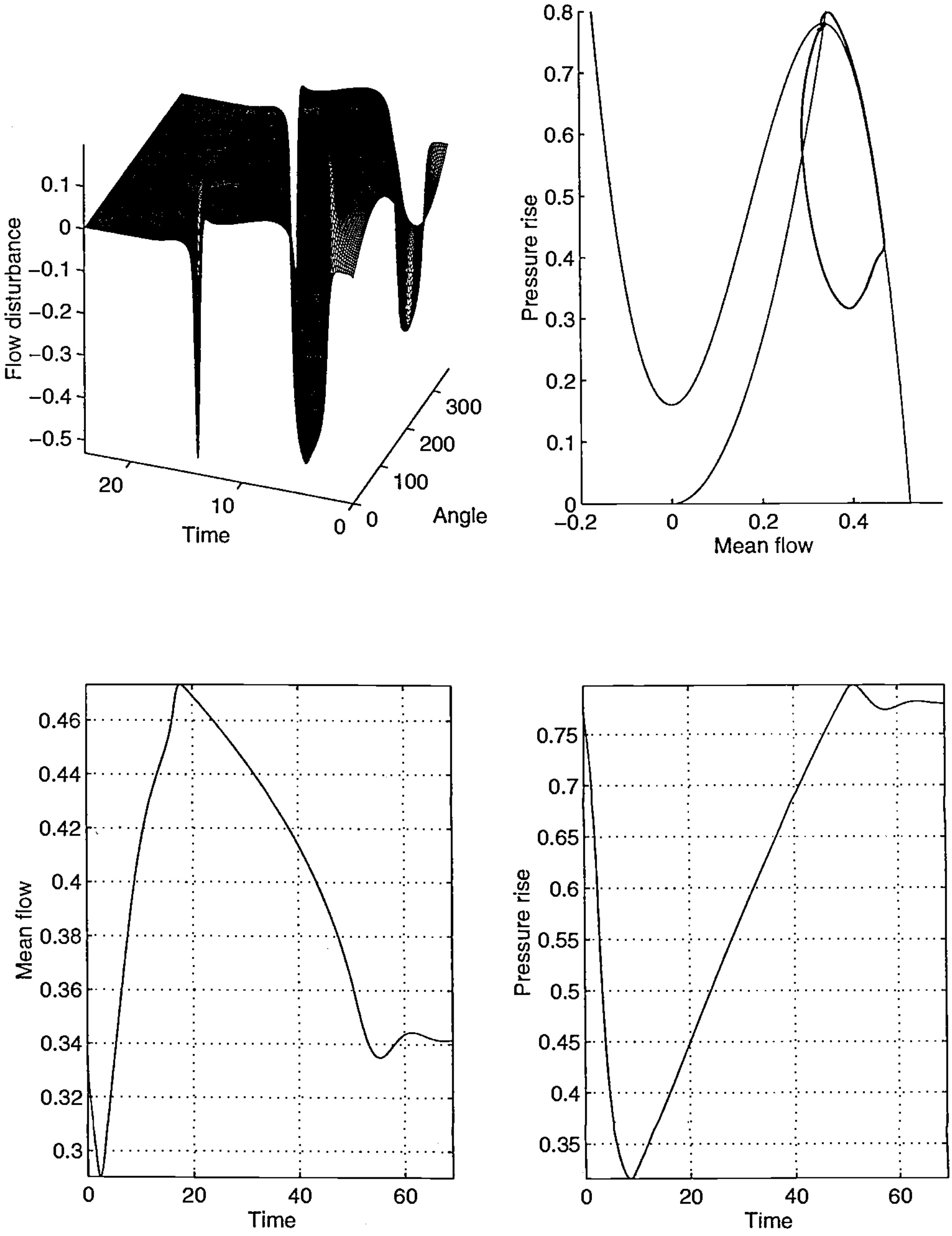

Stall cell evolution
High B, uncontrolled
t= 0
t= 1.697
t= 3.429
t= 5.16
t= 6.892
t= 8.623
t= 10.35
t= 12.09
t= 13.82
t= 15.55
t= 17.28
t= 19.01
0.6
0.4
0.2
0
−0.2
−0.4
0
100
200
300

FIG. 7.



Simulations: stall and surge dynamics
High B, controlled

FIG. 8.

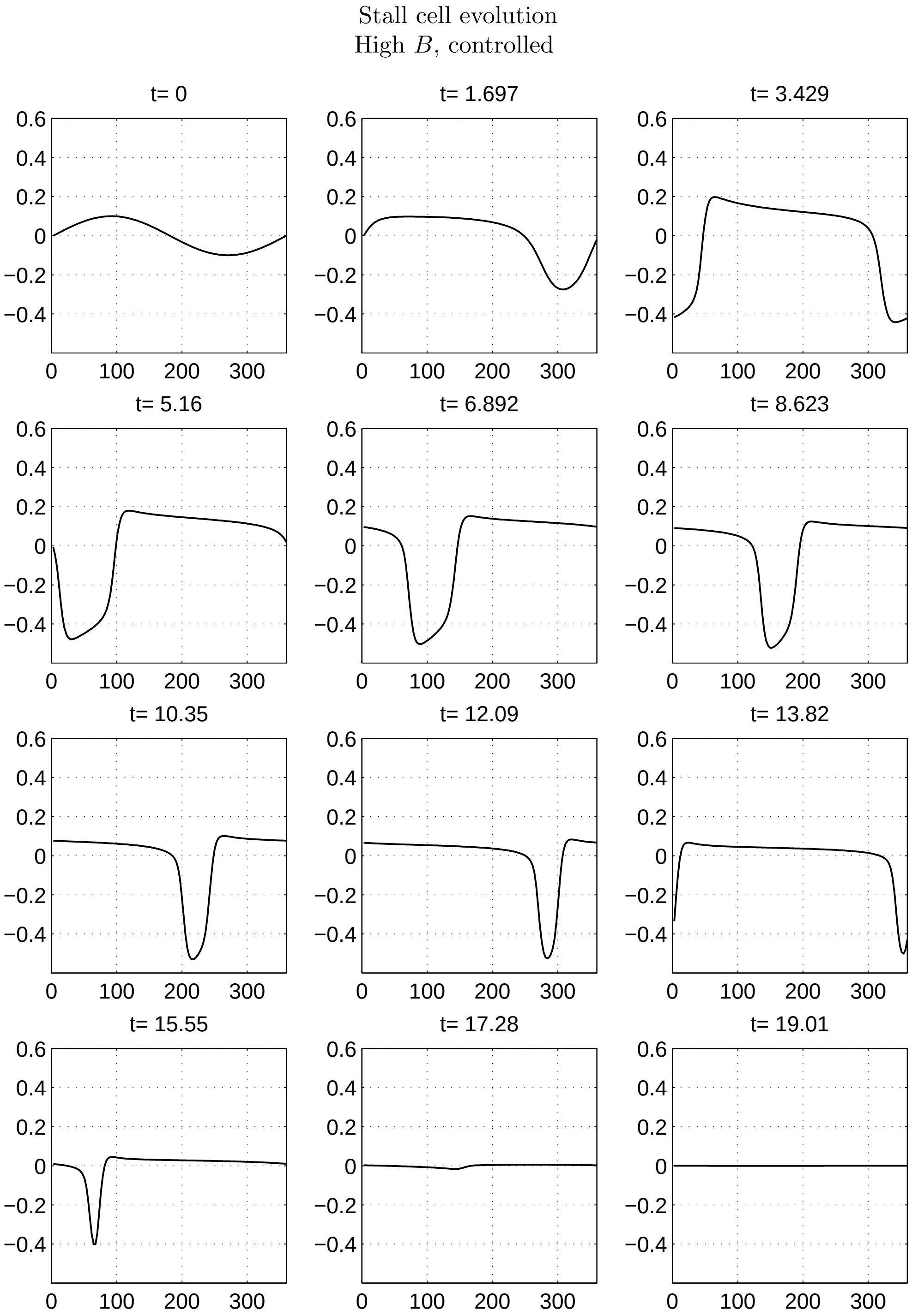

Stall cell evolution
High B, controlled
t= 0
t= 1.697
t= 3.429
t= 5.16
t= 6.892
t= 8.623
t= 10.35
t= 12.09
t= 13.82
t= 15.55
t= 17.28
t= 19.01
0.6
0.4
0.2
0
−0.2
−0.4
0
100
200
300

Fig. 9.

While one may argue that the necessity of finding the magnitudes of the Fourier modes of a stall cell for a feedback requires complicated implementation, let us observe that such information would be necessary anyway for any feedback law based on a Galerkin approximation of a Moore–Greitzer PDE model with a finite number of modes.

One feature of the H^1-controller is not desirable. Namely, its gain increases for higher order modes of the stall cell. This does not seem to be necessary (see Remark 4.1). We conjecture that one can replace the H^1_K norm of g with the L^∞ norm of g or with the minimum of g in the controller and have the same stabilizability property without using a higher gain for higher order modes of the stall cell. We are currently working on the proof of this conjecture.

Although we have concentrated on a specific model here, the methods developed in this paper can be used, with slight variations, in a variety of problems involving evolution equations.

Acknowledgments. The authors would like to thank Petar Kokotovic for initiating their collaboration on analysis and control of the full Moore–Greitzer model. The authors acknowledge input from Richard Murray on the stall cell measurements. The second author would also like to thank Björn Birnir and Tom Sideris for helpful discussions.

REFERENCES

[1] A. Banaszuk and A. Krener, *Design of controllers for MG3 compressor models with general characteristics using graph backstepping*, Automatica, 35 (1999), pp. 1343–1368; available on http://talon.colorado.edu/˜ banaszuk, 1996.

[2] N. A. Cumpsty, *Compressor Aerodynamics*, John Wiley and Sons, New York, 1989.

[3] K. M. Eveker, D. L. Gysling, C. N. Nett, and O. P. Sharma, *Integrated control of rotating stall and surge in aeroengines*, in Sensing, Actuation, and Control in Aeropropulsion; SPIE 1995 International Symposium on Aerospace/Defense Sensing and Dual-Use Photonics, 1995.

[4] A. Haraux, *Nonlinear Evolution Equations—Global Behavior of Solutions*, Lecture Notes in Math. 841, Springer, New York, 1981.

[5] T. Kato, *Spectral Theory and Differential Equations*, Lecture Notes in Math. 448, Springer, New York, 1975, pp. 25–70.

[6] A. Krener, *The Feedbacks That Soften the Primary Bifurcation of MG3*, 1995, preprint.

[7] M. Krstić, I. Kanellakopoulos, and P. Kokotović, *Nonlinear and Adaptive Control Design*, John Wiley and Sons, New York, 1995.

[8] M. Krstić, J. Protz, J. D. Paduano, and P. Kokotovic, *Backstepping design for jet engine stall and surge control*, in Proc. 1995 Conference on Decision and Control, 1995, pp. 3049–3055.

[9] H.-H. Wang, M. Krstić, and M. Larsen, *Control of deep hysteresis aeroengine compressors*, J. Dynam. Systems, Measurement, Control, accepted.

[10] D. Liaw and E. Abed, *Active control of compressor stall inception: A bifurcation theoretic approach*, Automatica, 32 (1996), pp. 109–115.

[11] C. A. Mansoux, D. L. Gysling, J. D. Setiawan, and J. D. Paduano, *Distributed nonlinear modeling and stability analysis of axial compressor stall and surge*, in Proc. American Control Conference, 1994, pp. 2305–2316.

[12] F. McCaughan, *Bifurcation analysis of axial flow compressor stability*, SIAM J. Appl. Math., 50 (1990), pp. 1232–1253.

[13] R. McOwen, *Partial Differential Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1996.

[14] I. Mezić, H. A. Hauksson, and A. Banaszuk, *Nonlinear Dynamics of the Moore–Greitzer Compression System Model*, 1997, preprint.

[15] F. Moore and E. Greitzer, *A theory of post-stall transients in axial compression systems: Part 1—Development of equations*, ASME J. Engrg. for Gas Turbines and Power, 108 (1986), pp. 68–76.

[16] R. Sepulchre and P. Kokotovic, *Shape signifiers for control of surge and stall in jet engines*, IEEE Trans. Automat. Control, 43 (1998), p. 1643.

SIAM J. Control Optim.
Vol. 37, No. 5, pp. 1538–1552

TWO-DIMENSIONAL PROPER RATIONAL MATRICES AND CAUSAL INPUT/OUTPUT REPRESENTATIONS OF TWO-DIMENSIONAL BEHAVIORAL SYSTEMS*

MARIATERESA NAPOLI† AND SANDRO ZAMPIERI‡

Abstract. The concept of proper rational matrix is strictly connected with the representation of causal transfer matrices. In the two-dimensional (2D) case there is much freedom in defining proper rational matrices. This freedom is connected to the fact that past and future in the 2D case can be determined by a 2D cone. In this way the concept of rational matrix which is proper with respect to a cone can be introduced. Moreover, an algorithm that checks the properness of a rational matrix is proposed. Finally, this algorithm is used for determining all possible causal input/output (I/O) representations of a behavior given by a kernel representation.

Key words. two-dimensional, behavioral approach, input/output representation, proper rational matrices, causality, cones

AMS subject classifications. 93A30, 93B25

PII. S0363012998338181

1. Introduction.

In the behavioral approach, a dynamical system is essentially described through the set of its admissible trajectories, without making any a priori distinction between input and output variables and without setting any causality relation between them.

This distinction, which is the characteristic feature of input/output (I/O) models, can be performed a posteriori, introducing the concept of free variables that are called in this way because their value can be arbitrarily assigned. As a consequence we have that, at least for the class of autoregressive (AR) systems, we can extract an I/O description [12, 9, 15, 13], starting from a behavioral model.

The first question that naturally arises when dealing with I/O descriptions is how to define causality. In case of discrete two-dimensional (2D) systems, which is the one we are interested in, the matter is complex, since the plane $\mathbb{Z}^2$ lacks a natural total ordering. As a consequence, the choice of the causality cone $\mathcal{C}$ is not as straightforward as in the one-dimensional (1D) case. In the classical I/O approach [4], the only admissible causality cone is $\mathcal{C} = \mathbb{N}^2$, so that causality is synonymous with quarter plane causality. In this paper we consider an extension of this notion of causality by assuming that $\mathcal{C}$ is an arbitrary cone in $\mathbb{Z}^2$.

The characterization of causality of 2D systems is based on the concept of the 2D proper rational matrix. This concept has been introduced and analyzed for a particular class of cones in [14, 3]. The characterization of 2D proper rational matrices allowed us to obtain some interesting existence results regarding causal I/O representations of 2D behavioral systems. The aim of this paper is to investigate the causal I/O representation of 2D behavioral systems in another direction. More precisely, starting from the kernel representation of a 2D behavioral system, we want to obtain

*Received by the editors May 1, 1998; accepted for publication (in revised form) December 2, 1998; published electronically August 2, 1999.
http://www.siam.org/journals/sicon/37-2/33818.html

†Department of Mechanical Engineering, University of California, Santa Barbara, CA 93106 (napoli@engineering.ucsb.edu).

‡Dipartimento di Elettronica ed Informatica, Università di Padova, Via Gradenigo 6/a, 35131 Padova, Italy (zampi@paola.dei.unipd.it).

an efficient method for determining all the causal relations between the variables of the system, given in terms of the set of all causality cones. This result provides a full characterization of the causality structure of the behavioral system. This problem is solved by extending the concept of the 2D proper rational matrix to general cones and by finding a suitable characterization of this class of rational matrices.

2. 1D proper rational matrices.

In this section we will recall some basic definitions and results on proper rational matrices in the 1D case (see [7, 6]).

In this paper we will consider only polynomials having real coefficients. Notice, however, that all the results we will present hold true for any field. A polynomial $p(z)$, in which we allow also negative powers of the indeterminates, is called a Laurent polynomial and can always be written as

$$p(z) = \sum_{i=n}^{N} p_i z^i,$$

where $n \leq N$ are suitable integers. The coefficient p_0 is called zero-degree coefficient of the polynomial p. The set of all the Laurent polynomials has a ring structure with respect to the usual addition and multiplication and it is denoted by the symbol $\mathbb{R}[z, z^{-1}]$. The rings $\mathbb{R}[z]$ and $\mathbb{R}[z^{-1}]$ are both subrings of $\mathbb{R}[z, z^{-1}]$. Consider, moreover, the ring $\mathbb{R}[[z]]$ of formal power series

$$s(z) = \sum_{i=0}^{+\infty} s_i z^i,$$

and define finally the field of rational functions

$$\mathbb{R}(z) := \left\{ \frac{q(z)}{p(z)} : q(z), p(z) \in \mathbb{R}[z] \text{ and } p(z) \neq 0 \right\},$$

which is the field of fractions of $\mathbb{R}[z]$ (see [1]). It is easy to verify that, up to isomorphism, $\mathbb{R}(z)$ coincides with the field of fractions of $\mathbb{R}[z, z^{-1}]$.

Definition 1. *A rational function $h \in \mathbb{R}(z)$ is said to be proper if there exist $p, q \in \mathbb{R}[z]$ such that $h = q/p$ and the zero-degree coefficient of p is nonzero.*

Notice, moreover, that in this paper the role of the indeterminates z and z^{-1} is inverted with respect to the standard notation used in most system theory books (see [6]). We prefer to follow the less standard notation proposed in [7, 5], because it is more convenient in the 2D case as we will see below (see also [4]).

We give now a theorem providing several equivalent characterizations of proper rational functions. The equivalence of these characterizations is easy to verify (see the first part of Chapter 2 in [7]).

Theorem 2. *Let $h \in \mathbb{R}(z)$. The following facts are equivalent:*

1. *h is proper.*

2. *There exists a unique formal power series $y \in \mathbb{R}[[z]]$ such that for all $p, q \in \mathbb{R}[z]$ such that $h = q/p$ we have that*

$$py = q.$$

3. *Let $p, q \in \mathbb{R}[z, z^{-1}]$ be coprime polynomials such that $h = q/p$. Then there exists $n \in \mathbb{Z}$ such that*

(a) $\hat{p} := z^n p, \quad \hat{q} := z^n q \in \mathbb{R}[z]$.

(b) *The zero-degree coefficient of $\hat{p}$ is nonzero.*

4. *Let $p, q \in \mathbb{R}[z]$ coprime in $\mathbb{R}[z]$ be such that $h = q/p$. Then the zero-degree coefficient of p is nonzero.*

Remarks. Notice that condition 1, which corresponds to the definition of a proper rational function, is an existence statement and thus does not give an algorithmic check of properness. Condition 2 connects proper rational functions with formal power series and so with causal impulse responses. Conditions 3 and 4 provide algorithmic checks of properness in the different rings $\mathbb{R}[z, z^{-1}]$ and $\mathbb{R}[z]$, which in this case are slightly different. The distinction between these two properties will be useful in the 2D case.

Now we consider the matrix case. A polynomial matrix P can be considered both as a matrix with polynomial entries and as a polynomial having matrix coefficients. This is the reason why it makes sense to introduce the concept of the degree-zero coefficient of a polynomial matrix that is in this case a matrix.

Definition 3. *A rational matrix $H \in \mathbb{R}(z)^{h \times m}$ is said to be proper if its entries are proper rational functions.*

We give also in the matrix case a theorem that is similar to the previous one and that provides several equivalent characterizations of a proper rational matrix. The characterization of properness for 1D rational matrices is usually given in terms of row proper matrix fractions (see [6]). The characterization that we will give below is based on coprime matrix fractions. This characterization is known [7], but it is less classical and for this reason we will give a brief proof of this result. The convenience of this characterization compared with the characterization in terms of the row proper matrix fractions is motivated by the fact that the extension of the concept of the row proper matrix fraction to the 2D polynomial matrices is rather involved, while the extension of the concept of coprime matrix fraction to the 2D case is straightforward [8].

Theorem 4. *Let $H \in \mathbb{R}(z)^{h \times m}$. The following facts are equivalent:*

1. *H is proper.*

2. *There exist $P \in \mathbb{R}[z]^{h \times h}$ and $Q \in \mathbb{R}[z]^{h \times m}$ such that $H = P^{-1}Q$ and such that the degree-zero coefficient of P is an invertible square matrix.*

3. *There exists a unique formal power series $Y \in \mathbb{R}[[z]]^{h \times m}$ such that for all $P \in \mathbb{R}[z]^{h \times h}$ and $Q \in \mathbb{R}[z]^{h \times m}$ such that $H = P^{-1}Q$ we have that*

$$PY = Q.$$

4. *Let $P \in \mathbb{R}[z]^{h \times h}$ and $Q \in \mathbb{R}[z]^{h \times m}$ be left coprime polynomial matrices such that $H = P^{-1}Q$. Then the degree-zero coefficient of P is an invertible square matrix.*

Proof. ($1 \Rightarrow 3$) By definition and by Theorem 2, condition 2, we know that if for $i = 1, \dots, h$ and $j = 1, \dots, m$ the polynomials $f_{ij}, g_{ij} \in \mathbb{R}[z]$ are such that $H = [p_{ij}/q_{ij}]$, then there exist $y_{ij} \in \mathbb{R}[[z]]$ such that $p_{ij} y_{ij} = q_{ij}$. Let $p := \prod p_{ij}$ and let $\bar{Q} := [p q_{ij}/p_{ij}] \in \mathbb{R}[z]^{h \times m}$ so that $H = \bar{Q}/p$. Let $P \in \mathbb{R}[z]^{h \times h}$ and $Q \in \mathbb{R}[z]^{h \times m}$ be such that $H = P^{-1}Q$. Then we have that

$$pQ = P\bar{Q} = pPY,$$

which, from the fact that $\mathbb{R}[[z]]$ is a domain, implies that $Q = PY$. We show finally the uniqueness of Y. Suppose that there exist $\hat{P} \in \mathbb{R}[z]^{h \times h}$, $\hat{Q} \in \mathbb{R}[z]^{h \times m}$, and $\hat{Y} \in \mathbb{R}[[z]]^{h \times m}$ such that $\hat{P}\hat{Y} = \hat{Q}$ and $\hat{P}^{-1}\hat{Q} = H = P^{-1}Q$. Since $H = \bar{Q}/p$, we have that

$$\hat{P}\bar{Q} = p\hat{Q} = p\hat{P}\hat{Y}.$$

Notice, moreover, that $\bar{Q} = pY$ implies $\hat{P}\bar{Q} = p\hat{P}Y$ and so $p\hat{P}\hat{Y} = p\hat{P}Y$. Since $\hat{P}$ is nonsingular, this implies that $\hat{Y} = Y$.

($3 \Rightarrow 4$) Let $P \in \mathbb{R}[z]^{h\times h}$ and $Q \in \mathbb{R}[z]^{h\times m}$ be left coprime polynomial matrices such that $H = P^{-1}Q$. Then by condition 3 there exists $Y \in \mathbb{R}[[z]]^{h\times m}$ such that $PY = Q$. Moreover, coprimeness ensures the existence of polynomial matrices $A \in \mathbb{R}[z]^{h\times h}$ and $B \in \mathbb{R}[z]^{m\times h}$, which satisfy the Bezout identity $PA + QB = I$. This implies that $P(A + YB) = I$ and hence that the degree-zero coefficient of P must be an invertible matrix.

($4 \Rightarrow 2$) This is trivial.

($2 \Rightarrow 1$) This follows from the fact that $H = \mathrm{adj}(P)Q/\det(P)$ and from the fact that the degree-zero coefficient of $\det(P)$ is nonzero. □

Notice that the only condition that was valid in the scalar case and that is not valid any more in the matrix case is the one involving the primeness of Laurent polynomials. Condition 4 still provides an algorithmic check of properness in the matrix case together with condition 1 (the definition), which, translating matrix properness into scalar properness, shows another way to verify whether a rational matrix is proper or not.

3. Cones and 2D proper rational matrices.

In this section we will extend the notions of proper rational function and matrix to the 2D case. Some results in this direction can be found also in [14].

Before giving the definition of properness in the 2D case we need to introduce the notion of cone and of regular cone in $\mathbb{Z}^2$.

DEFINITION 5. *A cone $\mathcal{C}$ is a subset of $\mathbb{Z}^2$ such that there exists a pair of elements $d_1, d_2 \in \mathbb{Z}^2$ satisfying*

$$\mathcal{C} = \mathbb{Z}^2 \cap \{\alpha d_1 + \beta d_2 \in \mathbb{R}^2 \ : \ \alpha, \beta \in \mathbb{R},\ \alpha, \beta \geq 0\},$$

and such that the matrix $D \in \mathbb{Z}^{2\times 2}$, whose columns coincide with d_1 and d_2, is nonsingular, i.e., $\det(D) \neq 0$*. A cone $\mathcal{C}$ is said to be* regular *if there exists a pair of elements $d_1, d_2 \in \mathbb{Z}^2$ such that*

$$\mathcal{C} = \{\alpha d_1 + \beta d_2 : \alpha, \beta \in \mathbb{N}\}$$

and such that $\det(D) = \pm 1$*, where D is the matrix defined from d_1, d_2 as above.*

It can be shown that a regular cone $\mathcal{C}$ is always isomorphic to $\mathbb{N}^2$; i.e., it is possible to perform a change of coordinates in such a way that $\mathcal{C}$ coincides with $\mathbb{N}^2$. Moreover, given a cone $\mathcal{C}$, there is always a regular cone $\mathcal{C}_r$ containing $\mathcal{C}$. Actually, it is easy to prove that, up to a change of coordinates, there is no loss of generality not only in assuming that any cone is contained in $\mathbb{N}^2$, but also in supposing that it is specified as

$$\mathcal{C} = \{(i,j) \in \mathbb{N}^2 \ : \ j \leq mi\}, \tag{1}$$

where m is a suitable positive rational number.

Given a Laurent polynomial in two indeterminates

$$p(z_1, z_2) = \sum_{(i,j)\in S} p_{ij} z_1^i z_2^j,$$

where S is a finite subset of $\mathbb{Z}^2$, by $\mathrm{supp}(p)$ we mean the set of points $(i,j) \in \mathbb{Z}^2$ corresponding to nonzero coefficients of $p(z_1, z_2)$

$$\mathrm{supp}(p) = \{(i,j) \in \mathbb{Z}^2 \ : \ p_{ij} \neq 0\}.$$

Let $\mathcal{C}$ be a cone. With the symbol $\mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]_{\mathcal{C}}$ we mean the ring of polynomials whose support is contained in $\mathcal{C}$. Similar definitions can be immediately extended to polynomial matrices and power series. More precisely, with the symbol $\mathbb{R}[[z_1, z_2, z_1^{-1}, z_2^{-1}]]_{\mathcal{C}}$ we mean the ring of formal power series

$$s(z_1, z_2) = \sum_{(i,j)\in\mathcal{C}} s_{ij} z_1^i z_2^j.$$

Notice that $\mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]_{\mathcal{C}}$ is always a ring, but unless $\mathcal{C}$ is regular, this ring lacks many of the properties usually possessed by polynomial rings. (It can be seen, for instance, that it is not, in general, a unique factorization domain.)

For the sake of simplicity, from now on we will denote by $\mathbf{z}$ the pair (z_1, z_2). Consequently, we will use the following shorthand notations:

$$\mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]_{\mathcal{C}} := \mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}]_{\mathcal{C}}, \tag{2}$$

$$\mathbb{R}[[\mathbf{z}, \mathbf{z}^{-1}]]_{\mathcal{C}} := \mathbb{R}[[z_1, z_2, z_1^{-1}, z_2^{-1}]]_{\mathcal{C}}, \tag{3}$$

$$\mathbb{R}(\mathbf{z}) := \mathbb{R}(z_1, z_2), \tag{4}$$

where the last notation denotes the field of rational functions in two indeterminates.

If we think of a polynomial matrix $A(z_1, z_2) \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]^{h\times m}$ as a polynomial with matrix coefficients, we can write it as

$$A(z_1, z_2) = \sum_{(i,j)\in S} A_{ij} z_1^i z_2^j, \tag{5}$$

where $A_{ij} \in \mathbb{R}^{h\times m}$ and S is a finite subset of $\mathbb{Z}^2$. By degree-zero coefficient of $A(z_1, z_2)$ we mean the matrix A_{00}.

Definition 6. *A* 2D *rational function $h \in \mathbb{R}(\mathbf{z})$ is said to be proper with respect to a cone $\mathcal{C}$ if there exist $p, q \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]_{\mathcal{C}}$ such that $h = q/p$ and the zero-degree coefficient of p is nonzero.*

Now we give a theorem providing several equivalent characterizations of 2D proper rational functions. These characterizations provide the extension to the 2D case of the analogous results valid in the 1D case presented in Theorem 2. Observe, moreover, that the theorem that follows has already been proved for regular cones in [14, Lemma 3].

Theorem 7. *Let $h \in \mathbb{R}(\mathbf{z})$ and let $\mathcal{C}$ be any cone in $\mathbb{Z}^2$. The following facts are equivalent:*

1. *h is proper with respect to $\mathcal{C}$.*
2. *There exists a unique formal power series $y \in \mathbb{R}[[\mathbf{z}, \mathbf{z}^{-1}]]_{\mathcal{C}}$ such that for all $p, q \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]_{\mathcal{C}}$ such that $h = q/p$ we have that*

$$py = q.$$

3. *Let $p, q \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]$ be coprime polynomials such that $h = q/p$. Then there exists $n_1, n_2 \in \mathbb{Z}$ such that*
   (a) $\hat{p} := z_1^{n_1} z_2^{n_2} p, \quad \hat{q} := z_1^{n_1} z_2^{n_2} q \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]_{\mathcal{C}}$.
   (b) *The zero-degree coefficient of $\hat{p}$ is nonzero.*

Proof. ($3 \Rightarrow 1$) It is sufficient to notice that $h = \hat{q}/\hat{p}$, which, by the properties imposed on $\hat{p}$ and $\hat{q}$, implies that h is proper with respect to $\mathcal{C}$.

($1 \Rightarrow 2$) If h is proper with respect to $\mathcal{C}$, then there exist polynomials

$$\hat{p} = \sum_{(i,j)\in\mathcal{C}} \hat{p}_{ij} z_1^i z_2^j, \quad \hat{q} = \sum_{(i,j)\in\mathcal{C}} \hat{q}_{ij} z_1^i z_2^j$$

such that $h = \hat{q}/\hat{p}$ and such that $\hat{p}_{00} \neq 0$. It is not restrictive to assume $\hat{p}_{00} = 1$. Let $y \in \mathbb{R}[[\mathbf{z}, \mathbf{z}^{-1}]]_{\mathcal{C}}$ be defined recursively as follows:

$$y_{hk} = -\sum_{\substack{(i,j)\in\mathcal{C}\\(i,j)\neq(0,0)}} \hat{p}_{ij} y(h-i, k-j) + \hat{q}_{hk}.$$

This equation implies that $\hat{p}y = \hat{q}$. Now let $p, q \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]_{\mathcal{C}}$ be such that $h = q/p$. Then we have $\hat{p}q = \hat{q}p = \hat{p}py$ and so, since $\mathbb{R}[[\mathbf{z}, \mathbf{z}^{-1}]]_{\mathcal{C}}$ is a domain, we can argue that $py = q$.

($2 \Rightarrow 3$) In the proof, we will explicitly suppose our cone to be specified as

$$\mathcal{C} = \left\{ (i,j) \in \mathbb{N}^2 \;:\; j \leq \frac{m_1}{m_2} i \right\},$$

where m_1, m_2 are coprime positive integers. This can be done without loss of generality.

Let $p, q \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]$ be coprime polynomials such that $h = q/p$. Then there exist $r_1, r_2 \in \mathbb{Z}$ such that $\hat{p} := z_1^{r_1} z_2^{r_2} p$, $\hat{q} := z_1^{r_1} z_2^{r_2} q \in \mathbb{R}[\mathbf{z}]$ and such that $\hat{p}, \hat{q}$ are coprime in $\mathbb{R}[\mathbf{z}]$. Using the fact that the thesis is true for regular cones [14, Lemma 3], we can argue that

$$\hat{p}y = \hat{q} \tag{6}$$

and $y \in \mathbb{R}[[\mathbf{z}]]$ imply that the zero-degree coefficient of $\hat{p}$ is nonzero. We want to show now that $\hat{p}, \hat{q} \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]_{\mathcal{C}}$.

Let $\bar{\mathcal{C}}$ be the smallest cone containing both $\mathcal{C}$ and the support of $\hat{p}$ (see Figure 1), and let $\bar{m}_1, \bar{m}_2$ be coprime positive integers such that

$$\bar{\mathcal{C}} = \left\{ (i,j) \in \mathbb{N}^2 \;:\; j \leq \frac{\bar{m}_1}{\bar{m}_2} i \right\}.$$

If we show that $\bar{\mathcal{C}} = \mathcal{C}$ or, equivalently, that $(m_1, m_2) = (\bar{m}_1, \bar{m}_2)$, then we are done. Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be two regular cones such that

$$\mathcal{C}_1 \cap \mathcal{C}_2 = \bar{\mathcal{C}}.$$

We can take (see Figure 2) $\mathcal{C}_1 = \mathbb{N}^2$ and $\mathcal{C}_2 = \{\alpha(\bar{m}_1, \bar{m}_2) + \beta(\bar{l}_1, \bar{l}_2) \;:\; \alpha, \beta \in \mathbb{N}\}$, where $\bar{l}_1 = l_1 - k\bar{m}_1, \bar{l}_2 = l_2 - k\bar{m}_2$ and where $l_2\bar{m}_1 - l_1\bar{m}_2 = -1$ and k is a big enough positive integer. Observe that, since $\mathcal{C}_2$ contains the supports of both $\hat{p}$ and y, $\mathcal{C}_2$ contains also the support of $\hat{q}$.

Perform a change of coordinates transforming $\mathcal{C}_2$ into $\mathbb{N}^2$. After this change of coordinates $\hat{p}, \hat{q}$ are still coprime polynomials in $\mathbb{R}[\mathbf{z}]$ and y is still in $\mathbb{R}[[\mathbf{z}]]$. Since $\hat{p}, \hat{q}$ are coprime in $\mathbb{R}[\mathbf{z}]$, there exists [8] $a, b \in \mathbb{R}[\mathbf{z}]$ such that $a\hat{p} + b\hat{q} = \psi \in \mathbb{R}[z_1]$. This fact together with (6) yields $\hat{p}\hat{y} = \psi$, where $\hat{y} := a + by \in \mathbb{R}[[\mathbf{z}]]$. Observe that, if we consider $\hat{p}, a, b, y, \hat{y}$ as polynomials or power series in z_1 having polynomials or power series in z_2 as coefficients, we have that

$$\hat{y} = \sum \hat{y}_h(z_2) z_1^h = \sum a_h(z_2) z_1^h + \left(\sum y_i(z_2) z_1^i\right)\left(\sum b_j(z_2) z_1^j\right),$$

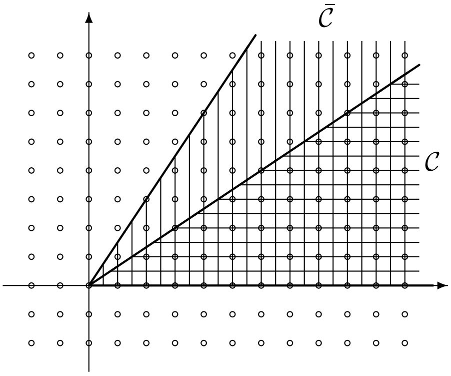


FIG. 1

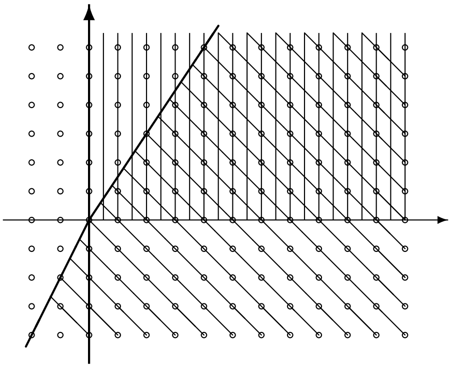


FIG. 2

and so we realize that

$$\hat{y}_h(z_2) = a_h(z_2) + \sum y_{h-i}(z_2) b_i(z_2).$$

This implies that

$$\left(\sum \hat{p}_i(z_2) z_1^i\right)\left(\sum \hat{y}_j(z_2) z_1^j\right) = \sum_{k=l}^{L} \psi_k z_1^k,$$

where $\psi_k \in \mathbb{R}$ and where we can assume that $l \in \mathbb{N}$ is such that $\psi_l \neq 0$. Now, by observing that $\hat{p}_0(z_2) \neq 0$, we can argue that

$$\hat{p}_0(z_2)\hat{y}_l(z_2) = \psi_l \in \mathbb{R} \setminus \{0\}. \tag{7}$$

Assume now by contradiction that $\bar{\mathcal{C}} \neq \mathcal{C}$. This has two consequences. On one hand this implies that the support of $\hat{p}_0(z_2)$ includes at least two points; on the other hand we have that all the coefficients $y_i(z_2)$ of the power series y and consequently also all the coefficients $\hat{y}_i(z_2)$ of the power series $\hat{y}$ are polynomials in z_2. These facts are in contradiction with (7). □

Remark. Notice that condition 4 of Theorem 2 does not extend to the 2D case for general cones in $\mathbb{Z}^2$. It can be seen that [14, Lemma 3] this extension holds true when the cone is regular. Consequently, for general cones we have that condition 3 provides the only way to check algorithmically the properness of a 2D rational function.

Notice, moreover, that the proof of the previous theorem is more difficult than the proof of the analogous result for regular cones. The reason is that for regular cones the ring $\mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]_{\mathcal{C}}$ is isomorphic to the ring $\mathbb{R}[\mathbf{z}]$ of polynomials in two variables, which has many nice properties such as a Bezout equation-like condition for coprimeness. When the cone $\mathcal{C}$ is not regular, the ring $\mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]_{\mathcal{C}}$ does not possess such properties any more. (It is not a unique factorization domain, and even the concept of coprime polynomials is not well defined.) The key idea in the proof of the previous theorem is that any cone is the intersection of two regular cones. In this way we can use the results that are known for regular cones for proving this theorem.

We consider now the matrix case.

DEFINITION 8. *A* 2D *rational matrix $H \in \mathbb{R}(\mathbf{z})^{h \times m}$ is said to be proper with respect to a cone $\mathcal{C}$ if its entries are* 2D *rational functions that are proper with respect to $\mathcal{C}$.*

We give also in this case a theorem providing several equivalent characterizations of a 2D proper rational matrix.

THEOREM 9. *Let $H \in \mathbb{R}(\mathbf{z})^{h\times m}$. The following facts are equivalent:*

1. *H is proper with respect to a cone $\mathcal{C}$.*
2. *There exist $P \in \mathbb{R}[\mathbf{z},\mathbf{z}^{-1}]_{\mathcal{C}}^{h\times h}$ and $Q \in \mathbb{R}[\mathbf{z},\mathbf{z}^{-1}]_{\mathcal{C}}^{h\times m}$ such that $H = P^{-1}Q$ and such that the degree-zero coefficient of P is an invertible square matrix.*
3. *There exists a unique formal power series $Y \in \mathbb{R}[[\mathbf{z},\mathbf{z}^{-1}]]_{\mathcal{C}}^{h\times m}$ such that for all $P \in \mathbb{R}[\mathbf{z},\mathbf{z}^{-1}]_{\mathcal{C}}^{h\times h}$ and $Q \in \mathbb{R}[\mathbf{z},\mathbf{z}^{-1}]_{\mathcal{C}}^{h\times m}$ such that $H = P^{-1}Q$ we have that*

$$PY = Q.$$

Proof. ($2 \Rightarrow 1 \Rightarrow 3$) These implications can be shown in the same way as we proved the same implications in Theorem 4. Notice that uniqueness again follows from the fact that $\mathbb{R}[\mathbf{z},\mathbf{z}^{-1}]_{\mathcal{C}}$ is a domain.

($3 \Rightarrow 2$) Let $P \in \mathbb{R}[\mathbf{z},\mathbf{z}^{-1}]_{\mathcal{C}}^{h\times h}$ and $Q \in \mathbb{R}[\mathbf{z},\mathbf{z}^{-1}]_{\mathcal{C}}^{h\times m}$ be such that $H = P^{-1}Q$. Then condition 3 ensures the existence of $Y \in \mathbb{R}[[\mathbf{z},\mathbf{z}^{-1}]]_{\mathcal{C}}^{h\times m}$ such that $PY = Q$. If $Y = [y_{ij}]$ and if we denote $p := \det(P)$ and $\bar{Q} := \operatorname{adj}(P)Q = [\bar{q}_{ij}]$, then we have that $py_{ij} = \bar{q}_{ij}$ and so, by Theorem 7, we argue that $h_{ij} = \bar{q}_{ij}/p$ is proper and hence that H is proper with respect to $\mathcal{C}$. □

The definition of 2D properness, by translating matrix properness into scalar properness, provides in this case the only way to verify algorithmically whether a rational matrix is proper or not. An efficient algorithmic check can be done as follows.

ALGORITHM. Given a rational matrix $H \in \mathbb{R}(\mathbf{z})^{h\times m}$.

Step 1. Represent it as $H = [q_{ij}/p_{ij}]$, where $q_{ij}, p_{ij} \in \mathbb{R}[\mathbf{z},\mathbf{z}^{-1}]$ are coprime.

Step 2. Let p be the least common multiple of the p_{ij} and let $\bar{q}_{ij} := q_{ij}p/p_{ij}$ so that $H = [\bar{q}_{ij}/p]$.

Step 3. We have that H is proper with respect to a cone $\mathcal{C}$ if and only if there exists $n_1, n_2 \in \mathbb{Z}$ such that

(a) $\hat{p} := z_1^{n_1} z_2^{n_2} p,\quad \hat{q}_{ij} := z_1^{n_1} z_2^{n_2} \bar{q}_{ij} \in \mathbb{R}[\mathbf{z},\mathbf{z}^{-1}]_{\mathcal{C}}$.

(b) The zero-degree coefficient of $\hat{p}$ is nonzero.

Proof of the algorithm. One direction of the proof is easy. Suppose conversely that H is proper. This implies that there exist monomials m_{ij} in z_1, z_2 such that $m_{ij}p_{ij}, m_{ij}q_{ij} \in \mathbb{R}[\mathbf{z},\mathbf{z}^{-1}]_{\mathcal{C}}$ and the zero-degree coefficients of $m_{ij}p_{ij}$ are nonzero. This implies that the polynomials p_{ij} belong to the set

$$\begin{aligned}\mathbf{U} \quad := \quad & \{g \in \mathbb{R}[\mathbf{z},\mathbf{z}^{-1}] \ : \ \exists\ h_1, h_2 \in \mathbb{Z}, z_1^{h_1} z_2^{h_2} g \in \mathbb{R}[\mathbf{z},\mathbf{z}^{-1}]_{\mathcal{C}} \\ & \qquad \text{and zero-degree coefficient of } z_1^{h_1} z_2^{h_2} g \text{ is nonzero}\}.\end{aligned}$$

It is easy to see that this set is a multiplicative set. It is less straightforward to show that it is saturated so that we have $p, q \in \mathbf{U}$ if and only if $pq \in \mathbf{U}$ [1]. This implies that the least common multiple p of p_{ij} is still in $\mathbf{U}$ and so there exists $n_1, n_2 \in \mathbb{Z}$ such that $\hat{p} := z_1^{n_1} z_2^{n_2} p$ and the zero-degree coefficient of $\hat{p}$ is nonzero. Observe finally that $m_{ij}p_{ij}$ divides $\hat{p}$ and that $\hat{p}/m_{ij}p_{ij} \in \mathbb{R}[\mathbf{z},\mathbf{z}^{-1}]_{\mathcal{C}}$. This implies that

$$\hat{q}_{ij} = z_1^{n_1} z_1^{n_2} \bar{q}_{ij} = \hat{p}\frac{q_{ij}}{p_{ij}} = \hat{p}\frac{m_{ij}q_{ij}}{m_{ij}p_{ij}} \in \mathbb{R}[\mathbf{z},\mathbf{z}^{-1}]_{\mathcal{C}}.$$

Remark. In Step 2 we can take as polynomial p any common multiple of the p_{ij} as, for instance, $p = \prod p_{ij}$. Notice, however, that in general the least common multiple is more convenient since often it has smaller support and, moreover, it can be computed efficiently (see [2]).

4. 2D systems in the behavioral approach.

In the remaining part of the paper we want to use the characterization of 2D proper rational matrices given in the previous section for the analysis of the causality structure of a 2D behavioral system given through a kernel representation. We start by giving a short introduction to the theory of 2D systems in the behavioral approach.

It is known that, given a dynamical system, we can associate with it different mathematical models, according to the aim the model was constructed for and to the theoretical approach that has been chosen. When using behavioral models, a dynamical system is characterized by the set of trajectories that constitute the so-called behavior of the system. More precisely, in this setup a dynamical system is described by a triple

$$\Sigma = (T, W, \mathcal{B}),$$

where T is the time domain, W is the signal alphabet, and $\mathcal{B} \subset W^T$, the *behavior*, is the set of admissible trajectories. For 2D systems we assume that $T = \mathbb{Z}^2$ and $W = \mathbb{R}^q$. We refer the interested reader to [9, 10, 11] for a more complete introduction to 2D behavioral systems theory.

An important subclass of 2D systems is constituted by the so-called AR 2D systems. They are 2D systems whose behavior is given by the set of solutions $w \in (\mathbb{R}^q)^{\mathbb{Z}^2}$ (set of all q-dimensional signals defined on $\mathbb{Z}^2$) of a linear difference equation of the following kind:

$$\sum_{(i,j)\in S} R_{ij} w(h+i, k+j) = 0 \qquad \forall (h,k) \in \mathbb{Z}^2, \tag{8}$$

where $R_{ij} \in \mathbb{R}^{l\times q}$ and S is a finite subset of $\mathbb{Z}^2$. Notice that any polynomial matrix

$$R = \sum_{(i,j)\in S} R_{ij} z_1^i z_2^j \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]^{l\times q}$$

naturally induces a polynomial linear operator

$$R(\sigma_1, \sigma_2) \ : \ (\mathbb{R}^q)^{\mathbb{Z}^2} \longrightarrow (\mathbb{R}^l)^{\mathbb{Z}^2}$$

in the following way:

$$(R(\sigma_1, \sigma_2) w)(h,k) = \sum R_{ij} w(h+i, k+j) \qquad \forall (h,k) \in \mathbb{Z}^2.$$

In this way we have that the behavior $\mathcal{B}$ determined by the difference equation (8) coincides with $\ker R(\sigma_1, \sigma_2)$ and that the behavior of an AR system can always be represented as the kernel of a polynomial linear operator, which is called *kernel representation.*

5. Passing from kernel to I/O representations.

Given a behavioral model of a dynamical system, we could wonder whether an I/O representation of the same system can be obtained or not. By answering this question we can check whether there exists a cause-effect relation between the components of the signal.

Roughly speaking, if the constraints imposed by (8) are few with respect to the number of components of the signal, some of them can be considered as inputs. In fact, under this assumption, their value is arbitrarily assignable and determines the value of the remaining components.

The mathematical translation of this intuitive consideration is a rank condition on the polynomial matrix R providing the kernel representation of the system. It can be proved (see [9, 13, 15]) that if

$$\operatorname{rank} R(z_1, z_2) = h,$$

then it is possible to split the components of w in $m := q - h$ inputs (free variables) and h outputs (nonfree variables). More precisely, if S is any permutation matrix such that

$$RS = [P| - Q],$$

where $P \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]^{l\times h}$, $Q \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]^{l\times(q-h)}$, and rank $P = h$, then we say that the pair of matrices (P, Q) provides an I/O representation of the system because they satisfy the properties of the following definition.

Definition 10 (see [9, 14]). *Given a* 2D *AR system* $\Sigma(\mathbb{Z}^2, \mathbb{R}^q, \ker R(\sigma_1, \sigma_2))$, *the difference equation*

$$P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u, \tag{9}$$

where $h + m = q$, $P \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]^{l\times h}$, $Q \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]^{l\times m}$, *and where* $y \in (\mathbb{R}^h)^{\mathbb{Z}^2}$ *and* $u \in (\mathbb{R}^m)^{\mathbb{Z}^2}$, *is an* I/O representation *of* Σ *if*

1. $\mathcal{B} = \{S[\begin{smallmatrix} y \\ u \end{smallmatrix}] : P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u\}$, *where* S *is a suitable* $q\times q$ *permutation matrix;*
2. *u is free, i.e., for all* $u \in (\mathbb{R}^m)^{\mathbb{Z}^2}$ *there exists* $y \in (\mathbb{R}^h)^{\mathbb{Z}^2}$ *such that* (9) *holds;*
3. *no other component in y is free.*

We often use the shorthand notation (P, Q) to denote the I/O representation (9). Observe that, starting from an AR behavioral model, it is possible to extract finitely many different I/O descriptions. They are obtained by choosing different permutation matrices S that satisfy only the rank condition. In other words, they are obtained selecting in different ways the inputs and the outputs among the components of w.

The concept of causality is strictly related to I/O representations. In the 2D case its definition is more involved than for 1D systems, since there are different possible ways to order the time domain $T = \mathbb{Z}^2$. As a consequence, there is more freedom in the choice of the causality cone. Given a cone $\mathcal{C}$, by the symbol $(\mathbb{R}^m)^{\mathbb{Z}^2}_{\mathcal{C}}$ we mean the set of all m-dimensional signals defined on $\mathbb{Z}^2$ and supported in $\mathcal{C}$.

Definition 11. *The I/O representation* (9) *is said to be causal with respect to the cone* $\mathcal{C}$ *if for any* $u \in (\mathbb{R}^m)^{\mathbb{Z}^2}_{\mathcal{C}}$ *there exists* $y \in (\mathbb{R}^h)^{\mathbb{Z}^2}_{\mathcal{C}}$ *such that* (9) *holds.*

Notice that the definition above suggests that the influence of u on y is causal with respect to $\mathcal{C}$. In can be shown, moreover, [14, Lemma 1] that y in the previous definition is uniquely determined from u.

6. Characterization of causal I/O representations. In [14], a characterization of causal I/O representations with respect to regular cones has been given. Our aim here is to extend and generalize those results to general cones. Some of these results can be generalized in a straightforward way. This is the case for Proposition 3 [14], which will be used next. This proposition, stated for regular cones, guarantees that the causality of an I/O representation

$$P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u \tag{10}$$

depends only on a coprime representation of the polynomial matrices specifying the system. Thus, if $\bar{P} \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]^{h\times h}$ and $\bar{Q} \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]^{h\times m}$ are coprime polynomial matrices such that

$$P = F\bar{P}, \quad Q = F\bar{Q},$$

with F a full column rank polynomial matrix of suitable dimensions, then (10) is causal with respect to a regular cone $\mathcal{C}_r$ if and only if

$$\bar{P}(\sigma_1, \sigma_2)y = \bar{Q}(\sigma_1, \sigma_2)u$$

is causal with respect to it. It is easy to see that the proof still holds if we consider general cones.

Let (P, Q) be an I/O representation of a 2D AR system that is causal with respect to a cone $\mathcal{C}$. Define the inputs $\delta^{(i)}$, $i = 1, \ldots, m$, as

$$\delta^{(i)}(t) := \begin{cases} e_i, & t = (0,0), \\ 0 & \text{otherwise}, \end{cases}$$

where e_i is the ith vector of the canonical base in $\mathbb{R}^m$. If $y^{(i)} \in (\mathbb{R}^h)^{\mathbb{Z}^2}_{\mathcal{C}}$ is the corresponding output, namely,

$$P(\sigma_1, \sigma_2)y^{(i)} = Q(\sigma_1, \sigma_2)\delta^{(i)}, \tag{11}$$

we define the impulse response of the 2D system to be the matrix-valued sequence

$$Y := [y^{(1)} \ldots y^{(m)}] \in (\mathbb{R}^{h\times m})^{\mathbb{Z}^2}_{\mathcal{C}}.$$

It is worth pointing out that, as shown in [14], the causality of an I/O representation is equivalent to the existence of the impulse response, since the impulse response determines the way in which the system maps input signals supported in $\mathcal{C}$ into output y by the convolution

$$y(h, k) := \sum_{(i,j)\in\mathbb{Z}^2} Y(h - i, k - j)u(i, j).$$

Notice that, since u and Y are both supported in $\mathcal{C}$, the sum is always finite and, moreover, also the support of y is included in $\mathcal{C}$.

Now we are in a position to state the following theorem, which allows us to characterize the causality structure of a 2D AR system.

THEOREM 12. *Let*

$$P(\sigma_1, \sigma_2)y = Q(\sigma_1, \sigma_2)u, \tag{12}$$

with $P \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]^{l\times h}$ *and* $Q \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]^{l\times m}$, *be an I/O representation of a* 2D *AR system. Then* (12) *is causal with respect to a cone* $\mathcal{C}$ *if and only if the rational matrix* $H \in \mathbb{R}(\mathbf{z})^{h\times m}$ *such that* $Q = PH$ *is proper with respect to the cone* $-\mathcal{C}$.

Proof. Let $\bar{P} \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]^{h\times h}$ and $\bar{Q} \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]^{h\times m}$ be coprime polynomial matrices such that $P = F\bar{P}, Q = F\bar{Q}$, where F is a full column rank polynomial matrix of suitable dimensions, then, as mentioned above, (12) is causal with respect to $\mathcal{C}$ if and only if

$$\bar{P}(\sigma_1, \sigma_2)y = \bar{Q}(\sigma_1, \sigma_2)u \tag{13}$$

is causal with respect to $\mathcal{C}$. Observe that $H = \bar{P}^{-1}\bar{Q}$ and that, moreover, it is not restrictive to assume that $\bar{P}, \bar{Q}$ have entries in $\mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]_{-\mathcal{C}}$. Notice that (13) is causal with respect to $\mathcal{C}$ if and only if there exists the impulse response, which is a matrix-valued sequence $Y \in (\mathbb{R}^{h\times m})^{\mathbb{Z}}_{\mathcal{C}}$ satisfying the matrix difference equation

$$\sum_{ij} \bar{P}_{ij} Y(h+i, k+j) = \bar{Q}_{-h,-k} \quad \forall (h,k) \in \mathbb{Z}^2.$$

If we define the power series $\bar{Y} := \sum_{ij} Y(i,j) z_1^{-i} z_2^{-j} \in \mathbb{R}[[\mathbf{z}, \mathbf{z}^{-1}]]^{h\times m}_{-\mathcal{C}}$, then we have that $\bar{Q} = \bar{P}\bar{Y}$. By Theorem 9 this is equivalent to the fact that H is proper with respect to $-\mathcal{C}$. □

Remarks. Notice that by using Theorem 9 the proof of the previous theorem is more direct than the proof of the analogous result [14, Theorem 1] for regular cones. In both cases the main idea is to characterize properness of a rational matrix in terms of the existence of a power expansion that is supported in the cone. For regular cones this has been done by using the properties of coprime matrix fraction descriptions of rational matrices. However, this method works only for regular cones. This difficulty has been overcome in Theorem 9 simply by using the definition of proper rational matrices, which is in terms of the properness of its scalar entries. This method, which obviously works also for regular cones, is simpler and more natural than the technique used in [14, Theorem 1].

7. Minimal causality cones and parametrization of causal I/O representations. Consider an I/O representation (P, Q). The theorem we proved in the previous section allows us to determine the set of all cones $\mathcal{C}$ such that (P, Q) is causal with respect to $\mathcal{C}$. These cones are called *causality cones* for the I/O representation. Notice that, if $\mathcal{C}$ is a causality cone and $\mathcal{C}' \supseteq \mathcal{C}$, then also $\mathcal{C}'$ is a causality cone. Therefore, the set of causality cones is completely determined by its finite subset $\mathcal{M}(P, Q)$ constituted by the *minimal* causality cones.

In practice the construction of this set reduces to a simple procedure based on the previous theorem. Let $H \in \mathbb{R}(\mathbf{z})^{h\times m}$ be the rational matrix such that $Q = HP$ and represent it as $H = [q_{ij}/p_{ij}]$, where $q_{ij}, p_{ij} \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]$ are coprime. Let p be the least common multiple of p_{ij} and $\bar{q}_{ij} := q_{ij}p/p_{ij}$ so that $H = [\bar{q}_{ij}/p]$. As suggested in the algorithmic check of properness proposed above, H is proper with respect to a cone $\mathcal{C}$ if and only if there exist $n_1, n_2 \in \mathbb{Z}$ such that $\hat{p} := z_1^{n_1} z_2^{n_2} p$, $\hat{q}_{ij} := z_1^{n_1} z_2^{n_2} \bar{q}_{ij} \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]_{\mathcal{C}}$ and the zero-degree coefficient of $\hat{p}$ is nonzero. For this reason the finite set of minimal causality cones can be obtained from the polynomials p and $\bar{q}_{ij}$ in the following way:

Step 1. Determine the convex hull of supp (p) and from this the finite set $V = \{v_1, \ldots, v_k\}$ of the vertices of this convex hull.

Step 2. For each $v_i \in V$ consider the following set of cones:

$$\mathcal{C}(v_i) = \left\{ \mathcal{C} \ : \ v_i - \mathcal{C} \supseteq \text{supp } (p) \cup \bigcup_{ij} \text{supp } (\bar{q}_{ij}) \right\}.$$

Step 3. It is clear that, when the set $\mathcal{C}(v_i)$ is nonempty, it contains a cone $\hat{\mathcal{C}}_i$ that is smaller than every other cone in $\mathcal{C}(v_i)$. Then by Theorem 12 the set $\mathcal{M}(P, Q)$ of the minimal causality cones for the I/O representation (P, Q) coincides with the set of all the cones $\hat{\mathcal{C}}_i$.

It may happen that, for a given I/O representation (P, Q), the set $\mathcal{M}(P, Q)$ is empty. However, there exists a certain freedom in constructing an I/O representation

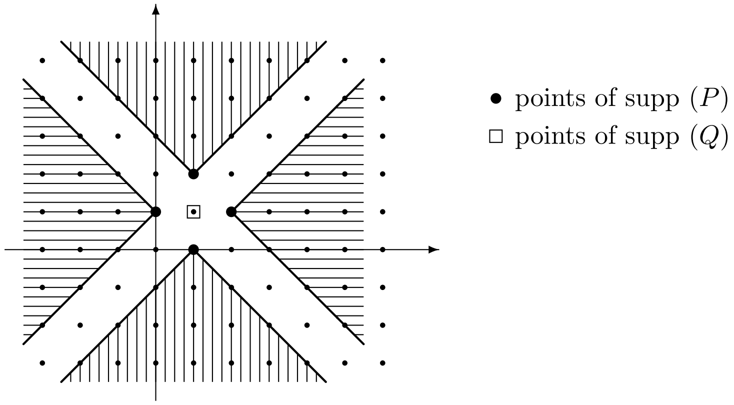


FIG. 3

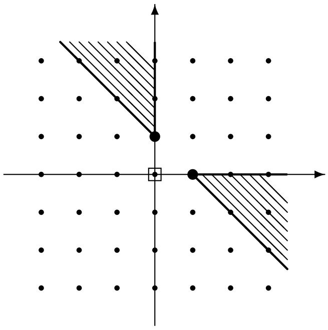


FIG. 4

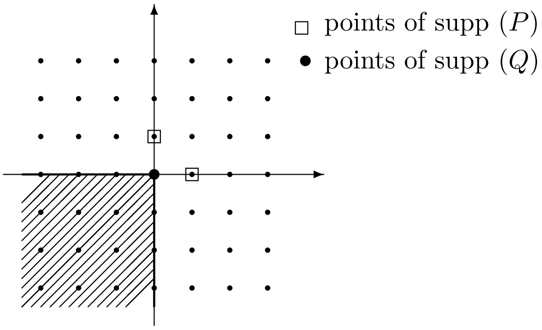


FIG. 5

from a kernel representation, which corresponds to the freedom that there exists in the choice of h linearly independent columns in a rank h polynomial matrix $R \in \mathbb{R}[\mathbf{z}, \mathbf{z}^{-1}]^{l \times q}$ providing the kernel representation of the AR system. The family of the sets $\mathcal{M}(P, Q)$, when (P, Q) varies in the set of all possible I/O representations of the AR system, provides a complete description of its causality structure. It is important to notice that, as a direct consequence of [14, Theorem 2], we have that there always exists an I/O representation (P, Q) such that $\mathcal{M}(P, Q)$ is nonempty.

Example 1. Let Σ be a 2D AR system whose behavior is the kernel of the polynomial matrix

$$R = [z_1 z_2 \mid -z_1 - z_2 - z_1^2 z_2 - z_1 z_2^2].$$

We can consider two I/O representations of Σ. If we let $P = z_1 z_2$ and $Q = z_1 + z_2 + z_1^2 z_2 + z_1 z_2^2$, we have that $\mathcal{M}(P, Q) = \emptyset$. If, conversely, we let $P = z_1 + z_2 + z_1^2 z_2 + z_1 z_2^2$ and $Q = z_1 z_2$, we obtain easily that $\mathcal{M}(P, Q)$ is constituted by four cones as shown in Figure 3. For convenience in this figure and in the figures relative to the example that follow the minimal causality cones are translated in such a way that their vertices coincide with the vertices of the convex hull of p.

Example 2. Let Σ be a 2D AR system whose behavior is the kernel of the polynomial matrix

$$R = [z_1 - z_2 \mid 1].$$

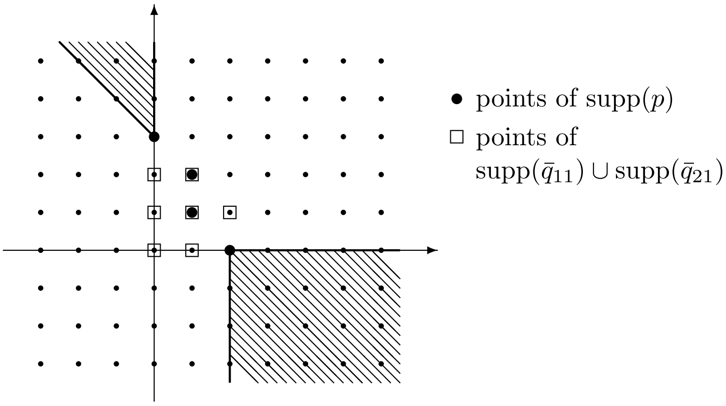


Fig. 6

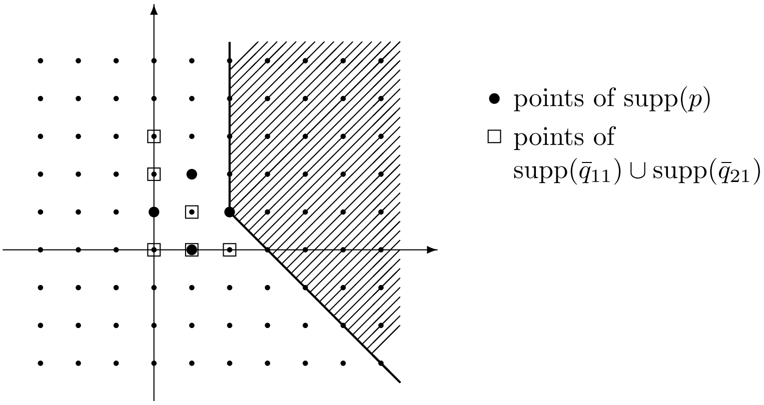


Fig. 7

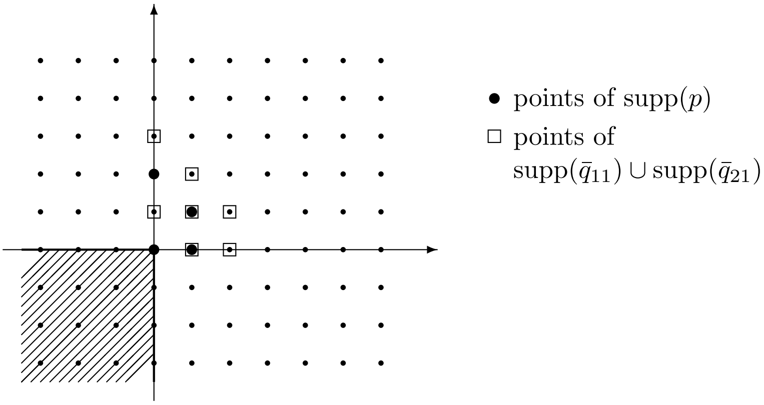


Fig. 8

This is the same 2D AR system considered in Example 1 in [14]. We can consider two I/O representations of Σ. If we let $P = z_1 - z_2$ and $Q = -1$, the set $\mathcal{M}(P,Q)$ contains the cones shown in Figure 4, while if we let $P = -1$ and $Q = z_1 - z_2$, we obtain easily that $\mathcal{M}(P,Q)$ is constituted by only one cone as shown in Figure 5.

Example 3. Let Σ be a 2D AR system whose behavior is the kernel of the polynomial matrix

$$R = \begin{bmatrix} z_1 - z_2^2 & 0 & 2z_1z_2 - 1 \\ 1 & z_1 - z_2 & 1 \end{bmatrix}.$$

This is the same 2D AR system considered in Example 2 in [14]. We can consider three I/O representations of Σ. If we let

$$P = \begin{bmatrix} z_1 - z_2^2 & 0 \\ 1 & z_1 - z_2 \end{bmatrix}, \qquad Q = \begin{bmatrix} 2z_1z_2 - 1 \\ 1 \end{bmatrix},$$

then following the algorithm presented above, we obtain that $p = z_1^2 - z_1z_2 - z_1z_2^2 + z_2^3$, $\bar{q}_{11} = -z_1 + z_2 + 2z_1^2z_2 - 2z_1z_2^2$, and $\bar{q}_{21} = 1 - 2z_1z_2 + z_1 - z_2^2$, and so, as shown in Figure 6, we see that the set $\mathcal{M}(P,Q)$ contains two cones.

Finally, if we consider the other two remaining I/O representations, we obtain the sets of minimal causality cones shown in Figures 7 and 8.

REFERENCES

[1] M. Atiyah and I. MacDonald, *Commutative Algebra*, Addison–Wesley, Reading, MA, 1969.

[2] N. Bose, *Applied Multidimensional Systems Theory*, Van Nostrand Reinhold, New York, 1982.

[3] R. Eising, *Realization and stabilization of* 2*-D systems*, IEEE Trans. Automat. Control, AC-23 (1980), pp. 793–799.

[4] E. Fornasini and G. Marchesini, *State-space realization theory of two-dimensional filters*, IEEE Trans. Automat. Control, 21 (1976), pp. 484–492.

[5] G. Forney, *Convolutional codes* I: *Algebraic structure*, IEEE Trans. Inform. Theory, IT-16 (1970), pp. 720–738.

[6] T. Kailath, *Linear Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[7] V. Kucera, *Discrete Linear Control, The Polynomial Equation Approach*, Wiley, New York, 1979.

[8] S. Kung, B. Lévy, M. Morf, and T. Kailath, *New results in* 2D *systems theory. Part* 1, Proc. IEEE, 65 (1977), pp. 861–872.

[9] P. Rocha, *Structure and Representation of* 2*-D Systems*, Ph.D. thesis, University of Groningen, Groningen, The Netherlands, 1990.

[10] P. Rocha and J. Willems, *Canonical computational forms for AR* 2*-D systems*, Multidimens. Systems Signal Process., 2 (1990), pp. 251–278.

[11] P. Rocha and J. Willems, *Controllability of* 2*-D systems*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 413–423.

[12] J. Willems, *Models for dynamics*, Dynam. Report., 2 (1988), pp. 171–269.

[13] S. Zampieri, *A solution of the Cauchy problem for multidimensional discrete linear shift-invariant systems*, Linear Algebra Appl., 202 (1994), pp. 143–162.

[14] S. Zampieri, *Causal input/output representation of* 2D *systems in the behavioral approach*, SIAM J. Control Optim., 36 (1998), pp. 1133–1146.

[15] E. Zerz and U. Oberst, *The canonical Cauchy problem for linear systems of partial difference equations with constant coefficients over the complete r-dimensional integral lattice $\mathbb{Z}^r$*, Acta Appl. Math., 31 (1993), pp. 249–273.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 5, pp. 1553–1567

STABILITY OF AN ADAPTIVE REGULATOR FOR PARTIALLY KNOWN NONLINEAR STOCHASTIC SYSTEMS[*]

A. BROCKWELL[†], K. BOROVKOV[†], AND R. EVANS[‡]

Abstract. We investigate the properties of a fast-identification style of control algorithm applied to a class of stochastic dynamical systems in continuous time which are sampled at a constant rate. The algorithm does not assume that the system dynamics are known and estimates them using a simple filter. Under a mild smoothness condition on the system dynamics, we show that when the sampling rate is sufficiently fast, the control algorithm stabilizes the system in the sense that the sampled closed-loop system becomes an ergodic Markov chain. Moreover, an explicit bound is given for the expected deviation of the system state from the origin. The result is also adapted for the case where state-measurement is subject to random noise.

Key words. adaptive, stochastic, control, nonlinear systems

AMS subject classifications. 60H30, 60J25, 93C10, 93C90, 93E15, 93E35

PII. S0363012997331512

1. Introduction.

Suppose that we have a continuous-time dynamical system given by

$$dx^{(p-1)}(t) = f(x^{(p-1)}(t), \dots, \dot{x}(t), x(t))dt + u(t)dt + \sigma dw(t), \tag{1.1}$$

where $x^{(k)} = \frac{d^k}{dt^k}x(t)$, $x(t) \in \mathbb{R}$, and $w(t)$ is a standard Brownian motion process. The real-valued control term $u(t)$ takes constant values on each time interval $t \in [j\delta, (j+1)\delta)$, $j = 0, 1, 2, \dots$, and can depend only on observed values of x and its derivatives at sampling points in time given by $\{k\delta,\ k = 0, 1, 2, \dots;\ k\delta \le t\}$, where $\delta > 0$ is a fixed sampling interval length which is to be specified in advance. Thus the control can be defined by the sequence $\{u_j\}$ with $u_j = u(t)$, $j\delta \le t < (j+1)\delta$. The function $f : \mathbb{R}^p \to \mathbb{R}$ is not known to the controller, but satisfies the global Lipschitz condition

$$|f(a) - f(b)| \le \beta \|a - b\| \tag{1.2}$$

for some $\beta < \infty$, $\|a\| = (aa^T)^{1/2}$ being the Euclidean norm for row vectors a.

Our objective is to make the sequence of vectors

$$\{\boldsymbol{x}(j\delta)\} := \{(x^{(p-1)}(j\delta), \dots, \dot{x}(j\delta), x(j\delta))\}, \quad j = 0, 1, 2, \dots,$$

behave in some desired fashion. We do this by choosing the (nonanticipative) control sequence $\{u_j, j = 0, 1, 2, \dots\}$ so as to make the system mimic as closely as possible a prespecified reference system. This type of approach is commonly used for adaptive control problems (see, e.g., [6]). In order to match the system dynamics to those of the reference system, the controller will need to estimate the values of the unknown function $f(\cdot)$ on each sampling interval. In fact, it will turn out that as the class

[*] Received by the editors December 12, 1997; accepted for publication (in revised form) November 18, 1998; published electronically August 2, 1999.

http://www.siam.org/journals/sicon/37-5/33151.html

[†] Department of Mathematics and Statistics, The University of Melbourne, Parkville VIC 3052, Australia (anthonyb@ms.unimelb.edu.au, kostya@ms.unimelb.edu.au).

[‡] Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville VIC 3052, Australia (r.evans@ee.mu.oz.au).

of possible systems to be controlled becomes larger (i.e., β increases), the sampling rate required to guarantee closed-loop stability also increases. Further aspects of this fast-sampling approach will be discussed below.

In this paper, we consider only the regulation problem, i.e., we choose reference systems which are stable around the origin. However, the results can be adapted for other types of problems.

There are relatively few results pertaining to the stability of control laws applied to nonlinear stochastic systems. In fact, to the authors' knowledge, this is the only current result which establishes closed-loop stability of a control law applied to an unknown nonlinear stochastic system. Florchinger [4] has obtained results which apply in the case where noise intensity vanishes as the system state approaches the origin, and Meyn and Guo [7] have obtained results on stability of control laws applied to linear time-varying stochastic systems (see also [2]). In this paper, we consider the case where the system dynamics are nonlinear and the noise intensity does not necessarily approach zero as the state approaches the origin.

The discrete-time system

$$\boldsymbol{y}_t := (y_t^{[1]}, y_t^{[2]}, \dots, y_t^{[p]}) \in \mathbb{R}^p$$

which matches the continuous-time system at the sampling points (in the sense that $\boldsymbol{y}_t = \boldsymbol{x}(t\delta)$) is governed by the system of nonlinear equations

$$\begin{aligned} y_{t+1}^{[1]} &= y_t^{[1]} + \delta(f(\boldsymbol{y}_t) + u_t) + \epsilon_{t+1} + d_{t+1}^{[1]}, \\ y_{t+1}^{[2]} &= y_t^{[2]} + \delta y_t^{[1]} + d_{t+1}^{[2]}, \\ &\cdots \\ y_{t+1}^{[p]} &= y_t^{[p]} + \delta y_t^{[p-1]} + d_{t+1}^{[p]}, \end{aligned} \tag{1.3}$$

where the noise terms ϵ_t are given by

$$\epsilon_t = \sigma(w(t\delta) - w((t-1)\delta)),$$

$w(t)$ being the standard Brownian motion generating the white noise in our original system, and the terms $\boldsymbol{d}_t := (d_t^{[1]}, d_t^{[2]}, \dots, d_t^{[p]})$ are "discretization-correction" terms clearly given by

$$\begin{aligned} d_{t+1}^{[1]} &= \int_{t\delta}^{(t+1)\delta} [f(\boldsymbol{x}(s)) - f(\boldsymbol{x}(t\delta))]\, ds, \\ d_{t+1}^{[2]} &= \int_{t\delta}^{(t+1)\delta} \left[x^{(p-1)}(s) - x^{(p-1)}(t\delta)\right] ds, \\ &\cdots \\ d_{t+1}^{[p]} &= \int_{t\delta}^{(t+1)\delta} [\dot{x}(s) - \dot{x}(t\delta)]\, ds. \end{aligned} \tag{1.4}$$

This choice of $\boldsymbol{d}_{t+1}$ ensures that the discrete-time system does in fact match the continuous time system at the sampling points. To simplify the expressions which follow, we will often refer to the quantities $y_t^{[1]}$ and $d_t^{[1]}$ simply as y_t and d_t, respectively. Note also that $\{\epsilon_t\}$ forms a sequence of independent identically distributed normal random variables, with first absolute moment $\gamma := \mathbf{E}[|\epsilon_t|] = \sigma\sqrt{2\delta/\pi}$.

The system (1.3) can be written more concisely in vector notation as

$$\boldsymbol{y}_{t+1} = \boldsymbol{y}_t(I + \delta B) + (\delta f(\boldsymbol{y}_t) + \delta u_t + \epsilon_{t+1})\boldsymbol{e}_1 + \boldsymbol{d}_{t+1}, \tag{1.5}$$

where B is the $p \times p$ shift matrix

$$B = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

and $\boldsymbol{e}_1$ is the row vector $(1, 0, \dots, 0)$.

To control our system, we will use a reference model specified by the linear relation

$$\boldsymbol{z}_{t+1} = \boldsymbol{z}_t L, \quad L = I + \delta H, \tag{1.6}$$

where

$$H = \begin{bmatrix} h_1 & 1 & 0 & \dots & 0 \\ h_2 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ h_{p-1} & 0 & 0 & \dots & 1 \\ h_p & 0 & 0 & \dots & 0 \end{bmatrix}. \tag{1.7}$$

Note that the structure of the reference system (1.6) matches that of the discrete-time system (1.5), i.e., there always exists some $u_t \in \mathbb{R}$ such that (1.5) is equivalent to $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t L + \boldsymbol{d}_{t+1}$. The reference system is in fact an Euler approximation to the continuous-time linear phase-space system $\dot{\boldsymbol{z}} = H\boldsymbol{z}$. Clearly, in order for it to be stable, the continuous-time system must also be stable; i.e., all eigenvalues of H must have negative real parts. Furthermore, δ must be sufficiently small (so that the approximation is "reasonably" good). We establish the maximal value of δ for which stability of (1.6) is guaranteed, along with a bound on the spectral radius of L, as follows.

Let $\{\lambda_i^L\}$ and $\{\lambda_i^H\}$ denote the sets of eigenvalues of L and H, respectively. It is easily seen that λ_i^H is an eigenvalue of H if and only if $1 + \delta\lambda_i^H$ is an eigenvalue of L. Let

$$c = \min_i \frac{|\mathrm{Re}(\lambda_i^H)|}{|\lambda_i^H|} > 0, \quad \delta_0 = \frac{c}{\max_i |\lambda_i^H|}, \tag{1.8}$$

and

$$h = \min_i \frac{|\lambda_i^H|(|\lambda_i^H| - |\mathrm{Im}(\lambda_i^H)|)}{|\mathrm{Re}(\lambda_i^H)|}. \tag{1.9}$$

It can then be shown, using an elementary geometric argument, that for all δ such that $0 \le \delta \le \delta_0$ and for each eigenvalue $\lambda_i^L = 1 + \delta\lambda_i^H$,

$$|\lambda_i^L| \le \rho := 1 - h\delta. \tag{1.10}$$

One possible choice of the quantities $\{h_j\}$ is

$$h_j = -\binom{p}{j} h^j, \quad j = 1, \dots, p,$$

for some $h > 0$. In this case, a direct computation shows that $\det(L - \lambda I) = (1 - \delta h - \lambda)^p$, so that L has a single eigenvalue $\lambda = 1 - \delta h$, with multiplicity p.

Our next step is to choose a control which will match the system dynamics (1.5) to the reference dynamics (1.6). If these were to be matched exactly, we would need to choose u_t so that

$$\delta f(\boldsymbol{y}_t) + \delta u_t + \epsilon_{t+1} + d_{t+1} = \delta h(\boldsymbol{y}_t)$$

or

$$u_t = h(\boldsymbol{y_t}) - f(\boldsymbol{y}_t) - (\epsilon_{t+1} + d_{t+1})/\delta. \tag{1.11}$$

Unfortunately, one cannot choose u_t in this manner. It must depend only on current and past observations of the system, and in (1.11) above, $f(\boldsymbol{y_t})$, ϵ_{t+1}, and d_{t+1} are unobservable at time t. We therefore use a simple filter to obtain an estimate of $f(\boldsymbol{y_t})$. Rather than making any assumptions on the form of $f(\cdot)$ and estimating parameters, we directly estimate the value of $f(\boldsymbol{y}_t)$ itself. Taking into account the fact that $\mathbf{E}\left[\epsilon_t\right] = 0$ and ignoring the presence of the discretization-correction terms $\boldsymbol{d}_t$, we can obtain from (1.3) an estimate ζ_{t-1} of $f(\boldsymbol{y}_{t-1})$ of the form

$$\zeta_{t-1} = (y_t - y_{t-1})/\delta - u_{t-1}$$

(recall that $y_t = y_t^{[1]}$). Since this filter is operating with a lag, we cannot compute ζ_t until time $t+1$. Hence, for control purposes, we rely on the continuity of $f(\cdot)$ and use ζ_{t-1} as an estimate of $f(\boldsymbol{y}_t)$.

Substituting our estimate into (1.11), and again replacing the terms ϵ_{t+1} and d_{t+1} with 0, we obtain the control law

$$u_t = h(\boldsymbol{y}_t) - \zeta_{t-1} = h(\boldsymbol{y}_t) - (y_t - y_{t-1})/\delta + u_{t-1}. \tag{1.12}$$

In terms of our original process (1.1), this control law takes the form

$$u_t = h(\boldsymbol{x}(t\delta)) - \left[x^{(p-1)}(t\delta) - x^{(p-1)}((t-1)\delta)\right]/\delta + u_{t-1}. \tag{1.13}$$

Since we cannot use this control law at time $t = 0$, we initialize the system by setting $u_0 = 0$.

This controller can be thought of as trying to eliminate any "external forces" (which are specified by the function $f(\cdot)$) and replacing them with the force which would be applied by the stable reference system were it at the current point of the phase space. With this in mind, we hope that the system will tend to "imitate" the behavior of the reference system. In particular, we could expect it to be stable around the origin.

The dynamics of our system (using this control law) are now given by the closed-loop equation (substituting (1.12) back into (1.5))

$$\begin{aligned}\boldsymbol{y}_{t+1} &= \boldsymbol{y}_t(I + \delta B) + (\delta f(\boldsymbol{y}_t) - \delta\zeta_{t-1})\boldsymbol{e}_1 + (\delta h(\boldsymbol{y}_t) + \epsilon_{t+1})\boldsymbol{e}_1 + \boldsymbol{d}_{t+1} \\ &= \boldsymbol{y}_t L + \delta(f(\boldsymbol{y}_t) - f(\boldsymbol{y}_{t-1}))\boldsymbol{e}_1 + (\epsilon_{t+1} - \epsilon_t)\boldsymbol{e}_1 + \boldsymbol{d}_{t+1} - d_t\boldsymbol{e}_1.\end{aligned} \tag{1.14}$$

The sequence $\{\boldsymbol{y}_t\}$ is not a Markov chain. However, it is easily seen (from (1.5) and (1.12)) that the sequence of vectors

$$Z_t = (\boldsymbol{y}_t, u_t) = (y_t, y_t^{[2]}, \ldots, y_t^{[p]}, u_t) \tag{1.15}$$

forms a (time-homogeneous) Markov chain.

2. The main result. Before stating the main result of this paper, we will require the following definitions.

Suppose that $\{X_t\}$ is a time-homogeneous Markov chain which takes values on the space $\mathbb{R}^q$. Let $\mathcal{B}(\mathbb{R}^q)$ denote the family of Borel sets in $\mathbb{R}^q$. Let $P^k(z, A)$ denote the k-step transition function of the Markov chain $\{X_t\}$, $P(z, A) = P^1(z, A)$. The Markov chain $\{X_t\}$ on $\mathbb{R}^q$ is said to be *ergodic* if there exists an invariant probability measure π on $\mathcal{B}(\mathbb{R}^q)$, that is,

$$\pi(A) = \int_{x\in\mathbb{R}^q} \pi(dx)P(x, A) \quad \text{for all } A \in \mathcal{B}(\mathbb{R}^q),$$

such that

$$\lim_{n\to\infty} \|P^n(x, \cdot) - \pi(\cdot)\|_{TV} = 0 \quad \text{for all } x \in \mathbb{R}^q,$$

where $\|\cdot\|_{TV}$ represents the total variation norm on signed measures on $\mathcal{B}(\mathbb{R}^q)$.

Our main result is the following theorem.

THEOREM 2.1. *Assume that the system* (1.1) *satisfies condition* (1.2), *and take any vector* $(h_1, \ldots, h_p)$ *such that all eigenvalues of the matrix* H *given by* (1.7) *have negative real parts. Suppose that the sampling interval* δ *is no greater than the quantity* δ_0 *given by* (1.8) *and satisfies the inequalities*

$$k_1(\delta) := \frac{\exp\{(\beta+1)\delta\} - 1 - (\beta+1)\delta}{(\beta+1)\delta} < 1 \tag{2.1}$$

and

$$\alpha := \delta\beta + \delta^2\beta\|H\|\varrho + \frac{2\varrho(k_4(\delta) + k_5(\delta))}{1 - k_1(\delta)} < 1, \tag{2.2}$$

where the quantities $k_4(\delta)$, $k_5(\delta)$, *and* ϱ *are given by*

$$k_4(\delta) = \delta k_1(\delta)(\|H\| + \beta), \quad k_5(\delta) = \delta k_1(\delta)\beta$$

and

$$\varrho = \frac{1}{\delta}\left(\frac{1}{h} + \frac{w_h}{h^2} + \cdots + \frac{w_h^{p-1}}{h^p}\right)$$

with

$$w_h = \sqrt{2}\left(\frac{p^3 - p}{3}\right)^{1/4}\left(p - 1 + \sum_{i=1}^{p} h_i^2\right)^{1/2}. \tag{2.3}$$

Then

(i) *the Markov chain* $\{Z_t\}$ *with control law* u_t *specified by* (1.12) *is ergodic and*

(ii) *under this control law, the continuous-time process* $\{\boldsymbol{x}(t)\}$ *satisfies*

$$\limsup_{t\to\infty} \mathbf{E}\left[\|\boldsymbol{x}(t)\|\right] \le k_7(\delta)\eta + k_8(\delta),$$

where

$$\eta = (1-\alpha)^{-1}\left((\sigma\sqrt{2\delta/\pi} + \delta|f(0)|)(1 + \delta\|H\|\varrho) + \frac{2k_6(\delta)\varrho}{1 - k_1(\delta)}\right),$$

$$k_6(\delta) = k_1(\delta)\sigma\sqrt{2\delta/\pi} + k_2(\delta),$$

$$k_7(\delta) = 1 + \frac{1+k_1(\delta)}{1-k_1(\delta)}\left(\delta(\|H\| + 2\beta)\right),$$

$$k_8(\delta) = \sigma\sqrt{8\delta/\pi} + k_2(\delta) + \frac{1+k_1(\delta)}{1-k_1(\delta)}(\sigma\sqrt{2\delta/\pi} + k_2(\delta)),$$

and

$$k_2(\delta) = \sqrt{8/(9\pi)}(\beta+1)\sigma\delta^{3/2}\exp\{(\beta+1)\delta\}.$$

Remark 1. The first part of Theorem 2.1 states that when the discrete-time controller (1.12) is applied to the continous-time system (1.1) with sampling interval δ, the state of the continuous-time system (and the controller) at the sampling times will approach a limiting distribution as time increases. The second part of the theorem places bounds on the deviation of the continuous-time system from the origin, thereby ensuring that the system will not behave "badly" in between the sampling times (this seems clear intuitively due to the smoothness of $f(\cdot)$).

Remark 2. The condition (2.1) will be satisfied if

$$\delta(\beta+1) < 1.2564. \tag{2.4}$$

By choosing a sufficiently small sampling interval δ, conditions (2.1) and (2.2) can always be met. Although this is a fast-sampling approach to control, which has some inherent disadvantages, the compromise between the level of accuracy of plant-knowledge and the sampling rate required for stability is an interesting issue which, to the authors' knowledge, is not well understood. We will not address this issue in this paper. However, the results obtained, in particular the required relationship between δ and β given by (2.4), suggest that this compromise is an issue worthy of further consideration.

Remark 3. The result can be extended to the case of noisy (but still complete) state measurements. Suppose that the controller only observes $\boldsymbol{y}'_t = \boldsymbol{y}_t + \boldsymbol{n}_t$, where $\{\boldsymbol{n}_t\}$ is a sequence of independently and indentically distributed (i.i.d.) random variables with covariance matrix $Q > 0$. Using the same control law (1.12), with $\boldsymbol{y}_t$ replaced by $\boldsymbol{y}'_t$, the sequence $\{Z_t = (\boldsymbol{y}_t, u_t, \boldsymbol{n}_t)\}$ is a Markov chain. If the quantities η, $k_6(\delta)$, and $k_8(\delta)$ are replaced with

$$\eta = (1-\alpha)^{-1}\Big[(\sigma\sqrt{2\delta/\pi} + \sqrt{\mathrm{tr}\ Q} + \delta|f(0)|)(1 + \delta\|H\|\varrho) + \frac{2k_6(\delta)\varrho}{1-k_1(\delta)} + \delta\|H\|\varrho\sqrt{\mathrm{tr}\ Q}\Big],$$

$$k_6(\delta) = k_1(\delta)\sigma\sqrt{2\delta/\pi} + k_2(\delta) + (2 + \delta\|H\|)k_1(\delta)\sqrt{\mathrm{tr}\ Q},$$

and

$$k_8(\delta) = \sigma\sqrt{8\delta/\pi} + k_2(\delta) + \frac{1+k_1(\delta)}{1-k_1(\delta)}\left(\sigma\sqrt{2\delta/\pi} + k_2(\delta) + 2\sqrt{\mathrm{tr}\ Q}\right),$$

then the results of Theorem 2.1 still hold. This means that the conditions for stability of the closed-loop system remain the same, but the bounds on the first absolute moment of the Markov chain (and on the first absolute moment of the continuous-time process) increase.

Remark 4. The result can also be extended to the multidimensional case where $y_t^{[i]} \in \mathbb{R}^q$, $u_t \in \mathbb{R}^q$, and $w(t)$ is a q-dimensional Brownian motion.

3. Proofs. First we bound $\mathbf{E}\left[\|\boldsymbol{d}_t\|\right]$ as a function of δ. Using this bound we will find a limiting upper bound for $\eta_t := \mathbf{E}\left[\|\boldsymbol{y}_t\|\right]$. This is a form of stability in itself. Then we will show that the Markov chain (1.15) is irreducible and aperiodic and has the (weak) Feller property. This allows us to use results from [8] to strengthen the stability result by establishing ergodicity of the chain.

3.1. Bounds on moments. In order to study the behavior of the discretization-correction terms $\boldsymbol{d}_t$, we establish a bound on the first absolute moment of the difference between processes $\boldsymbol{v}(t)$ and $\boldsymbol{n}(t)$ (starting at the same point $\boldsymbol{v}(0) = \boldsymbol{n}(0)$) which satisfy the equations

$$d\boldsymbol{v}(t) = G(\boldsymbol{v}(t))dt + \sigma\boldsymbol{e}_1 dw(t) \tag{3.1}$$

and

$$\boldsymbol{n}(t) - \boldsymbol{n}(0) = G(\boldsymbol{n}(0))t + \sigma\boldsymbol{e}_1 w(t), \tag{3.2}$$

where $G(\cdot)$ satisfies the global Lipschitz condition $\|G(a) - G(b)\| \leq \kappa\|a - b\|$ and $w(t)$ is a standard Brownian motion. We examine the error $\boldsymbol{v}(t) - \boldsymbol{n}(t)$ by integrating (3.1) and subtracting (3.2) to obtain

$$v(t) - n(t) = \int_0^t \left(G(\boldsymbol{v}(s)) - G(\boldsymbol{n}(s)) + G(\boldsymbol{n}(s)) - G(\boldsymbol{n}(0))\right) ds.$$

Taking norms and expectations, defining $Q(t) = \mathbf{E}\left[\sup_{0\leq s\leq t} \|\boldsymbol{v}(s) - \boldsymbol{n}(s)\|\right]$, and using the Lipschitz property of the function $G(\cdot)$, we obtain

$$\begin{aligned} Q(t) &\leq \kappa \int_0^t \mathbf{E}\left[\sup_{0\leq u\leq s} \|\boldsymbol{v}(s) - \boldsymbol{n}(s)\|\right] ds + \kappa\mathbf{E}\left[\sup_{0\leq u\leq t} \int_0^u \|sG(\boldsymbol{v}(0)) + \sigma\boldsymbol{e}_1 w(s)\| ds\right] \\ &= \kappa \int_0^t Q(s)ds + \frac{\kappa t^2}{2}\|G(\boldsymbol{v}(0))\| + \kappa\sigma \int_0^t \mathbf{E}\left[|w(s)|\right] ds \\ &= \kappa \int_0^t Q(s)ds + \frac{\kappa t^2}{2}\|G(\boldsymbol{v}(0))\| + \sqrt{\frac{8}{9\pi}}\kappa\sigma t^{3/2}. \end{aligned}$$

Using the Gronwall–Bellman inequality (see, e.g., [3]), we obtain the bound

$$Q(t) \leq \kappa\|G(\boldsymbol{v}_0)\| \left(\frac{\exp^{\kappa t} - 1 - \kappa t}{\kappa^2}\right) + \sqrt{\frac{8}{9\pi}}\kappa\sigma t^{3/2} \exp\{\kappa t\}. \tag{3.3}$$

Applying this result to our original process (1.1) and its discretization (1.3) (with $G(\boldsymbol{y}_t) = \boldsymbol{y}_t B + (f(\boldsymbol{y}_t) + u_t)\boldsymbol{e}_1$ so that $\kappa = \beta + 1$), we obtain the bound on the conditional expectation $\mathbf{E}\left[\|\boldsymbol{d}_{t+1}\| \mid \boldsymbol{y}_t, u_t\right]$ given by

$$\mathbf{E}\left[\|\boldsymbol{d}_{t+1}\| \mid \boldsymbol{y}_t, u_t\right] \leq k_1(\delta)\delta\|\boldsymbol{y}_t B + f(\boldsymbol{y}_t)\boldsymbol{e}_1 + u_t\boldsymbol{e}_1\| + k_2(\delta), \tag{3.4}$$

where the quantities $k_1(\delta)$ and $k_2(\delta)$ are defined in Theorem 2.1. For the sake of brevity, we will subsequently suppress the explicit reference to δ in the quantities $k_j(\delta)$, so that, for instance, when we refer to k_1 we really mean $k_1(\delta)$. Substituting the control (1.12) into (3.4), we obtain

$$\begin{aligned} \mathbf{E}\left[\|\boldsymbol{d}_{t+1}\| \mid \boldsymbol{y}_t, u_t\right] &\leq \delta k_1\|\boldsymbol{y}_t B + (f(\boldsymbol{y}_t) + h(\boldsymbol{y}_t) - \zeta_{t-1})\boldsymbol{e}_1\| + k_2 \\ &= \delta k_1\|\boldsymbol{y}_t H + (f(\boldsymbol{y}_t) - f(\boldsymbol{y}_{t-1}) + f(\boldsymbol{y}_{t-1}) - \zeta_{t-1})\boldsymbol{e}_1\| + k_2 \\ &= \delta k_1\|\boldsymbol{y}_t H + (f(\boldsymbol{y}_t) - f(\boldsymbol{y}_{t-1}))\boldsymbol{e}_1\| + k_1\|(\epsilon_t + d_t)\boldsymbol{e}_1\| + k_2. \end{aligned}$$

We now take expectations to get (recall that $\eta_t = \mathbf{E}\left[\|\boldsymbol{y}_t\|\right]$)

$$\begin{aligned}\mathbf{E}\left[\|\boldsymbol{d}_{t+1}\|\right] &\leq \delta k_1\big(\|H\|\mathbf{E}\left[\|\boldsymbol{y}_t\|\right] + \beta\mathbf{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|\right]\big) + k_1\gamma + k_1\mathbf{E}\left[\|\boldsymbol{d}_t\|\right] + k_2 \\ &\leq \delta k_1((\|H\| + \beta)\eta_t + \beta\eta_{t-1}) + k_1\gamma + k_1\mathbf{E}\left[\|\boldsymbol{d}_t\|\right] + k_2 \\ &\leq k_4\eta_t + k_5\eta_{t-1} + k_6 + k_1\mathbf{E}\left[\|\boldsymbol{d}_t\|\right], \qquad (3.5)\end{aligned}$$

where k_4, k_5, and k_6 are the quantities defined in Theorem 2.1 (note their dependence on δ).

Iterating the inequality (3.5), we find that

$$\mathbf{E}\left[\|\boldsymbol{d}_{t+1}\|\right] \leq k_1^t\mathbf{E}\left[\|\boldsymbol{d}_1\|\right] + \sum_{j=0}^{t-1} k_1^j[k_4\eta_{t-j} + k_5\eta_{t-j-1} + k_6]. \qquad (3.6)$$

We now set $r_t = \delta f(\boldsymbol{y}_{t-1})\boldsymbol{e}_1 + \epsilon_t\boldsymbol{e}_1$ and iterate the closed-loop system equation (1.14) to obtain

$$\begin{aligned}\boldsymbol{y}_{t+1} &= \boldsymbol{y}_1 L^t + \sum_{j=0}^{t-1}(r_{t+1-j} - r_{t-j} + \boldsymbol{d}_{t+1-j} - d_{t-j}^{[1]}\boldsymbol{e}_1)L^j \\ &= \left(\sum_{j=1}^{t-1} r_{t+1-j}L^{j-1}L - \sum_{j=1}^{t-1} r_{t+1-j}L^{j-1}\right) \\ &\qquad + \boldsymbol{y}_1 L^t + r_{t+1} - r_1 L^{t-1} + \sum_{j=0}^{t-1}(\boldsymbol{d}_{t+1-j} - d_{t-j}^{[1]}\boldsymbol{e}_1)L^j \\ &= \boldsymbol{y}_1 L^t + r_{t+1} - r_1 L^{t-1} + \delta H\sum_{j=1}^{t-1} r_{t+1-j}L^{j-1} + \sum_{j=0}^{t-1}(\boldsymbol{d}_{t+1-j} - d_{t-j}^{[1]}\boldsymbol{e}_1)L^j. \qquad (3.7)\end{aligned}$$

Using the inequality (3.5), we can bound the first absolute moment of the last sum on the right-hand side of (3.7) as follows (setting $\eta_{-1} = 0$):

$$\begin{aligned}\mathbf{E}\left[\left\|\sum_{j=0}^{t-1}(\boldsymbol{d}_{t+1-j} - d_{t-j}^{[1]}\boldsymbol{e}_1)L^j\right\|\right] &\leq \sum_{j=0}^{t-1}\|L^j\|\left(\mathbf{E}\left[\|\boldsymbol{d}_{t+1-j}\|\right] + \mathbf{E}\left[\|\boldsymbol{d}_{t-j}\|\right]\right) \\ &\leq \sum_{k=0}^{t-1-j} k_1^k[k_4\eta_{t-j-k} + (k_4 + k_5)\eta_{t-j-k-1} + k_5\eta_{t-j-k-2} + 2k_6] \\ &\quad + \sum_{j=0}^{t-1}\|L^j\|\left((k_1^{t-j} + k_1^{t-j-1})\mathbf{E}\left[\|\boldsymbol{d}_1\|\right]\right). \qquad (3.8)\end{aligned}$$

Recall that $u_0 = 0$, so it follows from inequality (3.4) that

$$\mathbf{E}\left[\|\boldsymbol{d}_1\|\right] \leq \delta k_1(\eta_0(1 + \beta) + \omega) + k_2, \qquad (3.9)$$

where $\omega = |f(0)|$. Taking norms and expectations on both sides of (3.7) and substituting (3.8) and (3.9), we have

$$\eta_{t+1} \leq \eta_1\|L^t\| + \delta\beta\eta_t + (\delta\beta\eta_0 + \gamma + \delta\omega)\|L^{t-1}\| + \gamma + \delta\omega$$

$$
\begin{aligned}
&+\delta\|H\|\sum_{j=0}^{t-1}\|L^{j-1}\|(\delta\beta\eta_{t-j}+\gamma+\delta\omega)\\
&+\sum_{j=0}^{t-1}\|L^j\|(k_1^{t-j}+k_1^{t-j-1})(k_1\delta(\eta_0(1+\beta)+\omega)+k_2)\\
&+\sum_{j=0}^{t-1}\|L^j\|\sum_{k=0}^{t-1-j}k_1^k[k_4\eta_{t-j-k}+(k_4+k_5)\eta_{t-j-k-1}+k_5\eta_{t-j-k-2}+2k_6]. \qquad (3.10)
\end{aligned}
$$

Now we need to bound the norms of L^k, $k=0,1,2,\ldots$. If the matrix L were normal (i.e., $LL^*=L^*L$, where L^* is the conjugate-transpose of L), then we would have $\|L^k\|\le\rho^k$, ρ being the spectral radius (maximum of magnitudes of eigenvalues) of L bounded by (1.10). Unfortunately, the matrix $L=I+\delta H$ is not normal, so we will have to make use of a different bound.

If UTU^* is a Schur decomposition of L (i.e., T is upper triangular and U is unitary), then the matrix T can be expressed as the sum of a diagonal matrix and a strictly upper triangular matrix : $T=D+M$. Let $\mathcal{M}$ be the set of matrices M obtained from every possible Schur decomposition of L. The quantity $\Delta(L)=\inf_{M\in\mathcal{M}}\|M\|$ can be used as a measure of nonnormality of the matrix M. Let $\|A\|_\epsilon$ denote the Euclidean norm of the matrix $A=[a_{i,j}]$ defined by

$$\|A\|_\epsilon=\left(\sum_{i,j}a_{i,j}^2\right)^{1/2}.$$

Then $\Delta(L)$ satisfies (see [5, Theorem 1])

$$
\begin{aligned}
\Delta(L)&\le\left(\frac{p^3-p}{12}\right)^{1/4}\sqrt{\|L^*L-LL^*\|_\epsilon}=\left(\frac{p^3-p}{12}\right)^{1/4}\sqrt{\delta^2\|HH^*-H^*H\|_\epsilon}\\
&\le\left(\frac{p^3-p}{12}\right)^{1/4}\delta\sqrt{2\|H\|_\epsilon^2}=\left(\frac{p^3-p}{3}\right)^{1/4}\delta\sqrt{2}\left(p-1+\sum_{i=1}^p h_i^2\right)^{1/2}\le\delta w_h,
\end{aligned}
$$

where w_h is defined by (2.3).

As a direct consequence of Theorem 2 of [5], we have the following lemma.

LEMMA 3.1. *Let $m=w_h\delta$, with w_h defined by* (2.3). *Then for all integers $k\ge1$, we have $\|L^k\|\le\varrho_k$, where*

$$\varrho_k=\rho^k+\binom{k}{1}\rho^{k-1}m+\cdots+\binom{k}{p-1}\rho^{k-p+1}m^{p-1} \qquad (3.11)$$

and ρ is the bound on the spectral radius of L defined by (1.10).

Summing the inequality (3.11), we obtain the bound

$$
\begin{aligned}
\sum_{k=0}^\infty\|L^k\|\le\varrho:=\sum_{k=0}^\infty\varrho_k&=\frac{1}{1-\rho}+m\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right)+\cdots+\frac{m^{p-1}}{(p-1)!}\frac{d^{p-1}}{d\rho^{p-1}}\left(\frac{1}{1-\rho}\right)\\
&\le\frac{1}{\delta h}+\frac{\delta w_h}{(\delta h)^2}+\cdots+\frac{(\delta w_h)^{p-1}}{(\delta h)^p}=\frac{1}{\delta}\left(\frac{1}{h}+\frac{w_h}{h^2}+\cdots+\frac{w_h^{p-1}}{h^p}\right).
\end{aligned}
$$

In order to bound the behavior of η_t, we will also make use of the following version of the discrete-time Gronwall–Bellman lemma.

LEMMA 3.2. *Suppose that* $\{b_t, t = 0, 1, \ldots\}$, $\{\alpha_t, t = 0, 1, \ldots\}$, *and* $\{g_t, t = 0, 1, \ldots\}$ *are sequences of nonnegative real values such that*

$$b_{t+1} \leq \sum_{j=0}^{t} \alpha_j b_{t-j} + g_t, \tag{3.12}$$

with $\sum_{j=0}^{\infty} \alpha_j \leq 1 - \epsilon < 1$. *If* $\{g_t\}$ *is a bounded sequence and* $\lim_{t\to\infty} g_t = g$, *then*

$$\limsup_{t\to\infty} b_t \leq \frac{g}{\epsilon}.$$

It is easily shown that the inequality (3.10) can be expressed in the form (3.12) with

$$g_t = (\gamma + \delta\omega) + (\delta\beta\eta_0 + \gamma + \delta\omega)\|L^{t-1}\| + \eta_1\|L^t\| + \delta\|H\|\varrho(\gamma + \delta\omega) + \frac{2k_6\varrho}{1 - k_1}$$
$$+ \sum_{j=0}^{t-1} \|L^j\|(k_1^{t-j} + k_1^{t-j-1})(k_1\delta(\eta_0(1 + \beta) + \omega) + k_2)$$

so that

$$\lim_{t\to\infty} g_t = g := (\gamma + \delta\omega)(1 + \delta\|H\|\varrho) + \frac{2k_6\varrho}{1 - k_1}.$$

We can sum the coefficients $\{\alpha_j,\ j = 0, 1, \ldots\}$ of $\{\eta_{t-j},\ j = 0, 1, \ldots\}$ in (3.10) to obtain the bound (using (3.12))

$$\sum_{j=0}^{t} \alpha_j \leq \delta\beta + \delta^2\|H\|\beta \sum_{j=0}^{t-2} \|L^j\| + \sum_{j=0}^{t-1} \|L^j\| \sum_{k=0}^{t-1-j} k_1^k(2k_4 + 2k_5)$$
$$\leq \delta\beta + \delta^2\|H\|\beta\varrho + \frac{2\varrho(k_4 + k_5)}{1 - k_1} = \alpha, \tag{3.13}$$

where α is defined in the statement of Theorem 2.1.

Applying Lemma 3.2, we have, if $\alpha < 1$,

$$\mathbf{E}\left[\|\boldsymbol{y}_t\|\right] \leq b_t,$$

with

$$\limsup_{t\to\infty} b_t = \eta := \frac{g}{1 - \alpha}. \tag{3.14}$$

In addition, we can rewrite (1.3) to get

$$u_t = \delta^{-1}(y_{t+1} - y_t - f(\boldsymbol{y}_t)\delta - \epsilon_{t+1} - d_{t+1}).$$

It follows directly from this representation of u_t, the Lipschitz property of $f(\cdot)$, and representation (1.4) that if $\mathbf{E}\left[\|\boldsymbol{y}_t\|\right]$ is bounded, then so is $\mathbf{E}\left[\|u_t\|\right]$. Consequently, if $\alpha < 1$, then $\mathbf{E}\left[\|Z_t\|\right] \leq \mathbf{E}\left[\|\boldsymbol{y}_t\|\right] + \mathbf{E}\left[\|u_t\|\right]$ is bounded.

3.2. Bounds for the continuous-time process. In the previous subsection we established bounds on the first absolute moments of the Markov chain $\{Z_t\}$. Clearly, the behavior of the continuous-time process $\{\boldsymbol{x}(t)\}$ at times other than the sampling instants $\{k\delta,\ k = 0, 1, 2, \ldots\}$ is also of importance. Intuitively, one might expect that due to the smoothness of $f(\cdot)$, the process will not be "badly" behaved in these time intervals. We construct a bound as follows.

Let us consider the process $\boldsymbol{x}(t)$ restricted to the interval $\mathcal{T}_k := [k\delta, (k+1)\delta)$. For all $t \in \mathcal{T}_k$, we have the integral representation

$$\begin{aligned}\boldsymbol{x}(t) &= \boldsymbol{x}(k\delta) + \int_{k\delta}^{t} G_k(\boldsymbol{x}(s))ds + (w(t) - w(k\delta))\boldsymbol{e}_1 \\ &= \boldsymbol{x}(k\delta) + (t - k\delta)G_k(\boldsymbol{x}(k\delta)) + \sigma(w(t) - w(k\delta))\boldsymbol{e}_1 + E(t),\end{aligned}$$

where

$$G_k(\boldsymbol{x}) = \boldsymbol{x}B + (f(\boldsymbol{x}) + u_k)\boldsymbol{e}_1 \tag{3.15}$$

and $E(t)$ is the correction term. Taking norms, suprema over the interval $\mathcal{T}_k$, and expectations, we obtain

$$\begin{aligned}\mathbf{E}\left[\sup_{t\in\mathcal{T}_k} \|\boldsymbol{x}(t)\|\right] \le \mathbf{E}\left[\|\boldsymbol{x}(k\delta)\|\right] + \delta\mathbf{E}\left[\|G_k(\boldsymbol{x}(k\delta))\|\right] + \sigma\mathbf{E}\left[\sup_{t\in\mathcal{T}_k} |w(t) - w(k\delta)|\right] \\ +\mathbf{E}\left[\sup_{t\in\mathcal{T}_k} \|E(t)\|\right].\end{aligned}$$

From (3.3) (with $Q(t) = \mathbf{E}\left[\sup \|E(t)\|\right]$) we have

$$\mathbf{E}\left[\sup_{t\in\mathcal{T}_k} \|E(t)\|\right] \le \delta k_1 \mathbf{E}\left[\|G_k(\boldsymbol{x}(k\delta))\|\right] + k_2,$$

and using the standard argument based on the reflection principle, it can be shown that

$$\mathbf{E}\left[\sup_{t\in\mathcal{T}_k} |w(t) - w(k\delta)|\right] \le \sqrt{8\delta/\pi}.$$

Hence,

$$\begin{aligned}\mathbf{E}\left[\sup_{t\in\mathcal{T}_k} \|\boldsymbol{x}(t)\|\right] \le \mathbf{E}\left[\|\boldsymbol{x}(k\delta)\|\right] + \delta\mathbf{E}\left[\|G_k(\boldsymbol{x}(k\delta))\|\right] + \sigma\sqrt{8\delta/\pi} \\ +\delta k_1\mathbf{E}\left[\|G_k(\boldsymbol{x}(k\delta)\|\right] + k_2.\end{aligned} \tag{3.16}$$

Substituting the equation

$$u_k = h(\boldsymbol{x}(k\delta)) - f(\boldsymbol{x}((k-1)\delta)) - (\epsilon_k + d_k)/\delta$$

into (3.15) and using (3.4), we have (recalling that $\eta_k = \mathbf{E}\left[\|\boldsymbol{x}(k\delta)\|\right]$)

$$\begin{aligned}\mathbf{E}\left[\|G_k(\boldsymbol{x}(k\delta))\|\right] &= \mathbf{E}[\|\boldsymbol{x}(k\delta)B + f(\boldsymbol{x}(k\delta))\boldsymbol{e}_1 - f(\boldsymbol{x}((k-1)\delta))\boldsymbol{e}_1 \\ &\quad +h(\boldsymbol{x}(k\delta))\boldsymbol{e}_1 - (\epsilon_k + d_k)\boldsymbol{e}_1/\delta\|] \\ &\le \|H\|\eta_k + \beta\eta_k + \beta\eta_{k-1} + \sigma\sqrt{2/(\pi\delta)} + \mathbf{E}\left[|d_k|\right]/\delta \\ &\le (\|H\| + \beta)\eta_k + \beta\eta_{k-1} + \sigma\sqrt{2/(\pi\delta)} \\ &\quad +k_1\mathbf{E}\left[\|G_{k-1}(\boldsymbol{x}((k-1)\delta))\|\right] + k_2/\delta.\end{aligned}$$

Multiplying by δ and taking the limit superior on both sides (using (3.14)), we have

$$\limsup_{k\to\infty} \delta\mathbf{E}\left[\|G_k(\boldsymbol{x}(k\delta))\|\right] \le (1-k_1)^{-1}(\delta(\|H\|+2\beta)\eta+\sigma\sqrt{2\delta/\pi}+k_2).$$

Substituting this expression into (3.16) and again taking limits superior on both sides, we find that

$$\begin{aligned}\limsup_{k\to\infty}\mathbf{E}\left[\sup_{t\in\mathcal{T}_k}\|\boldsymbol{x}(t)\|\right] &\le \eta+\sigma\sqrt{8\delta/\pi}+k_2\\ &+\frac{1+k_1}{1-k_1}\left(\delta(\|H\|+2\beta)\eta+\sigma\sqrt{2\delta/\pi}+k_2\right)=k_7\eta+k_8.\end{aligned}$$

As one might expect, this quantity approaches η (cf. (3.14)) as δ approaches 0. This is consistent with the intuitive argument that as the sampling intervals shorten, within each interval the continuous-time process has less time to deviate from the starting point.

The second part of Theorem 2.1 follows from the fact that

$$\limsup_{t\to\infty}\mathbf{E}\left[\|\boldsymbol{x}(t)\|\right] \le \limsup_{k\to\infty}\sup_{t\in\mathcal{T}_k}\mathbf{E}\left[\|\boldsymbol{x}(t)\|\right] \le \limsup_{k\to\infty}\mathbf{E}\left[\sup_{t\in\mathcal{T}_k}\|\boldsymbol{x}(t)\|\right].$$

3.3. Ergodicity. To prove that the chain (1.15) is ergodic, we first need to establish several further properties of the chain.

It is a well-known fact that solutions of stochastic differential equations for which the coefficients satisfy the Lipschitz condition (1.2) also satisfy the Feller property (see, for instance, [9, p. 126]). It is a relatively simple matter to establish the (weak) Feller property for the chain (1.15), since the first p elements of the chain are simply successive samplings of the solution of a stochastic differential equation, and the distribution of the last element (the control term) at time $(t+1)$ depends continuously on the state of the chain at time t.

We will also show that the chain $\{Z_t\}$ is irreducible with respect to μ^{p+1}, the $(p+1)$-dimensional Lebesgue measure, that is,

$$\text{(3.17)}\quad \mu^{p+1}(A)>0 \Rightarrow \sum_{k=1}^{\infty}P^k(x,A)>0 \qquad \text{for all } x\in\mathbb{R}^q, \text{for all } A\in\mathcal{B}(\mathbb{R}^{p+1}),$$

and that the chain is aperiodic. The following lemma will assist us in doing this.

LEMMA 3.3. *Let $\{X_t,\ t=0,1,2,\dots\}$ be a time-homogeneous Markov chain taking values in $\mathbb{R}^q$. Suppose that for some $k>0$ and for each initial value $X_0\in\mathbb{R}^q$, X_k has a probability density which is positive everywhere in $\mathbb{R}^q$. Then the chain is μ^q-irreducible and aperiodic.*

Proof. Relation (3.17) clearly holds for $\phi=\mu^q$ since $P^k(x,\cdot)$ has a positive density. Aperiodicity follows from the observation that, for any $y\in\mathbb{R}^q$, $P^{k+1}(y,\cdot)$ has the same property due to the Chapman–Kolmogorov equation, and hence $P^k(y,\cdot)$ and $P^{k+1}(y,\cdot)$ are equivalent measures, which precludes any periodic behavior. □

In order to use this lemma, first note that $\boldsymbol{y}_1$ has a conditional probability density everywhere-positive in $\mathbb{R}^p$ (for each $Z_0=(\boldsymbol{y}_0,u_0)$). To see this, recall that $\boldsymbol{y}_1=\boldsymbol{x}(\delta)$ is simply the solution of the stochastic differential equation (1.1) at the fixed time $t=\delta$. By Girsanov's theorem (see, e.g., [9]), the distribution of this solution (in the space $C_p[0,\delta]$ of $\mathbb{R}^p$-valued continuous functions) is equivalent to that of the solution $\tilde{\boldsymbol{x}}(t)$ of the system

$$\begin{aligned} d\tilde{x}_1(t) &= \sigma dw(t), \\ d\tilde{x}_j(t) &= \tilde{x}_{j-1}(t)dt, \qquad j = 2, \dots, p. \end{aligned}$$

Therefore the distributions of $\boldsymbol{y}_1$ and $\tilde{\boldsymbol{x}}(\delta)$ are also equivalent, being the projections of equivalent measures. Furthermore, it can be shown (see, e.g., Lemma 3.2 of [10]) that $\tilde{\boldsymbol{x}}(t)$ has a nondegenerate p-variate Gaussian distribution, and hence has a density which is positive everywhere. It is then a direct consequence of the Radon–Nikodym theorem (see, e.g., [1]) that the density of $\boldsymbol{y}_1$ has the same property.

Now we will show that the two-step transition probabilities of the chain (1.15) have everywhere-positive densities in $\mathbb{R}^{p+1}$ (in fact the one-step transition functions are degenerate). From the control law (1.12), we know that given any $(\boldsymbol{y}_0, u_0)$, u_1 is a linear function of $\boldsymbol{y}_1$. This implies that, given any $(\boldsymbol{y}_0, u_0)$, the conditional distribution of $(\boldsymbol{y}_1, u_1)$ is concentrated on the hyperplane

$$u = \boldsymbol{a} \cdot \boldsymbol{y} + \delta^{-1}\boldsymbol{e}_1 \cdot \boldsymbol{y}_0 + u_0,$$

$$\boldsymbol{a} = (h_1 - \delta^{-1}, h_2, \dots, h_p)$$

and has an everywhere-positive density with respect to the respective p-dimensional volume measure on this hyperplane. Hence the random variable $v_1 = \delta^{-1}\boldsymbol{e}_1 \cdot \boldsymbol{y}_1 + u_1$ also has a (conditional) density ν_{y_0,u_0} which is positive everywhere on the real line.

The same argument shows that the conditional distribution of $(\boldsymbol{y}_2, u_2)$ given $(\boldsymbol{y}_1, u_1)$ is concentrated on the hyperplane $u = \boldsymbol{a} \cdot \boldsymbol{y} + v_1$ and also has a density on this hyperplane which is everywhere positive. Hence the conditional distribution of $(\boldsymbol{y}_2, u_2)$ given v_1, being a mixture of such distributions, also has a density $f_{v_1}(\boldsymbol{y}, u)$ which is positive everywhere on the hyperplane. From this it follows that the joint conditional distribution of $(\boldsymbol{y}_2, u_2)$ given the initial state $(\boldsymbol{y}_0, u_0)$ has a density (with respect to μ^{p+1}) which is proportional to the product $f_v(\boldsymbol{y}, u)\nu_{y_0,u_0}(v)$, with $v = \delta^{-1}\boldsymbol{e}_1 \cdot \boldsymbol{y} + u$. This density is clearly positive everywhere in $\mathbb{R}^{p+1}$.

Applying Lemma 3.3 (with $k = 2$), we see that the chain (1.15) is irreducible and aperiodic.

In addition, we have already seen that the first absolute moments $\mathbf{E}[\|Z_t\|]$ are bounded. It follows then from Chebyshev's inequality that the chain must be bounded in probability, that is, its transition probability functions are tight: for any $\epsilon > 0$ and any $x \in \mathbb{R}^q$, there exists a compact set C such that

$$\liminf_{k\to\infty} P^k(x, C) \geq 1 - \epsilon.$$

Finally, having established that the chain $\{Z_t\}$ is an irreducible, aperiodic Feller chain which is bounded in probability, we can apply Theorems 6.0.1 (iii) and 18.0.2 (i) of [8] to see that the chain is ergodic. This completes the proof of Theorem 2.1.

4. Simulation. To illustrate our results, we simulate the control of a particle subjected to a force (in one dimension). This force consists of a constant component, a frictional component which is proportional to velocity, a nonlinearity, and an external control. More precisely, we consider the second order ($p = 2$) system specified by

$$f(x, \dot{x}) = 2 - 0.5\dot{x} + 10\max(0, (1 - |x - 6|)),$$

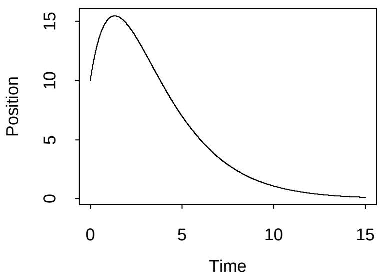


FIG. 4.1. *Position vs. time.*

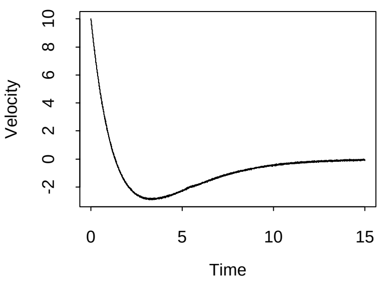


FIG. 4.2. *Velocity vs. time.*

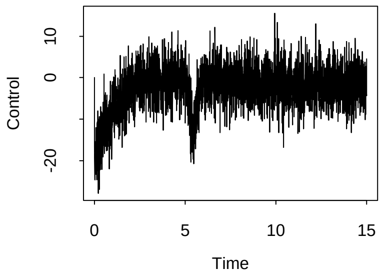


FIG. 4.3. *Control vs. time.*

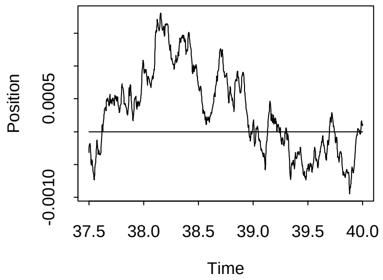


FIG. 4.4. *Position vs. time (small-scale).*

with noise intensity $\sigma = 0.3$ and sampling interval $\delta = 0.005$. The sharp nonlinearity around the point $x = 6$ is included so as to test the adaptive behavior of the controller and to verify stability of the closed-loop system in the presence of such irregularities (recall that the only information available to the controller is the order of the system and a Lipschitz constant associated with the function $f(\cdot,\cdot)$). We use the reference system specified by

$$h(x, \dot{x}) = -\dot{x} - x/4,$$

so that the matrix H has one eigenvalue $\lambda_H = -1/2$ with multiplicity two, and $\|H\| \simeq 1.4254$. This choice of reference system is arbitrary (subject to its stability, of course). Clearly, by choosing a reference system with different eigenvalues, our closed-loop system could be made to exhibit different characteristics, such as a different rate of convergence to the stable regime, more or less control effort expended, etc. It is easily verified that the system and reference system satisfy the conditions of Theorem 2.1, with $\beta = 10$, $k_1(\delta) \simeq 0.3332$, $\delta_0 = 2$, and $h = 1/2$. Figures 4.1 and 4.2 display position and velocity as a function of time for a particular realization of the controlled system. The "settling down" behavior which becomes apparent as time increases is exactly what Theorem 2.1 leads us to expect.

Figure 4.3 illustrates the control action (specified by (1.12)) used to stabilize the system. The adaptive action of the controller becomes apparent just after the time $t = 5$, when the state enters the region $5 \le x \le 7$. The small-scale plot of position for $37.5 \le t \le 40$ given in Figure 4.4 shows that the controller does indeed restrict the system to a narrow neighborhood around the origin as time increases.

The result appears to be quite conservative, since the sampling rate requirements can be relaxed significantly without destabilizing the system. In the example above, the maximal value of δ which satisfies the condition (2.2) is approximately 0.0056.

However, simulations indicate that the critical value of δ for which the system becomes unstable is between 1.5 and 1.6. In part, this could be explained by the fact that the system only spends a minimal amount of time in the region of the nonlinearity ($5 \leq x \leq 7$), so the "effective Lipschitz constant" could be much smaller than 10. For systems in which $f(\cdot)$ varies more around 0, we could expect a smaller discrepancy between the critical value of δ determined by (2.2) and the critical value indicated by simulation.

Acknowledgment. The first author is grateful to Kais Hamza for valuable discussions and guidance during the early stages of this research.

REFERENCES

[1] P. Billingsley, *Probability and Measure*, 2nd ed., John Wiley, New York, 1986.

[2] P. Caines, *Linear Stochastic Systems*, John Wiley, New York, 1988.

[3] W. Fleming and R. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[4] P. Florchinger, *Lyapunov-like techniques for stochastic stability*, SIAM J. Control Optim., 33 (1995), pp. 1151–1169.

[5] P. Henrici, *Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices*, Numer. Math., 4 (1962), pp. 24–40.

[6] I. Mareels, H. Penfold, and R. Evans, *Controlling nonlinear time-varying systems via Euler approximations*, Automat. J. IFAC, 28 (1992), pp. 681–696.

[7] S.P. Meyn and L. Guo, *Stability, convergence, and performance of an adaptive control algorithm applied to a randomly varying system*, IEEE Trans. Automat. Control, 37 (1992), pp 535–540.

[8] S.P. Meyn and R.L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, New York, 1993.

[9] B. Øksendal, *Stochastic Differential Equations*, Springer-Verlag, New York, 1995.

[10] O. Stramer, R.L. Tweedie, and P.J. Brockwell, *Existence and stability of continuous-time threshold ARMA processes*, Statist. Sinica, 6 (1996) pp. 15–732.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 5, pp. 1568–1599

ON THE OBSERVABILITY INEQUALITIES FOR EXACT CONTROLLABILITY OF WAVE EQUATIONS WITH VARIABLE COEFFICIENTS*

PENG-FEI YAO†

Abstract. This paper deals with boundary exact controllability for the dynamics governed by the wave equation with nonconstant coefficients in the principal part, subject to Dirichlet or Neumann boundary controls. The observability inequalities are established by the Riemannian geometry method under some geometric condition for the Dirichlet problem and for the Neumann problem, respectively. Next, a number of nontrivial examples are presented to verify the observability inequality. In particular, a counterexample is given without boundary exact controllability, where the control is exerted on the whole boundary.

Key words. wave equation, exact controllability, Riemannian manifold, Hessian comparison theorem, geometric optics

AMS subject classifications. 35A, 35l, 35q, 49A, 49B, 49E

PII. S0363012997331482

1. Introduction.

Let Ω be a bounded domain in R^n with smooth boundary $\partial\Omega{=}\Gamma$. It is assumed that Γ consists of two parts: Γ_0 and Γ_1, $\Gamma_0\cup\Gamma_1{=}\Gamma$, with Γ_0 nonempty and relatively open in Γ. In the cylinder $\Omega\times(0,T)$, we consider the exact controllability for the mixed problem

$$\begin{cases} y_{tt}-\displaystyle\sum_{ij=1}^{n}\frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial y}{\partial x_j}\right)=0 \quad \text{in} \quad \Omega\times(0,T),\\ y(x,0)=y^0(x), \quad y_t(x,0)=y^1(x) \quad \text{in} \quad \Omega, \end{cases} \tag{1.1}$$

$$y=0 \quad \text{in} \quad \Gamma_1\times(0,T), \qquad y=v \quad \text{in} \quad \Gamma_0\times(0,T), \tag{1.1a}$$

where y_{tt} stands for $\partial^2y/\partial t^2$, $a_{ij}=a_{ji}$ are C^∞ functions in $\mathbb{R}^n$, and

$$\sum_{ij=1}^{n}a_{ij}(x)\xi_i\xi_j\geq a\sum_{i=1}^{n}\xi_i^2, \quad x\in\Omega, \tag{1.2}$$

for some constant $a>0$.

We ask whether there is some constant $T_0>0$ such that if $T>T_0$, the following steering property of (1.1) and (1.1a) holds true: for all initial data y^0, y^1 in $L^2(\Omega)\times H^{-1}(\Omega)$, there exists a suitable control function v in $L^2(0,T;L^2(\Gamma_0))$, whose corresponding solution of (1.1), (1.1a) satisfies

$$y(\cdot,T)\equiv y_t(\cdot,T)\equiv 0. \tag{1.3}$$

When the answer is in the affirmative, we then say that the dynamics (1.1) and (1.1a) is exactly controllable in the interval $[0,T]$ on $L^2(\Omega)\times H^{-1}(\Omega)$ by means of the Dirichlet control function $v\in L^2(0,T;L^2(\Gamma_0))$.

*Received by the editors December 12, 1997; accepted for publication (in revised form) October 22, 1998; published electronically August 2, 1999. This research was partially supported by the National Natural Science Foundation of China and the National Key Project of China.

http://www.siam.org/journals/sicon/37-5/33148.html

†Institute of Systems Science, Academia Sinica, Beijing 100080, China (pfyao@iss03.iss.ac.cn).

Exact controllability problems are best handled by duality or transposition, rather than directly.

Set

$$\mathcal{A}u = -\sum_{ij=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right)$$

and consider the corresponding homogeneous problem

$$\text{(1.4)} \qquad \begin{cases} \phi_{tt} + \mathcal{A}\phi = 0 \quad \text{in} \quad \Omega \times (0,T), \\ \phi(0) = \varphi^0, \quad \phi_t(0) = \varphi^1 \quad \text{in} \quad \Omega, \\ \phi = 0 \quad \text{on} \quad \Gamma \times (0,T). \end{cases}$$

Given $\varphi^0 \in H_0^1(\Omega)$, $\varphi^1 \in L^2(\Omega)$, (1.4) admits a unique solution $\{\,\phi,\, \phi_t\,\} \in C([0,T]; H_0^1(\Omega) \times L^2(\Omega))$ satisfying the trace regularity $\frac{\partial \phi}{\partial v_{\mathcal{A}}} \in L^2(0,T;L^2(\Gamma))$ (Lasiecka, Lions, and Triggiani [9]); i.e., more precisely, for all $T > 0$,

$$\text{(1.5)} \qquad \int_0^T \int_{\Gamma_0} \left(\frac{\partial \phi}{\partial v_{\mathcal{A}}}\right)^2 d\sigma dt \le cT\left(\|\varphi^0\|^2_{H_0^1(\Omega)} + \|\varphi^1\|^2_{L^2(\Omega)}\right),$$

where

$$\frac{\partial \phi}{\partial v_{\mathcal{A}}} = \sum_{ij=1}^{n} a_{ij}\frac{\partial \phi}{\partial x_j} v_i$$

is the conormal derivative, $v = (v_1, v_2, \ldots, v_n)$ is the unit normal of Γ pointing towards the exterior of Ω, and $d\sigma$ is the Euclidean surface element on Γ.

Let ψ be the solution of the following problem:

$$\text{(1.6)} \qquad \begin{cases} \psi_{tt} + \mathcal{A}\psi = 0 \quad \text{in} \quad \Omega \times (0,T), \\ \psi(T) = \psi_t(T) = 0 \quad \text{in} \quad \Omega, \\ \psi|_{\Gamma_1} = 0, \qquad 0 < t < T, \\ \psi|_{\Gamma_0} = \dfrac{\partial \phi}{\partial v_{\mathcal{A}}}, \qquad 0 < t < T, \end{cases}$$

with $\frac{\partial \phi}{\partial v_{\mathcal{A}}}$ produced by (1.4). Given $\varphi^0 \in H_0^1(\Omega)$ and $\varphi^1 \in L^2(\Omega)$, we have defined, in a unique fashion, $\Lambda : H_0^1(\Omega) \times L^2(\Omega) \longrightarrow H^{-1}(\Omega) \times L^2(\Omega)$ by

$$\text{(1.7)} \qquad \Lambda \begin{pmatrix} \varphi^0 \\ \varphi^1 \end{pmatrix} = \begin{pmatrix} \psi_t(0) \\ -\psi(0) \end{pmatrix}.$$

A formal use of Green's formula gives, after we multiply (1.6) by ϕ and integrate by parts,

$$\text{(1.8)} \qquad \left\langle \Lambda \begin{pmatrix} \varphi^0 \\ \varphi^1 \end{pmatrix}, \begin{pmatrix} \varphi^0 \\ \varphi^1 \end{pmatrix} \right\rangle = \int_0^T \int_{\Gamma_0} \left(\frac{\partial \phi}{\partial v_{\mathcal{A}}}\right)^2 d\sigma dt,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product over $L^2(\Omega) \times L^2(\Omega)$.

By (1.4)–(1.8), getting the exact controllability for (1.1) and (1.1a) over space $L^2(\Omega) \times H^{-1}(\Omega)$ is equivalent to showing that there are constants c, $T_0 > 0$ such that the following observability inequality holds true for $T > T_0$:

$$\text{(1.9)} \qquad \int_0^T \int_{\Gamma_0} \left(\frac{\partial \phi}{\partial v_{\mathcal{A}}}\right)^2 d\sigma dt \ge c\left(\|\varphi^0\|^2_{H_0^1(\Omega)} + \|\varphi^1\|^2_{L^2(\Omega)}\right)$$

for all $\varphi^0 \in H_0^1(\Omega)$, $\varphi^1 \in L^2(\Omega)$, where ϕ solves problem (1.4).

This problem has received considerable attention in the literature, with numerous contributions achieved over the past several years. For the constant coefficient case, see Chen [3], Lagnese [7, 8], Lions [12], and Triggiani [18]. For the variable coefficient case, the observability inequalities were obtained in Bardos, Lebeau, and Rauch [1] and in Tataru [15, 16, 17]. In [1], the observability inequality is established subject to some geometric optics conditions that are almost necessary conditions for the exact controllability. However, as explicitly recognized by the authors, for $a_{ij}(x)$, Ω given, except for the constant coefficient case, it is not an easy matter to verify the (sharp) condition that all rays hit Γ_0 at a nondiffractive point since rays are solutions of a system of nonlinear ordinary equations (Hamilton). In Tataru [15, 16, 17], the observability inequality is obtained by the Carleman estimate, subject to the existence of a pseudoconvex function. Both works above succeed in handling arbitrary first-order terms (energy level) as well as time-dependent coefficients.

We note here that nontrivial examples have not been available in the literature until now.

In the present paper, we consider observability inequality (1.9) by the Riemannian geometry method, subject to a different geometric condition (1.16), which is motivated by geometric multiplier identities. Several multiplier identities, which have been built for constant coefficient wave equation (Lions [12]), are generalized to the variable coefficient case by some computational techniques in Riemannian geometry so that observability inequality (1.9) is derived from those identities. The situation here is very similar to that in the constant coefficient case.

If there is a strictly convex function ϕ on $\overline{\Omega}$, then vector field $H = D\phi$ meets condition (1.16), where D is a connection that will be specified later. At the same time, since $\phi - ct^2$ is pseudoconvex for some constant c, Tataru's work can be applied so that some of the examples, given in section 3, can be derived from Tataru's work. However, in general, vector field H is not the covariant differential of a function (for instance, see Example 3.4).

An interesting problem for exact controllability is to give the time T_0 that the control needs. Similar to the case of constant coefficients, our estimates present an explicit formula for T_0. In addition, our approach also works well for the Euler–Bernoulli equation (Yao [20]) and is available for some nonlinear boundary feedback problems.

The sufficient and necessary conditions, given by Bardos, Lebeau, and Rauch [1], show that if Γ_0 is "small," the exact controllability may not hold true. Such a counterexample is presented by Ralston [13] for the constant coefficient case. It is, however, well known that in the case of constant coefficients one can always find Γ_0 ("large" enough) such that the exact controllability holds true. The situation is very different in the case of variable coefficients. Indeed, by a combination of Riemannian geometry and geometric optics, section 4 presents a counterexample without exact controllability, where the control is exerted on the whole boundary; that is, there exists a ray that never hits the boundary.

We note that for the case $n = 1$ the exact controllability always holds true without geometric condition; that is, the one-dimensional problem (space dependent) fits all the theories above. We give a sketch of the proof for the case $n = 1$ in the appendix at the end of this paper.

Now we are back to our main problem and we focus on observability inequality (1.9), or equivalently, on the following question, raised by Lions [12]. Set

$$F_{T,\Gamma_0} = \left\{ (\varphi^0, \varphi^1) \,|\, (\varphi^0, \varphi^1) \in H_0^1(\Omega) \times L^2(\Omega), \int_0^T \int_{\Gamma_0} \left(\frac{\partial \phi}{\partial v_{\mathcal{A}}} \right)^2 dx dt < \infty \right\}. \tag{1.10}$$

Open question. How can we characterize the Hilbert space F_{T,Γ_0}? Is F_{T,Γ_0} always equal to $H_0^1(\Omega) \times L^2(\Omega)$?

We introduce some notation. Suppose that

$$\sum_{ij=1}^{n} a_{ij}(x)\xi_i\xi_j > 0 \quad \forall\, x \in \mathbb{R}^n, \quad \xi = (\xi_1, \xi_2, \ldots, \xi_n)^\tau \in \mathbb{R}^n, \quad \xi \neq 0. \tag{1.11}$$

Set

$$A(x) = (a_{ij}(x)), \tag{1.12}$$

$$G(x) = (g_{ij}(x)) = A(x)^{-1}. \tag{1.13}$$

Let $\mathbb{R}^n$ have the usual topology and $x = (x_1, x_2, \ldots, x_n)$ be the natural coordinate system. For each $x \in \mathbb{R}^n$, define the inner product and norm over the tangent space $\mathbb{R}^n_x = \mathbb{R}^n$ by

$$g\,(X,\, Y) = \langle X,\, Y \rangle_g = \sum_{ij=1}^{n} g_{ij}(x)\alpha_i\beta_j, \tag{1.14}$$

$$|X|_g = \langle X,\, X \rangle_g^{1/2} \quad \forall \quad X = \sum_{i=1}^{n} \alpha_i \frac{\partial}{\partial x_i}, \quad Y = \sum_{i=1}^{n} \beta_i \frac{\partial}{\partial x_i} \in \mathbb{R}^n_x.$$

It is easily checked from (1.11) that $(\mathbb{R}^n, g)$ is a Riemannian manifold with Riemannian metric g. Denote the Levi–Civita connection in metric g by D. Let H be a vector field on $(\mathbb{R}^n, g)$. The covariant differential DH of H determines a bilinear form on $\mathbb{R}^n_x \times \mathbb{R}^n_x$, for each $x \in \mathbb{R}^n$, by

$$DH\,(X,\, Y) = \langle D_X H,\, Y \rangle_g \quad \forall\; X, Y \in \mathbb{R}^n_x,$$

where $D_X H$ is the covariant derivative of vector field H with respect to X.

For each $x \in \mathbb{R}^n$, denote

$$X \cdot Y = \sum_{i=1}^{n} \alpha_i\beta_i, \quad |X|_0 = (X \cdot X)^{1/2} \quad \forall\, X = \sum_{i=1}^{n} \alpha_i \frac{\partial}{\partial x_i}, \quad Y = \sum_{i=1}^{n} \beta_i \frac{\partial}{\partial x_i} \in \mathbb{R}^n_x,$$

the usual dot product over $\mathbb{R}^n$. For $x \in \mathbb{R}^n$, set

$$A(x)X = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij}(x)\alpha_j \right) \frac{\partial}{\partial x_i} \quad \forall \quad X = \sum_{i=1}^{n} \alpha_i \frac{\partial}{\partial x_i} \in \mathbb{R}^n_x. \tag{1.15}$$

Let

$$X = \sum_{i=1}^{n} \alpha_i(x) \frac{\partial}{x_i}$$

be a vector field on $\mathbb{R}^n$. Denote the divergence of X in the Euclidean metric by $\operatorname{div}_0(X)$. Then

$$\operatorname{div}_0(X) = \sum_{i=1}^{n} \frac{\partial \alpha_i(x)}{\partial x_i}.$$

For $f \in C^1(\overline{\Omega})$, denote the gradients of f by ∇_0 and ∇_g in the Euclidean metric and in Riemannian metric g, respectively.

Let $x^0 \in \mathbb{R}^n$. Let r: $(0,\infty) \to \mathbb{R}^n$ be a geodesic with $r(0) = x^0$ parameterized by arc length in the Riemannian metric g. Denote the distance on $(\mathbb{R}^n, g)$ by d_g. For sufficiently small $t > 0$, we know that $d_g(r(t), x^0) = t$, since the exponential map $\exp_{x^0}$: $\mathbb{R}^n_{x^0} \to \mathbb{R}^n$ is injective on a sufficiently small ball in $(\mathbb{R}^n, g)$. We recall that $r(t_0)$ is called the cut point of r with respect to x^0, if $t_0 > 0$ is such that $d_g(r(t), x^0) = t$, $0 \le t < t_0$, and $d_g(r(t), x^0) < t$ for all $t > t_0$ (Cheeger and Ebin [2]). The union of all cut points is called the cut locus of x^0 and is denoted by $cut(x^0)$. For any $X \in \mathbb{R}^n_{x^0}$, $|X|_g = 1$, there is at most one cut point on the geodesic $\exp_{x^0} tX$ $(t \ge 0)$. Thus $cut(x^0)$ is the image of the exponential map on some closed subset of S^{n-1} and the n-dimensional measure of $cut(x^0)$ is zero. Set $\mu(X) = d_g(x^0, r(t_0))$ if $r(t_0)$ is the cut point of x^0 along $r(t) = \exp_{x^0} tX$; set $\mu(X) = \infty$ when there is no cut point of x^0 along r, where $X \in S^{n-1} \subset \mathbb{R}^n_{x^0}$. We define

$$E = \left\{ tX \,|\, 0 \le t < \mu(X),\ X \in S^{n-1} \subset \mathbb{R}^n_{x^0} \right\}.$$

Then $\exp_{x^0}$: $E \to \exp_{x^0}(E)$ is a diffeomorphism. It is obvious that $\exp_{x^0}(E)$ is a star domain and

$$\mathbb{R}^n = \exp_{x^0}(E) \cup cut(x^0).$$

Now we are in a position to state our main results.

THEOREM 1.1. *Let H be a vector field on Riemannian manifold $(\mathbb{R}^n, g)$ such that*

$$\langle D_X H,\, X\rangle_g \ge a|X|_g^2 \quad \forall\, X \in \mathbb{R}^n_x, \quad x \in \Omega, \tag{1.16}$$

for some constant $a > 0$. Then there exists $T_0 > 0$ such that for any $T > T_0$,

$$F_{T,\Gamma_0} = H_0^1(\Omega) \times L^2(\Omega),$$

where

$$\Gamma_0 = \left\{ x \;\middle|\; x \in \Gamma,\ H(x)\cdot \upsilon(x) > 0 \right\}, \tag{1.17}$$

$$T_0 = \frac{2}{a} \sup_{x \in \overline{\Omega}} |H|_g(x). \tag{1.18}$$

Remark 1.1. In the case of constant coefficients, where $a_{ij}(x) = \delta_{ij}$, the above results were obtained by Lions [12] for the radial field $H = x - x_0$. Komornik [6] improved the estimate T_0.

In general, it is not easy to find a vector field H verifying condition (1.16). In the following, we relate the existence of such a vector field to sectional curvature by several corollaries that are useful for the examples in section 3.

If $(\mathbb{R}^n, g)$ is a noncompact complete Riemannian manifold of everywhere positive sectional curvature, there exists a C^∞ strictly convex function h on (R^n, g) by Greene and Wu [4]. All the eigenvalues of its second covariant differential are everywhere positive. Take $H = Dh$, and we have

$$DH\,(X,\,X) = D^2h\,(X,\,X) > 0 \quad \forall \quad X \in \mathbb{R}^n_x, \quad x \in \mathbb{R}^n,$$

where D^2h is the Hessian of h in Riemannian metric g. There is a constant $a > 0$ such that inequality (1.16) holds true, since $\overline{\Omega}$ is compact.

COROLLARY 1.1. *Let $(\mathbb{R}^n, g)$ be noncompact complete and of everywhere positive sectional curvature. h is a C^∞ strictly convex function on $(\mathbb{R}^n, g)$. Then vector field $H = Dh$ meets condition* (1.16) *for any $\Omega \subset \mathbb{R}^n$.*

For $x \in \mathbb{R}^n$, let $\Pi \subset \mathbb{R}^n_x$ be a two-dimensional subspace. Denote the sectional curvature by $K_x(\Pi)$. Let $B(x^0, \gamma)$ be a geodesic ball in $(\mathbb{R}^n, g)$ with the center at x^0 and the radius $\gamma > 0$. Set

$$k_B = \sup_{x \in B(x^0,\gamma),\, \Pi \subset \mathbb{R}^n_x} K_x(\Pi). \tag{1.19}$$

For $x^0 \in \mathbb{R}^n$, set $\rho(x) = d_g(x^0, x)$ $\forall x \in \mathbb{R}^n$.

COROLLARY 1.2. *Let $x^0 \in \mathbb{R}^n$. Let $\gamma > 0$ be such that $B(x^0, \gamma) \subset \exp_{x^0}(E)$. If $\Omega \subset B(x^0, \gamma)$ such that $4\gamma^2 k_B < \pi^2$, then vector field $H = \rho D\rho$ meets condition* (1.16) *with $a = \sqrt{k_B}\gamma cot(\sqrt{k_B}\gamma)$ if $k_B > 0$ and with $a = 1$ when $k_B \leq 0$.*

Remark 1.2. This corollary is a global result. For any a_{ij}, however, we can derive local boundary exact controllability from it since, for any $x^0 \in \mathbb{R}^n$, there exists a $\gamma > 0$ such that $B(x^0, \gamma) \subset \exp_{x^0}(E)$ (Cheeger and Ebin [2]). In addition, if g, defined by (1.14), is the usual dot product of $\mathbb{R}^n$, it is easily checked that

$$H = \rho D\rho = x - x^0.$$

If $(\mathbb{R}^n, g)$ is of everywhere nonpositive sectional curvature, then, for all $x^0 \in \mathbb{R}^n$, $\exp_{x^0}(E) = \mathbb{R}^n$ and $k_B \leq 0$ (see Spivak [14]). The following corollary is immediate from Corollary 1.2.

COROLLARY 1.3. *Let $(\mathbb{R}^n, g)$ be of nonpositive sectional curvature. For any $x^0 \in \mathbb{R}^n$, vector field $H = \rho D\rho$ satisfies condition* (1.16) *for any $\Omega \subset \mathbb{R}^n$ where $a = 1$.*

COROLLARY 1.4. *Assume that*

$$K \geq \sup_{x \in \mathbb{R}^n,\, \pi \subset \mathbb{R}^n_x} K_x(\pi) \geq \inf_{x \in \mathbb{R}^n,\, \pi \subset \mathbb{R}^n_x} K_x(\pi) > 0. \tag{1.20}$$

By the Bonnet theorem (Cheeger and Ebin [2]*), the completion $(\mathbb{R}^n \cup \{\infty\}, g)$ of $(\mathbb{R}^n, g)$ is a compact Riemannian manifold.*

Set

$$d_g(\Omega) = \sup_{x,y \in \Omega} d_g(x, y).$$

If

$$d_g(\Omega) < \frac{1}{2} \min \left\{ \frac{\pi}{\sqrt{K}}, \right.$$

$$\left. \textit{half the length of the shortest closed geodesic in } \mathbb{R}^n \right\}, \tag{1.21}$$

then vector field $H = \rho D\rho$ meets condition (1.16).

Here we turn to the Neumann action problem and examine problem (1.1) with, instead of the Dirichlet action in (1.1a), the Neumann action

$$\frac{\partial y}{\partial v_{\mathcal{A}}} = v \quad \text{on} \quad \Gamma \times (0,T). \tag{1.1b}$$

We say that the dynamic (1.1), (1.1b) is exactly controllable in the interval $[0,T]$ by means of control function v with Neumann action if, for φ^0, φ^1 in some Hilbert spaces, one can find v such that the solution of (1.1) and (1.1b) satisfies (1.3).

Theorem 1.2. *Let H be a vector field on Riemannian manifold* $(\mathbb{R}^n, g)$ *such that inequality* (1.16) *holds true. Let T_0 be given by* (1.18) *and $T > T_0$. Then, for any $\varphi^0 \in L^2(\Omega)$, $\varphi^1 \in (H^1(\Omega))'$, one can find v on $\Gamma \times (0,T)$ such that the solution of* (1.1) *and* (1.1b) *satisfies* (1.3)*. The control v has the following structure:*

$$v = \begin{cases} v_0 + \dfrac{\partial v_1}{\partial t} & on \quad \Gamma_0 \times (0,T), \\ v_2 & on \quad \Gamma_1 \times (0,T), \end{cases} \tag{1.22}$$

where

$$\begin{aligned} &v_0,\, v_1 \in L^2(\Gamma_0 \times (0,T)), \\ &v_2 \in L^2(0,T; (H^1(\Gamma_1))'). \end{aligned}$$

Similar to Lions [12], if Ω and H are subject to the geometrical conditions

$$H(x) \cdot v(x) \geq 0 \quad \forall\, x \in \Gamma, \tag{1.23}$$

then the control v can be taken in $L^2(\Gamma \times (0,T))$.

Theorem 1.3. *Let H be a vector field on Riemannian manifold* $(\mathbb{R}^n, g)$ *such that conditions* (1.16) *and* (1.23) *hold true. Let T_0 be given in* (1.18) *and $T > T_0$. Then, for any $\varphi^0 \in H^1(\Omega)$, $\varphi^1 \in L^2(\Omega)$, one can find $v \in L^2(\Gamma \times (0,T))$ such that the solution of* (1.1) *and* (1.1b) *satisfies* (1.3).

Finally, we consider the exact controllability problem for

$$\begin{cases} y_{tt} + \mathcal{A}y = 0 \quad \text{in} \quad \Omega \times (0,T), \\ y(0) = y^0, \quad y_t(0) = y^1, \\ y = 0 \quad \text{in } \Gamma_1, \quad \dfrac{\partial y}{\partial v_{\mathcal{A}}} = v \quad \text{on } \Gamma_0. \end{cases} \tag{1.24}$$

We have the Dirichlet condition (no action) on Γ_1 and a Neumann action on Γ_0. Set

$$H^1_{\Gamma_1}(\Omega) = \{\, u \,|\, u \in H^1(\Omega),\, u|_{\Gamma_1} = 0 \,\}.$$

In this case, the preceding approach leads to the following result. For the case of the Laplace $\mathcal{A} = -\Delta$, the same result was obtained by Lions [12] and Lasiecka and Triggiani [10]. It is mentioned that, unlike the Dirichlet case (see (1.5)), the corresponding regularity inequality does not hold true for dimension $\Omega \geq 2$.

Theorem 1.4. *Let H be a vector field on Riemannian manifold* $(\mathbb{R}^n, g)$ *such that condition* (1.16) *holds true. Let T_0 be given in* (1.18)*. Then, for any $T > T_0$, $y^0 \in L^2(\Omega)$, $y^1 \in (H^1_{\Gamma_1}(\Omega))'$, we can find v such that the solution y of problem* (1.24) *satisfies* (1.3)*, where*

$$v = v_0 + \frac{\partial v_1}{\partial t}, \quad v_0,\, v_1 \in L^2(\Gamma_0 \times (0,T)).$$

2. The proof of the results.

2.1. Multiplier identities. To estimate the observability inequality (1.9), we need some multiplier identities which have been built for the classical wave equations, where $a_{ij} = \delta_{ij}$ (Lions [12], Komornik [6], Lasiecka and Triggiani [10]). We now consider their generalizations in the variable coefficient situations.

With two metrics on $\mathbb{R}^n$ in mind, one the Euclidean metric and the other Riemannian metric g, we have to deal with various notations carefully.

We recall that $E_1, E_2, \dots, E_n$ is a frame field normal at x (Wu, Shen, and Yu [19]) on Riemannian manifold $(\mathbb{R}^n, g)$ if and only if it is a local basis for vector fields with $\langle E_i, E_j\rangle_g = \delta_{ij}$ in some neighborhood of x and with $(D_{E_i}E_j)(x) = 0$ for $1 \le i, j \le n$. Let H be a vector field on $\mathbb{R}^n$ and $f \in C^1(\overline{\Omega})$. We have the formulae for divergence in the Euclidean metric

$$\operatorname{div}_0(fH) = f\operatorname{div}_0(H) + H(f) \tag{2.1}$$

and

$$\int_\Omega \operatorname{div}_0(H)\,dx = \int_\Gamma H\cdot v\,d\sigma. \tag{2.2}$$

We shall write $(\phi,\, \psi)$ for $\int_\Omega \phi\psi\,dx$ and $\|\phi\|$ for $(\phi,\, \phi)^{1/2}$.

LEMMA 2.1. *Let $x = (x_1, x_2, \dots, x_n)$ be the natural coordinate system in $\mathbb{R}^n$, f, $h \in C^1(\overline{\Omega})$, and H, X vector fields. Then*

(1)

$$\langle H(x),\, A(x)X(x)\rangle_g = H(x)\cdot X(x), \quad x \in \mathbb{R}^n; \tag{2.3}$$

(2)

$$\nabla_g f(x) = \sum_{i=1}^n \left(\sum_{j=1}^n a_{ij}(x)\frac{\partial f}{\partial x_j}\right)\frac{\partial}{\partial x_i}, \quad x \in \mathbb{R}^n; \tag{2.4}$$

(3)

$$\langle \nabla_g f,\, \nabla_g h\rangle_g = \nabla_g f(h) = \nabla_0 f\cdot A(x)\nabla_0 h, \quad x\in\mathbb{R}^n; \tag{2.5}$$

(4)

$$\begin{aligned}\langle \nabla_g f,\, \nabla_g(H(f))\rangle_g(x) &= DH(\nabla_g f,\, \nabla_g f)(x) + \frac{1}{2}\operatorname{div}_0\left(|\nabla_g f|_g^2 H\right)(x)\\ &\quad -\frac{1}{2}|\nabla_g f|_g^2(x) div_0(H)(x) \quad \forall\, x\in\mathbb{R}^n,\end{aligned} \tag{2.6}$$

where $A(x)$ is given by (1.12).

Proof. Set

$$H(x) = \sum_{i=1}^n h_i\frac{\partial}{\partial x_i} \quad \text{and} \quad X(x) = \sum_{i=1}^n f_i\frac{\partial}{\partial x_i}.$$

It follows from (1.15) that

$$\langle H(x),\, A(x)X(x)\rangle_g = \sum_{ij=1}^n\sum_{k=1}^n a_{ik}f_k h_j g_{ij} = \sum_{k=1}^n f_k h_k = H(x)\cdot X(x), \tag{2.7}$$

where

$$g_{ij} = \left\langle \frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j} \right\rangle_g .$$

Set

$$\nabla_g x_i = \sum_{j=1}^n x_{ij} \frac{\partial}{\partial x_j}, \quad 1 \le i \le n, \tag{2.8}$$

$$C(x) = (x_{ij}(x)), \quad x \in \mathbb{R}^n, \tag{2.9}$$

an $n \times n$ matrix. It follows from (2.8) that

$$x_{ij} = \nabla_g x_i(x_j) = \langle \nabla_g x_i, \nabla_g x_j \rangle_g = \sum_{k,l=1}^n g_{kl} x_{ik} x_{jl}, \quad 1 \le i, j \le n. \tag{2.10}$$

Equation (2.10) gives

$$C(x) = C(x)^\tau G(x) C(x), \quad x \in \mathbb{R}^n,$$

where $C(x)^\tau$ is the transpose of the matrix $C(x)$ and $G(x)$ is defined by (1.13); that is,

$$C(x) = G^{-1}(x) = A(x), \quad x \in \mathbb{R}^n. \tag{2.11}$$

From (2.8) and (2.11), we obtain

$$\nabla_g f = \sum_{i=1}^n \nabla_g f(x_i) \frac{\partial}{\partial x_i} = \sum_{i=1}^n \nabla_g x_i(f) \frac{\partial}{\partial x_i} = \sum_{i=1}^n \left(\sum_{j=1}^n a_{ij} \frac{\partial f}{\partial x_j} \right) \frac{\partial}{\partial x_i}, \quad x \in \mathbb{R}^n.$$

This yields

$$\langle \nabla_g f, \nabla_g h \rangle_g = \nabla_g f(h) = \sum_{ij=1}^n a_{ij} \frac{\partial f}{\partial x_i} \frac{\partial h}{\partial x_j} = \nabla_0 f(x) \cdot A(x) \nabla_0 h(x), \quad x \in \mathbb{R}^n.$$

We now prove part 4. Let $x \in \mathbb{R}^n$. Let $E_1, E_2, \ldots, E_n$ be a frame field normal at x. There are functions $h_1, h_2, \ldots, h_n$ on some neighborhood of x such that $H = \sum_{i=1}^n h_i E_i$. In addition, we have

$$H(f) = \sum_{i=1}^n h_i E_i(f), \tag{2.12}$$

$$\nabla_g f = \sum_{i=1}^n E_i(f) E_i, \quad \text{and} \quad |\nabla_g f|_g^2 = \sum_{j=1}^n (E_j(f))^2, \tag{2.13}$$

where $E_i(f)$ are the covariant differential of f with regard to E_i in Riemannian metric g. For the covariant differential of vector field H, we obtain

$$\begin{aligned} DH(E_i, E_j)(x) &= \langle D_{E_i} H(x), E_j(x) \rangle_g \\ &= \sum_{k=1}^n \langle E_i(h_k) E_k + h_k D_{E_i} E_k, E_j \rangle_g \\ &= E_i(h_j)(x), \end{aligned} \tag{2.14}$$

since $\langle E_k, E_j\rangle_g = \delta_{kj}$ and $D_{E_i}E_k(x) = 0$. It follows from (2.12)–(2.14) and (2.1) that

$$\begin{aligned}\langle \nabla_g f,\, \nabla_g (H(f))\rangle_g (x) &= \sum_{j=1}^n E_j(f) E_j(H(f))\\ &= \sum_{j=1}^n E_j(f)\left(\sum_{i=1}^n E_j(h_i)E_i(f) + \sum_{i=1}^n h_i E_j E_i(f)\right)\\ &= \sum_{ij=1}^n E_j(h_i)E_i(f)E_j(f) + \sum_{i=1}^n h_i\left(\sum_{j=1}^n E_j(f)E_iE_j(f)\right)\\ &= DH(\nabla_g f,\, \nabla_g f) + \frac{1}{2}H\left(|\nabla_g f|_g^2\right)\\ &= DH(\nabla_g f,\, \nabla_g f)(x) + \frac{1}{2}\mathrm{div}_0\left(|\nabla_g f|_g^2 H\right)(x)\\ &\qquad - \frac{1}{2}|\nabla_g f|_g^2(x)\mathrm{div}_0(H)(x) \quad \forall\, x \in \mathbb{R}^n,\end{aligned}$$

where $E_iE_j(f)(x) = E_jE_i(f)(x)$ are the second covariant differential of f. □

PROPOSITION 2.1. *Let ϕ solve problem*

$$\phi_{tt} + \mathcal{A}\phi = 0 \quad on \quad \Omega \times (0, T). \tag{2.15}$$

Suppose that H is a vector field on $\overline{\Omega}$. Then

(1)

$$\begin{aligned}&\int_0^T\int_\Gamma \frac{\partial\phi}{\partial\upsilon_{\mathcal{A}}}H(\phi)\,d\sigma dt + \frac{1}{2}\int_0^T\int_\Gamma\left(\phi_t^2 - |\nabla_g\phi|_g^2\right)H\cdot\nu\,d\sigma dt\\ &= (\phi_t,\, H(\phi))\Big|_0^T + \int_0^T\int_\Omega DH(\nabla_g\phi,\, \nabla_g\phi)\,dxdt\\ &+\frac{1}{2}\int_0^T\int_\Omega\left(\phi_t^2 - |\nabla_g\phi|_g^2\right)\mathrm{div}_0(H)\,dxdt.\end{aligned} \tag{2.16}$$

(2) *Let $P \in C^2(\overline{\Omega})$. Then*

$$\begin{aligned}&\int_0^T\int_\Omega P\left(\phi_t^2 - |\nabla_g\phi|_g^2\right)\,dxdt\\ &= (\phi_t,\, \phi P)\Big|_0^T + \frac{1}{2}\int_0^T\int_\Omega \phi^2\mathcal{A}P\,dxdt + \frac{1}{2}\int_0^T\int_\Gamma \phi^2\nabla_g P\cdot v\,d\sigma dt\\ &-\int_0^T\int_\Gamma\frac{\partial\phi}{\partial\upsilon_{\mathcal{A}}}\phi P d\sigma dt.\end{aligned} \tag{2.17}$$

Proof. (1) We multiply (2.15) by $H(\phi)$ and integrate by parts. Using Green's formula and Lemma 2.1, parts 2–4, we find that

$$\begin{aligned}\int_\Omega \mathcal{A}\phi H(\phi)\,dx &= \int_\Omega\sum_{ij=1}^n a_{ij}(x)\frac{\partial\phi}{\partial x_j}\frac{\partial}{x_i}(H(\phi))\,dx - \int_\Gamma\frac{\partial\phi}{\partial\upsilon_{\mathcal{A}}}H(\phi)\,d\sigma\\ &= \int_\Omega \nabla_g\phi\,(H(\phi))\,dx - \int_\Gamma\frac{\partial\phi}{\partial\upsilon_{\mathcal{A}}}H(\phi)\,d\sigma\end{aligned}$$

$$
\begin{aligned}
&= \int_\Omega \langle \nabla_g \phi,\, \nabla_g\,(H(\phi))\rangle_g\,dx - \int_\Gamma \frac{\partial \phi}{\partial v_{\mathcal{A}}} H(\phi)\,d\sigma \\
&= \int_0^T \int_\Omega DH\,(\nabla_g \phi,\, \nabla_g \phi)\;dxdt + \frac{1}{2}\int_0^T \int_\Gamma |\nabla_g \phi|_g^2 H\cdot v\,d\sigma dt \\
&- \frac{1}{2}\int_0^T \int_\Omega |\nabla_g \phi|_g^2 \mathrm{div}_0(H)\,dxdt - \int_\Gamma \frac{\partial \phi}{\partial v_{\mathcal{A}}} H(\phi)\,d\sigma. \qquad (2.18)
\end{aligned}
$$

On the other hand, we obtain from integration by parts and (2.1)

$$
\begin{aligned}
&\int_0^T \int_\Omega \phi_{tt} H(\phi)\,dxdt \\
&= \int_\Omega \phi_t H(\phi)\,\Big|_0^T\,dx - \int_\Omega \int_0^T \phi_t H(\phi_t)\,dtdx \\
&= (\phi_t,\, H(\phi))\,\Big|_0^T - \frac{1}{2}\int_0^T \int_\Omega H(\phi_t^2)\,dxdt \\
&= (\phi_t,\, H(\phi))\,\Big|_0^T + \frac{1}{2}\int_0^T \int_\Omega \phi_t^2 \mathrm{div}_0(H)\,dxdt - \frac{1}{2}\int_0^T \int_\Gamma \phi_t^2 H\cdot v\,d\sigma dt. \qquad (2.19)
\end{aligned}
$$

Equations (2.19) and (2.18), together with (2.15), yield (2.16).

(2) Lemma 2.1, part 2, gives

$$
\mathcal{A}P = -\sum_{ij=1}^n \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial P}{\partial x_j}\right) = -\mathrm{div}_0(\nabla_g P). \qquad (2.20)
$$

From (2.20) and formula (2.1), we thus obtain

$$
\begin{aligned}
\langle \nabla_g \phi,\, \nabla_g (P\phi)\rangle_g(x) &= P|\nabla_g \phi|_g^2(x) + \phi\langle \nabla_g \phi, \nabla_g P\rangle_g(x) \\
&= P|\nabla_g \phi|_g^2 + \frac{1}{2}\nabla_g P(\phi^2) \\
&= P|\nabla_g \phi|_g^2 + \frac{1}{2}\mathrm{div}_0(\phi^2 \nabla_g P) + \frac{1}{2}\phi^2 \mathcal{A}P \quad \forall x\in\Omega. \qquad (2.21)
\end{aligned}
$$

It follows from (2.15), (2.21), (2.2), and Green's formula that

$$
\begin{aligned}
&(\phi_t,\, \phi P)\,\Big|_0^T = \int_0^T [(\phi_{tt}, \phi P) + (\phi_t, \phi_t P)]\,dt \\
&= \int_0^T \int_\Omega \Big[-\langle \nabla_g \phi,\, \nabla_g(\phi P)\rangle_g + \phi_t^2 P\Big]\,dxdt + \int_0^T \int_\Gamma \frac{\partial \phi}{\partial v_{\mathcal{A}}}\phi P\,d\sigma dt \\
&= \int_0^T \int_\Omega P\left(\phi_t^2 - |\nabla_g \phi|_g^2\right)dxdt - \frac{1}{2}\int_0^T \int_\Gamma \phi^2 \nabla_g P\cdot v\,d\sigma dt \\
&- \frac{1}{2}\int_0^T \int_\Omega \phi^2 \mathcal{A}P\,dxdt + \int_0^T \int_\Gamma \frac{\partial \phi}{\partial v_{\mathcal{A}}}\phi P\,d\sigma dt. \qquad (2.22)
\end{aligned}
$$

Equation (2.17) follows from (2.22). □

2.2. The proofs of the results. Now we derive some inequalities that will be used later on.

Let H be a vector field on $(\mathbb{R}^n, g)$ satisfying condition (1.16) and a, which is given by (1.16). Set

$$b = \sup_{x\in\overline{\Omega}} |H|_g(x), \qquad P = \mathrm{div}_0 H - a, \tag{2.23}$$

$$c = \sup_{x\in\overline{\Omega}} |(\mathrm{div}_0 H)^2 - a^2 + 2H(P)|. \tag{2.24}$$

LEMMA 2.2. *Let ϕ solve* (1.4). *Then*

$$\left|\left(\phi_t,\, H(\phi) + \frac{1}{2}P\phi\right)\right| \le bE(0) + \frac{\sqrt{c}}{2}\|\phi_t\|\|\phi\|, \tag{2.25}$$

where

$$E(t) = \frac{1}{2}\int_\Omega (\phi_t^2 + |\nabla_g\phi|_g^2)\,dx = E(0) = \frac{1}{2}(\||\nabla_g\varphi^0|_g\|^2 + \|\varphi^1\|^2). \tag{2.26}$$

Proof. We use a technique given in Komornik [6]. Applications of divergence formulae (2.1) and (2.2) yield

$$\begin{aligned}\left\|H(\phi) + \frac{1}{2}P\phi\right\|^2 &= \|H(\phi)\|^2 + (H(\phi),\, P\phi) + \frac{1}{4}\|P\phi\|^2\\ &= \|H(\phi)\|^2 + \frac{1}{2}\int_\Omega PH(\phi^2)\,dx + \frac{1}{4}\|P\phi\|^2\\ &= \|H(\phi)\|^2 - \frac{1}{4}\int_\Omega \phi^2[(\mathrm{div}_0 H)^2 - a^2 + 2H(P)]\,dx\\ &\le \|H(\phi)\|^2 + \frac{c}{4}\|\phi\|^2, \qquad (2.27)\end{aligned}$$

since $\phi|_\Gamma = 0$. It follows from (2.27) and (2.26) that

$$\begin{aligned}\left|\left(\phi_t,\, H(\phi) + \frac{1}{2}P\phi\right)\right| &\le \|\phi_t\|\left\|H(\phi) + \frac{1}{2}P\phi\right\|\\ &\le \|\phi_t\|\left(\|H(\phi)\| + \sqrt{\frac{c}{4}}\|\phi\|\right)\\ &\le bE(0) + \frac{\sqrt{c}}{2}\|\phi_t\|\|\phi\|. \qquad \Box\end{aligned}$$

Consider the operator $\mathcal{A}$ on $L^2(\Omega)$, given by

$$\mathcal{D}(\mathcal{A}) = \{\, u \,|\, u \in H^2(\Omega),\, u|_\Gamma = 0\,\},$$

$$\mathcal{A}u = -\sum_{ij=1}^n \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right), \quad u \in \mathcal{D}(\mathcal{A}). \tag{2.28}$$

It is well known that the spectrum of positive self-adjoint $\mathcal{A}$ consists of eigenvalues

$$0 < \lambda_1 < \lambda_2 < \cdots < \lambda_k < \cdots,$$

with $\lim_{k\to\infty}\lambda_k = \infty$. Denote by Z_k the eigenspaces of $\mathcal{A}$ corresponding to λ_k for $k = 1, 2, \ldots$.

The crux of the proof of Theorem 1.1 is the following.

LEMMA 2.3. *Set*

$$\vartheta = \sup_{x\in\Gamma_0}\left[(H\cdot\upsilon)/|\upsilon_{\mathcal{A}}|_g^2\right] \quad \textit{and} \quad d = \sup_{x\in\Omega}|\mathcal{A}(\mathrm{div}_0 H)|. \tag{2.29}$$

Then

$$\int_0^T\int_{\Gamma_0}\left(\frac{\partial\phi}{\partial\upsilon_{\mathcal{A}}}\right)^2 d\sigma dt \geq \frac{2}{\vartheta}\left[aT - 2b - \sqrt{\frac{c}{\lambda_m}} - \frac{d}{2\lambda_m}T\right]E(0)$$

$$\begin{gathered}\forall T > 0, \quad \varphi^0 \in H_0^1(\Omega), \quad \varphi^1 \in L^2(\Omega),\\ \varphi^0,\, \varphi^1 \perp Z_k, \quad 1 \leq k \leq m-1, \quad m = 1, 2, \ldots,\end{gathered} \tag{2.30}$$

where ϕ solves (1.4) *for initial data* (φ^0, φ^1).

Proof. First, we deal with the values of $|\nabla_g\phi|^2$ and $H(\phi)$ on boundary Γ, respectively. Let $x \in \Gamma$. We decompose $\nabla_g\phi$ into a direct sum in $(\mathbb{R}^n_x, g(x))$:

$$\nabla_g\phi(x) = \left\langle\nabla_g\phi(x),\, \frac{\upsilon_{\mathcal{A}}(x)}{|\upsilon_{\mathcal{A}}|_g}\right\rangle_g\frac{\upsilon_{\mathcal{A}}(x)}{|\upsilon_{\mathcal{A}}|_g} + Y(x), \tag{2.31}$$

where $Y(x) \in \mathbb{R}^n_x$ with $\langle Y(x),\, \upsilon_{\mathcal{A}}(x)\rangle_g = 0$. Lemma 2.1, part 1, and (2.31) yield

$$Y(x)\cdot\upsilon(x) = \langle Y(x),\, \upsilon_{\mathcal{A}}(x)\rangle_g = 0, \tag{2.32}$$

that is, $Y(x) \in \Gamma_x$, the tangent space of Γ at x. It follows from (2.31) and (2.32) that

$$\begin{aligned}|\nabla_g\phi|_g^2 = \nabla_g\phi(\phi) &= \frac{1}{|\upsilon_{\mathcal{A}}(x)|_g^2}\langle\nabla_g\phi(x),\, \upsilon_{\mathcal{A}}(x)\rangle_g^2 + Y(\phi)\\ &= \frac{1}{|\upsilon_{\mathcal{A}}|_g^2}\left(\frac{\partial\phi}{\partial\upsilon_{\mathcal{A}}}\right)^2,\end{aligned} \tag{2.33}$$

since $\phi|_\Gamma = 0$. Similarly, H can be decomposed into a direct sum

$$H = \left\langle H(x),\, \frac{\upsilon_{\mathcal{A}}(x)}{|\upsilon_{\mathcal{A}}(x)|_g}\right\rangle_g\frac{\upsilon_{\mathcal{A}}(x)}{|\upsilon_{\mathcal{A}}(x)|_g} + Z(x), \tag{2.34}$$

where $Z(x) \in \Gamma_x$. Formula (2.34) and Lemma 2.1, part 1, give

$$H(\phi)(x) = \frac{\langle H(x),\, \upsilon_{\mathcal{A}}(x)\rangle_g}{|\upsilon_{\mathcal{A}}(x)|_g^2}\left(\frac{\partial\phi}{\partial\upsilon_{\mathcal{A}}}\right) = \frac{H(x)\cdot\upsilon(x)}{|\upsilon_{\mathcal{A}}(x)|_g^2}\left(\frac{\partial\phi}{\partial\upsilon_{\mathcal{A}}}\right). \tag{2.35}$$

By Proposition 2.1, parts 1 and 2, (1.17), (1.16), (2.33), (2.35), and Lemma 2.2, we find that

$$\begin{aligned}&\frac{\vartheta}{2}\int_0^T\int_{\Gamma_0}\left(\frac{\partial\phi}{\partial\upsilon_{\mathcal{A}}}\right)^2 d\sigma dt \geq \frac{1}{2}\int_0^T\int_\Gamma\left(\frac{\partial\phi}{\partial\upsilon_{\mathcal{A}}}\right)^2\frac{H\cdot\upsilon}{|\upsilon_{\mathcal{A}}|_g^2}\,d\sigma dt\\ &= (\phi_t,\, H(\phi))\Big|_0^T + \int_0^T\int_\Omega\langle D_{\nabla_g\phi}H,\, \nabla_g\phi\rangle_g\,dxdt + \frac{1}{2}\int_0^T\int_\Omega(\phi_t^2 - |\nabla_g\phi|_g^2)\mathrm{div}_0(H)\,dxdt\end{aligned}$$

$$
\begin{aligned}
&= \frac{a}{2}\int_0^T\int_\Omega(\phi_t^2+|\nabla_g\phi|_g^2)\,dxdt+\int_0^T\int_\Omega\langle D_{\nabla_g\phi}H,\,\nabla_g\phi\rangle\,dxdt\\
&-a\int_0^T\int_\Omega|\nabla_g\phi|_g^2\,dxdt+(\phi_t,\,H(\phi))\Big|_0^T+\frac{1}{2}\int_0^T\int_\Omega(\phi_t^2-|\nabla_g\phi|_g^2)P\,dxdt\\
&\geq aTE(0)+\left(\phi_t,\,H(\phi)+\frac{1}{2}P\phi\right)\Big|_0^T+\frac{1}{4}\int_0^T\int_\Omega\phi^2\mathcal{A}(\mathrm{div}_0H)\,dxdt\\
&\geq aTE(0)-2bE(0)-\sqrt{c}\|\phi_t\|\|\phi\|-\frac{d}{4}\int_0^T\|\phi\|^2\,dt\\
&\geq aTE(0)-2bE(0)-\sqrt{\frac{c}{\lambda_m}}\|\phi_t\|\||\nabla_g\phi|_g\|-\frac{d}{4\lambda_m}\int_0^T\||\nabla_g\phi|_g\|^2\,dt\\
&\geq\left[aT-2b-\sqrt{\frac{c}{\lambda_m}}-\frac{d}{2\lambda_m}T\right]E(0),
\end{aligned}
$$

where inequalities

$$\|\phi\|^2\leq\frac{1}{\lambda_m}\||\nabla_g\phi|_g\|^2$$

$$\forall\,\varphi_0\in H_0^1(\Omega),\quad\varphi_1\in L^2(\Omega),\quad\varphi^0,\varphi^1\ \perp Z_k,\quad 1\leq k\leq m-1$$

are used. This completes the proof. □

Consider the operator $\mathcal{A}_0$ on $L^2(\Omega)$, defined by

$$\mathcal{D}(\mathcal{A}_0)=\left\{u\ \Big|\ u\in H^2(\Omega),\ \frac{\partial u}{\partial v_{\mathcal{A}}}\Big|_\Gamma=0\right\},$$

$$\mathcal{A}_0u=-\sum_{ij=1}^n\frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right),\quad u\in\mathcal{D}(\mathcal{A}_0).\tag{2.36}$$

It is well known that $\mathcal{A}_0$ is nonnegative, self-adjoint on $L^2(\Omega)$ and that its spectrum consists of eigenvalues

$$0=\gamma_1<\gamma_2<\cdots<\gamma_k<\cdots$$

with $\lim_{k\to\infty}\gamma_k=\infty$. Denote by Z_k^0 the eigenspaces of $\mathcal{A}_0$ corresponding to γ_k for $k=1,2,\ldots$.

LEMMA 2.4. *Let $m\geq 2$ be a positive integer, with P and d as given in* (2.23) *and* (2.29), *respectively. Set*

$$e=\sup_{x\in\Omega}|P|.\tag{2.37}$$

Suppose that ϕ solves problem

$$\begin{cases}\phi_{tt}+\mathcal{A}\phi=0\quad in\quad\Omega\times(0,T),\\ \phi(0)=\varphi^0,\quad\phi_t(0)=\varphi^1\quad in\quad\Omega,\\ \dfrac{\partial\phi}{\partial v_{\mathcal{A}}}=0\quad on\quad\Gamma\times(0,T).\end{cases}\tag{2.38}$$

Then

$$\int_0^T\int_{\Gamma_0}\phi_t^2 H\cdot \upsilon\, d\sigma dt-\int_0^T\int_{\Gamma_1}|\nabla_g\phi|_g^2 H\cdot\upsilon\, d\sigma dt-\frac{1}{2}\int_0^T\int_{\Gamma}\phi^2\nabla_g(\mathrm{div}_0(H))\cdot\upsilon\, d\sigma dt$$

$$\geq 2\left[aT-2b-\frac{e}{\sqrt{\gamma_m}}-\frac{Td}{2\gamma_m}\right]E(0)\quad \forall T>0, \tag{2.39}$$

for all $(\varphi^0,\,\varphi^1)\in H^1(\Omega)\times L^2(\Omega)$ *with* $\varphi^0,\,\varphi^1\ \perp Z_k^0$, $1\leq k\leq m-1$, *for which the left-hand side of* (2.39) *is finite, where this time*

$$E(0)=\frac{1}{2}(\|\varphi^0\|^2_{H^1(\Omega)}+\|\varphi^1\|^2). \tag{2.40}$$

Proof. Since $\frac{\partial\phi}{\partial\upsilon_{\mathcal{A}}}|_\Gamma=0$ and $H(x)\cdot\upsilon(x)\leq 0$ for $x\in\Gamma_1$, we have, from parts 1 and 2 of Proposition 2.1, that

$$\begin{aligned}&\int_0^T\int_{\Gamma_0}\phi_t^2 H\cdot\upsilon\, d\sigma dt-\int_0^T\int_{\Gamma_1}|\nabla_g\phi|_g^2 H\cdot\upsilon\, d\sigma dt-\frac{1}{2}\int_0^T\int_\Gamma\phi^2\nabla_g P\cdot\upsilon\, d\sigma dt\\&=2aTE(0)-\int_0^T\int_{\Gamma_1}\phi_t^2 H\cdot\upsilon\, d\sigma dt+\int_0^T\int_{\Gamma_0}|\nabla_g\phi|_g^2 H\cdot\upsilon\, d\sigma dt+2(\phi_t,\,H(\phi))\Big|_0^T\\&+2\int_0^T\int_\Omega\langle D_{\nabla_g\phi}H,\,\nabla_g\phi\rangle_g\, dxdt-2a\int_0^T\int_\Omega|\nabla_g\phi|_g^2\, dxdt+(\phi_t,\,\phi P)\Big|_0^T\\&+\frac{1}{2}\int_0^T\int_\Omega\phi^2\mathcal{A}P\, dxdt\\&\geq 2aTE(0)-4\|\phi_t\|\|H(\phi)\|-2\|\phi_t\|\|\phi P\|-\frac{d}{2}\int_0^T\int_\Omega\phi^2\, dxdt\\&\geq 2\left[aT-2b-\frac{e}{\sqrt{\gamma_m}}-\frac{Td}{2\gamma_m}\right]E(0)\quad\forall T>0,\end{aligned}$$

for all $(\varphi^0,\,\varphi^1)\in H^1(\Omega)\times L^2(\Omega)$ with $\varphi^0,\,\varphi^1\ \perp Z_k^0$, $1\leq k\leq m-1$, for which the left-hand side of (2.39) is finite, where inequalities

$$\|\phi\|^2\leq\frac{1}{\gamma_m}\||\nabla_g\phi|_g\|^2\quad\forall\,(\varphi^0,\varphi^1)\in H^1(\Omega)\times L^2(\Omega),$$

$$\varphi^0,\,\varphi^1\ \perp Z_k^0,\ 1\leq k\leq m-1$$

are used. □

Consider the operator $\mathcal{A}_{\Gamma_1}$ on $L^2(\Omega)$:

$$\mathcal{D}(A_{\Gamma_1})=\left\{u\ \Big|\ u\in H^2(\Omega),\,\frac{\partial u}{\partial\upsilon_{\mathcal{A}}}\Big|_{\Gamma_0}=0,\,u|_{\Gamma_1}=0\right\},$$

$$A_{\Gamma_1}u=-\sum_{ij=1}^n\frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right),\quad u\in\mathcal{D}(A_{\Gamma_1}). \tag{2.41}$$

The spectrum of $\mathcal{A}_{\Gamma_1}$ consists of eigenvalues

$$0\leq\beta_1<\beta_2<\cdots<\beta_k<\cdots$$

with $\lim_{k\to\infty}\beta_k=\infty$. Denote by $Z_k^{\Gamma_1}$ the eigenspaces of $\mathcal{A}_{\Gamma_1}$ corresponding to β_k for $k=1,2,\ldots$.

Lemma 2.5. *Let m be a positive integer and e, d the same as those in Lemma 2.4. Let ϕ solve problem*

$$\begin{cases} \phi_{tt}+\mathcal{A}\phi=0 \quad \textit{in} \quad \Omega\times(0,T), \\ \phi(0)=\varphi^0, \quad \phi_t(0)=\varphi^1 \quad \textit{in} \quad \Omega, \\ \dfrac{\partial\phi}{\partial\upsilon_{\mathcal{A}}}=0 \quad \textit{on} \quad \Gamma_0, \quad \phi=0 \quad \textit{on} \quad \Gamma_1. \end{cases} \tag{2.42}$$

Then

$$\begin{aligned} &\int_0^T\int_{\Gamma_0}\phi_t^2 H\cdot\upsilon\, d\sigma dt-\frac{1}{2}\int_0^T\int_{\Gamma_0}\phi^2\nabla_g(\mathrm{div}_0(H))\cdot\upsilon\, d\sigma dt \\ &\qquad\geq 2\left[aT-2b-\frac{e}{\sqrt{\beta_m}}-\frac{Td}{2\beta_m}\right]E(0) \quad \forall\, T>0, \end{aligned} \tag{2.43}$$

for all $(\varphi^0,\varphi^1)\in H^1_{\Gamma_1}(\Omega)\times L^2(\Omega)$ with φ^0, $\varphi^1 \perp Z_k^{\Gamma_1}$, $1\leq k\leq m-1$, for which the left-hand side of (2.43) *is finite, where*

$$E(0)=\frac{1}{2}(\|\varphi^0\|^2_{H^1_{\Gamma_1}(\Omega)}+\|\varphi^1\|^2). \tag{2.44}$$

Proof. Since $\frac{\partial\phi}{\partial\upsilon_{\mathcal{A}}}|_{\Gamma_0}=0$ and $\phi|_{\Gamma_1}=0$, we obtain, from (2.33), (2.35), and Proposition 2.1, parts 1 and 2, that

$$\begin{aligned} &\int_0^T\int_{\Gamma_0}\phi_t^2 H\cdot\upsilon\, d\sigma dt-\frac{1}{2}\int_0^T\int_{\Gamma_0}\phi^2\nabla_g(\mathrm{div}_0(H))\cdot\upsilon\, d\sigma dt=\int_0^T\int_{\Gamma_0}|\nabla_g\phi|_g^2 H\cdot\upsilon\, d\sigma dt \\ &-\int_0^T\int_{\Gamma_1}\left(\frac{\partial\phi}{\upsilon_{\mathcal{A}}}\right)^2\frac{H\cdot\upsilon}{|\upsilon_{\mathcal{A}}|_g^2}\, d\sigma dt+2aTE(0)+2\int_0^T\int_\Omega\langle D_{\nabla_g\phi}H,\ \nabla_g\phi\rangle_g\, d\sigma dt \\ &-2a\int_0^T\int_\Omega|\nabla_g\phi|_g^2\, d\sigma dt+2(\phi_t,\ H(\phi))\Big|_0^T+(\phi_t,\ \phi P)\Big|_0^T+\frac{1}{2}\int_0^T\int_\Omega\phi^2\mathcal{A}P\, dxdt \\ &\geq 2aTE(0)-4\|\phi_t\|\|H(\phi)\|-2\|\phi_t\|\|\phi P\|-\frac{d}{2}\int_0^T\int_\Omega\phi^2\, dxdt \\ &\geq 2\left[aT-2b-\frac{e}{\sqrt{\beta_m}}-\frac{Td}{2\beta_m}\right]E(0) \quad \forall\, T>0, \end{aligned} \tag{2.45}$$

for all $(\varphi^0,\ \varphi^1)\in H^1_{\Gamma_1}(\Omega)\times L^2(\Omega)$ with φ^0, $\varphi^1 \perp Z_k^{\Gamma_1}$, $1\leq k\leq m-1$, for which the left-hand side of (2.45) is finite, where this time inequalities

$$\|\phi\|^2\leq\frac{1}{\beta_m}\||\nabla_g\phi|_g\|^2 \quad \forall\,(\varphi^0,\ \varphi^1)\in H^1_{\Gamma_1}(\Omega)\times L^2(\Omega),$$

$$\varphi^0,\ \varphi^1 \perp Z_k^{\Gamma_1}, \qquad 1\leq k\leq m-1$$

are used. □

Proof of Theorem 1.1. Let T_0 be as defined in (1.18). Let $T>T_0$. Let m be given large enough such that

$$T-T_0-\frac{1}{a}\sqrt{\frac{c}{\lambda_m}}-\frac{Td}{2a\lambda_m}>0. \tag{2.46}$$

Lemma 2.3 yields the following inequality:

$$\int_0^T \int_{\Gamma_0} \left(\frac{\partial \phi}{\partial v_{\mathcal{A}}}\right)^2 d\sigma dt \geq c(T) E(0)$$

$$\forall\, \varphi^0 \in H_0^1(\Omega), \quad \varphi^1 \in L^2(\Omega), \quad \varphi^0,\, \varphi^1 \perp Z_k, \quad 1 \leq k \leq m-1, \tag{2.47}$$

where ϕ solves (1.4) for initial data (φ^0, φ^1) and

$$c(T) = \frac{2}{\vartheta} a \left(T - T_0 - \frac{1}{a}\sqrt{\frac{c}{\lambda_m}} - \frac{Td}{2a\lambda_m}\right) > 0.$$

Inequalities (1.5) and (2.47), together with Komornik [6, Thm. 5.2], give Theorem 1.1. $\square$

Proof of Corollary 1.2. We show that vector field $H = \rho \nabla_g \rho$ verifies condition (1.16) under the assumption of Corollary 1.2.

Let M be a space form of constant curvature k_B with Riemannian metric $\langle \cdot, \cdot \rangle_M$ and let $\widetilde{D}$ be the Levi–Civita connection on M. Let $p \in M$. Let $\widetilde{\rho}(q)$ be the distance function on M from p to $q \in M$. For convenience, we introduce a calculation of $\widetilde{D}^2 \widetilde{\rho}$ which is presented in Wu, Shen, and Yu [19]. Let $\widetilde{r} : [0, b] \to M$ be a normal geodesic from p to q such that there is no point conjugate to p on $\widetilde{r}$. Then $\widetilde{\rho}(q) = b$. Let $\widetilde{X} \in M_q$ such that $\langle X, \widetilde{r}'(b)\rangle_M = 0$ and $|\widetilde{X}|_M = 1$. Let $J(t)$ be a proper Jacobi field along $\widetilde{r}$ such that $J(0) = 0$ and $J(b) = \widetilde{X}$. Since M is of constant curvature, $J(t)$ is an almost parallel field along $\widetilde{r}$. Let $E(t)$ be a unit parallel field such that $E(b) = \widetilde{X}$. Then $J(t) = f(t)E(t)$, where f solves problem

$$\begin{cases} f''(t) + k_B f(t) = 0, \quad 0 < t < b, \\ f(0) = 0, \quad f(b) = 1. \end{cases} \tag{2.48}$$

Expression (2.48) gives

$$f(t) = \begin{cases} \dfrac{t}{b}, \quad k_B = 0, \\ \dfrac{\sin\sqrt{k_B}t}{\sin\sqrt{k_B}b}, \quad k_B > 0, \\ \dfrac{\sinh\sqrt{-k_B}t}{\sinh\sqrt{-k_B}b}, \quad k_B < 0. \end{cases} \tag{2.49}$$

Denote the unit sphere in M_p by S. Suppose that $\sigma : [0, \epsilon] \to S$ is a curve such that $\sigma(0) = \widetilde{r}'(0)$ and such that the transversal vector field of $\{\widetilde{r}_s\}$ is J, where $\widetilde{r}_s(t) = \exp_p t\sigma(s)$. Let $F : [0, b] \times [0, \epsilon] \to M$ be defined by $F(t, s) = \widetilde{r}_s(t)$. Set $T = dF(\frac{\partial}{\partial t})$ and $U = dF(\frac{\partial}{\partial s})$. Then $U(\widetilde{r}(t)) = J(t)$ and $T(\widetilde{r}(t)) = \widetilde{r}'(t) = \frac{\partial}{\partial \widetilde{\rho}}(\widetilde{r}(t))$. From (2.49), we then obtain

$$\begin{aligned} \widetilde{D}^2 \widetilde{\rho}(q)(\widetilde{X}, \widetilde{X}) &= UU\widetilde{\rho} - (\widetilde{D}_U)U\widetilde{\rho} \\ &= U\left\langle U, \frac{\partial}{\partial \widetilde{\rho}}\right\rangle_M - \left\langle \widetilde{D}_U U, \frac{\partial}{\partial \widetilde{\rho}}\right\rangle_M \\ &= U\langle U, T\rangle_M - \langle \widetilde{D}_U U, T\rangle_M \\ &= \langle U, \widetilde{D}_U T\rangle_M = \langle U, \widetilde{D}_T U\rangle_M = \langle J(b), J'(b)\rangle_M = f'(b) \\ &= \begin{cases} \dfrac{1}{b}, \quad k_B = 0, \\ \sqrt{k_B}\cot(\sqrt{k_B}b), \quad k_B > 0, \\ \sqrt{-k_B}\coth(\sqrt{-k_B}b), \quad k_B < 0, \end{cases} \end{aligned} \tag{2.50}$$

where $UT = TU$ is used.

Let $x \in \Omega$. Let $r : [0,\, b] \to B(x^0, \gamma)$ be a normal minimal geodesic from x^0 to x. Then $b = \rho(x)$. Let $X \in \mathbb{R}^n_x$ such that $\langle X,\, r'(b)\rangle_g = 0$. Since $\Omega \subset B(x^0, \gamma)$ and $B(x^0, \gamma) \subset \exp_x(E)$, $\rho(x) = d_g(x^0, x)$ is smooth on $\Omega/\{x^0\}$. We obtain from the Hessian comparison theorem (Greene and Wu [5]) and (2.50) that

$$D^2\rho(X,\, X)(x) = |X|_g^2 D^2\rho\left(\frac{X}{|X|_g},\, \frac{X}{|X|_g}\right) \geq |X|_g^2 \widetilde{D}^2\widetilde{\rho}(\widetilde{X},\, \widetilde{X})$$

$$= \begin{cases} \dfrac{1}{\rho(x)}|X|_g^2, \quad k_B = 0, \\ \sqrt{k_B}\cot(\sqrt{k_B}\rho(x))|X|_g^2, \quad k_B > 0, \\ \sqrt{-k_B}\coth(\sqrt{-k_B}\rho(x))|X|_g^2, \quad k_B < 0, \end{cases}$$

$$\geq \begin{cases} \dfrac{1}{\gamma}|X|_g^2, \quad k_B = 0, \\ \sqrt{k_B}\cot(\sqrt{k_B}\gamma)|X|_g^2, \quad k_B > 0 \quad \forall x \in \Omega, \\ \dfrac{1}{\gamma}|X|_g^2, \quad k_B < 0, \end{cases} \tag{2.51}$$

where inequality

$$\sqrt{-k_B}\rho\left(e^{\sqrt{-k_B}\rho} + e^{-\sqrt{-k_B}\rho}\right) \geq e^{\sqrt{-k_B}\rho} - e^{-\sqrt{-k_B}\rho}, \quad k_B < 0,\, \rho > 0,$$

is used.

Set

$$a = \begin{cases} 1, \quad K_B \leq 0, \\ \sqrt{k_B}\gamma\cot(\sqrt{k_B}\gamma), \qquad K_B > 0. \end{cases} \tag{2.52}$$

It is easily checked that the function $t\sqrt{k_B}\cot(\sqrt{k_B}t)$ is monotonically decreasing in $t \in (0, \infty)$ and

$$t\sqrt{k_B}\cot(\sqrt{k_B}t) \leq 1 \qquad \forall\, t \in (0, \infty). \tag{2.53}$$

We thus obtain from (2.51), (2.52), and (2.53), for any $Y \in \mathbb{R}^n_x$, that

$$\begin{aligned} DH(Y,\, Y)(x) =& \rho D^2\rho(Y,\, Y) + \langle Y,\, r'(b)\rangle_g^2 \\ &= \rho D^2\rho(X,\, X) + \langle Y,\, r'(b)\rangle_g^2 \\ &\geq a\left(|X|_g^2 + \langle Y,\, r'(b)\rangle_g^2\right) = a|Y|_g^2 \quad \forall\, x \in \Omega, \end{aligned} \tag{2.54}$$

where $H = \rho\nabla_g\rho$, $Y = X + \langle Y,\, r'(b)\rangle_g r'(b)$, and $\langle X,\, r'(b)\rangle_g = 0$. □

Proof of Corollary 1.4. Set

$$\gamma_0 = \min\left\{\frac{\pi}{\sqrt{K}},\right.$$

$$\left.\text{half the length of the shortest closed geodesic in } \mathbb{R}^n\right\}.$$

Let $x^0 \in \Omega$. Then $\Omega \subset B(x^0, \frac{\gamma_0}{2})$ since $d_g(\Omega) < \frac{\gamma_0}{2}$. By the Klingenberg theorem (Cheeger and Ebin [2, Cor. 5.7]), we have

$$B\left(x^0, \frac{\gamma_0}{2}\right) \subset \exp_{x^0}(E).$$

Since

$$4\left(\frac{\gamma_0}{2}\right)^2 K = \gamma_0^2 K < \pi^2,$$

Corollary 1.2 completes our proof. □

Proof of Theorem 1.2. It is well known that Γ is a submanifold of Riemannian manifold $(\mathbb{R}^n, g)$ with the induced Riemannian metric g_Γ in $(\mathbb{R}^n, g)$. Denote gradient, divergence, and the Laplace operator on (Γ, g_Γ) by ∇_Γ, div_Γ, and Δ_Γ, respectively. Set $\mathcal{G}(x) = \det G(x)$ for $x \in \mathbb{R}^n$. Let Δ_g be the Laplace operator on $\mathbb{R}^n$ in Riemannian metric g. Then, in the natural system $x = (x_1, x_2, \dots, x_n)$,

$$(2.55)\quad \Delta_g f = \frac{1}{\sqrt{\mathcal{G}(x)}} \sum_{ij=1}^{n} \frac{\partial}{\partial x_i}\left(\sqrt{\mathcal{G}(x)} a_{ij}(x) \frac{\partial f}{\partial x_j}\right) \quad \forall f \in C^2(\mathbb{R}^n), \quad x \in \mathbb{R}^n.$$

By the formula of divergence in Riemannian metric g and from part 1 of Lemma 2.1, we have

$$(2.56)\quad \int_\Omega \Delta_g f\, d\Omega_g = \int_\Gamma \left\langle \nabla_g f, \frac{\upsilon_{\mathcal{A}}}{|\upsilon_{\mathcal{A}}|_g}\right\rangle_g d\sigma_\Gamma = \int_\Gamma \frac{1}{|\upsilon_{\mathcal{A}}|_g} \nabla_g f \cdot \upsilon d\sigma_\Gamma \quad \forall f \in C^2(\overline{\Omega}),$$

where Ω_g and $d\sigma_\Gamma$ are the metric volume element and the metric surface element in Riemannian metric g, respectively. It is easily checked that $d\Omega_g = \sqrt{\mathcal{G}(x)} dx$. We thus obtain from (2.55) and part 2 of Lemma 2.1 that

$$(2.57)\quad \int_\Omega \Delta_g f\, d\Omega_g = \int_\Omega \mathrm{div}_0(\sqrt{\mathcal{G}(x)} \nabla_g f)\, dx = \int_\Gamma \sqrt{\mathcal{G}(x)} \nabla_g f \cdot \upsilon\, d\sigma \quad \forall f \in C^2(\overline{\Omega}).$$

Formulae (2.57) and (2.56) give

$$(2.58)\qquad d\sigma_\Gamma = |\upsilon_{\mathcal{A}}|_g \sqrt{\mathcal{G}(x)} d\sigma.$$

In addition, by the Stokes theorem for Riemannian manifold (Γ, g_Γ), we have

$$(2.59)\qquad -\int_\Gamma f \Delta_\Gamma f\, d\sigma_\Gamma = \int_\Gamma |\nabla_\Gamma f|_g^2\, d\sigma_\Gamma$$

for any $f \in H^1(\Gamma)$.

In the following, let ϕ solve problem (2.38) for initial data (φ^0, φ^1). Then

$$(2.60)\qquad |\nabla_g \phi|_g^2 = \left(\frac{\partial \phi}{\partial \upsilon_{\mathcal{A}}}\right)^2 + |\nabla_\Gamma \phi|_g^2 = |\nabla_\Gamma \phi|_g^2 \quad \forall x \in \Gamma,$$

since $\frac{\partial \phi}{\upsilon_{\mathcal{A}}} = 0$ on Γ.

We define ψ by

$$(2.61)\qquad \begin{cases} \psi_{tt} + \mathcal{A}\psi = 0 \quad \text{in} \quad \Omega \times (0,T), \\ \psi(T) = \psi_t(T) = 0, \\ \dfrac{\partial \psi}{\partial \upsilon_{\mathcal{A}}} = \begin{cases} (\phi_{tt} - \phi)|\upsilon_{\mathcal{A}}|_g \sqrt{\mathcal{G}(x)} & \text{on} \quad \Gamma_0 \times (0,T), \\ (\Delta_\Gamma \phi - \phi)|\upsilon_{\mathcal{A}}|_g \sqrt{\mathcal{G}(x)} & \text{on} \quad \Gamma_1 \times (0,T). \end{cases} \end{cases}$$

The solution of (2.61) is a weak solution, defined by transposition. Therefore, given φ^0, φ^1, we have defined, in a unique fashion,

$$(2.62)\qquad \Lambda_1(\varphi^0, \varphi^1) = (\psi_t(0), -\psi(0)).$$

Then, from (2.38), (2.61), (2.60), and (2.58), we obtain, in the sense of distribution,

$$\langle \Lambda_1(\varphi^0,\, \varphi^1),\, (\varphi^0,\, \varphi^1)\rangle = \int_0^T \int_{\Gamma_0} (\phi^2 + \phi_t^2)\, d\sigma_\Gamma dt + \int_0^T \int_{\Gamma_1} (\phi^2 + |\nabla_\Gamma \phi|_g^2)\, d\sigma_\Gamma dt. \tag{2.63}$$

We define on initial data (φ_0, φ^1)

$$\|(\varphi_0, \varphi^1)\|_F = \left(\int_0^T \int_{\Gamma_0} (\phi^2 + \phi_t^2)\, d\sigma_\Gamma dt + \int_0^T \int_{\Gamma_1} (\phi^2 + |\nabla_\Gamma \phi|_g^2)\, d\sigma_\Gamma dt \right)^{1/2}. \tag{2.64}$$

Set

$$\mathcal{X}(T) = \left\{ (\varphi^0, \varphi^1) \;\middle|\; (\varphi^0, \varphi^1) \in H^1(\Omega) \times L^2(\Omega),\, \|(\varphi^0, \varphi^1)\|_F < \infty \right\}. \tag{2.65}$$

Then $\mathcal{X}(T) \subset F$. It is easily checked that $\|\cdot\|_F$ is a seminorm on $\mathcal{X}(T)$.

Let T_0 be given in (1.18). Let $T > T_0$. Suppose that m is given large enough such that

$$T - T_0 - \frac{e}{a\sqrt{\gamma_m}} - \frac{Td}{2a\gamma_m} > 0. \tag{2.66}$$

Set

$$\varepsilon = \max \left(\sup_{x \in \Gamma} \frac{|H \cdot v|}{|v_{\mathcal{A}}|_g \sqrt{\mathcal{G}}},\, \sup_{x \in \Gamma} \frac{|\nabla_g (\mathrm{div}_0 H) \cdot v|}{2|v_{\mathcal{A}}|_g \sqrt{\mathcal{G}}} \right), \tag{2.67}$$

$$c_m = \frac{2a}{\varepsilon} \left(T - T_0 - \frac{e}{a\sqrt{\gamma_m}} - \frac{Td}{2a\gamma_m} \right). \tag{2.68}$$

Lemma 2.4, (2.58), (2.67), and (2.68) then yield the following:

$$\int_0^T \int_{\Gamma_0} (\phi^2 + \phi_t^2)\, d\sigma_\Gamma dt + \int_0^T \int_{\Gamma_1} (\phi^2 + |\nabla_\Gamma \phi|_g^2)\, d\sigma_\Gamma dt \geq c_m E(0)$$

$$\forall\, (\varphi^0, \varphi^1) \in \mathcal{X}(T), \quad \varphi^0,\, \varphi^1 \;\perp\; Z_k^0, \quad 1 \leq k \leq m - 1. \tag{2.69}$$

By the compactness-uniqueness argument, we now show that there is a $c(T) > 0$ such that

$$\|(\varphi_0, \varphi^1)\|_F^2 \geq c(T) E(0) \quad \forall\, (\varphi^0, \varphi^1) \in \mathcal{X}(T). \tag{2.70}$$

Suppose that inequality (2.70) does not hold true. There is a sequence $\{(\varphi_k^0, \varphi_k^1)\} \subset H^1(\Omega) \times L^2(\Omega)$ with

$$E_k(0) = \frac{1}{2}(\|\varphi_k^0\|_{H^1(\Omega)}^2 + \|\varphi_k^1\|^2) = 1 \tag{2.71}$$

for all $k \geq 1$ such that

$$\|(\varphi_k^0, \varphi_k^1)\|_F \leq \frac{1}{k}, \quad k = 1, 2, \ldots. \tag{2.72}$$

Let $H^1(\Omega)\times L^2(\Omega)$ be decomposed into a direct sum:

$$H^1(\Omega)\times L^2(\Omega)=\big((\oplus_{k=1}^{m-1}Z_k^0)\times(\oplus_{k=1}^{m-1}Z_k^0)\big)\oplus\widetilde{W}. \tag{2.73}$$

It is easily checked that

$$\dim\big((\oplus_{k=1}^{m-1}Z_k^0)\times(\oplus_{k=1}^{m-1}Z_k^0)\big)<\infty. \tag{2.74}$$

Set

$$(\varphi_k^0,\varphi_k^1)=\widetilde{u}_k+\widetilde{w}_k,\quad k=1,2,\dots, \tag{2.75}$$

where $\widetilde{u}_k\in(\oplus_{k=1}^{m-1}Z_k^0)\times(\oplus_{k=1}^{m-1}Z_k^0)$ and $\widetilde{w}_k\in\widetilde{W}$. From (2.74), we can assume that

$$\widetilde{u}_k\to\widetilde{u}_0\in\mathcal{X}(T)\quad\text{in}\quad\|\cdot\|_F,\quad\text{as}\quad k\to\infty. \tag{2.76}$$

Since $\|\cdot\|_F$ is a seminorm on $\mathcal{X}(T)$, we obtain from (2.40), (2.69), (2.72), and (2.75) that

$$\sqrt{c_m}\|\widetilde{w}_k-\widetilde{w}_j\|_{H^1(\Omega)\times L^2(\Omega)}\le\|\widetilde{w}_k-\widetilde{w}_j\|_F\le\left(\frac{1}{k}+\frac{1}{j}\right)+\|\widetilde{u}_k-\widetilde{u}_j\|_F\quad\forall\, k\ge1,\ j\ge1.$$

This means that by (2.76),

$$(\varphi_k^0,\varphi_k^1)\to(u^0,u^1)\quad\text{in}\quad H^1(\Omega)\times L^2(\Omega)\quad\text{as}\quad k\to\infty. \tag{2.77}$$

It follows from (2.71) and (2.77) that

$$\|(u^0,u^1)\|_{H^1(\Omega)\times L^2(\Omega)}=1. \tag{2.78}$$

Let ϕ^0 solve problem (2.38) for $(\varphi^0,\varphi^1)=(u^0,u^1)$. Then $\frac{\partial\phi^0}{\partial v_{\mathcal{A}}}|_\Gamma=0$. In addition, by the trace theorem of Sobolev spaces, (2.64), and (2.72), we know that there is a constant $\beta>0$ such that

$$\begin{aligned}\int_0^T\|\phi^0\|_{L^2(\Gamma)}^2\,dt&\le\int_0^T\|\phi^0-\phi^k\|_{L^2(\Gamma)}^2\,dt+\int_0^T\|\phi^k\|_{L^2(\Gamma)}^2\,dt\\&\le\beta\int_0^T\|\phi^0-\phi^k\|_{H^1(\Omega)}^2dt+\|(\varphi_k^0,\varphi_k^1)\|_F^2\\&\le\beta T\|(u^0,u^1)-(\varphi_k^0,\varphi_k^1)\|_{H^1(\Omega)\times L^2(\Omega)}^2+\frac{1}{k^2},\end{aligned} \tag{2.79}$$

where ϕ^k solve problems (2.38) with initial data $(\varphi^0,\varphi^1)=(\varphi_k^0,\varphi_k^1)$ for $k=1,2,\dots$. Expressions (2.77) and (2.79) give

$$\phi^0|_\Gamma=0. \tag{2.80}$$

From (2.80), we may apply observability inequality (1.9) to ϕ^0 and obtain

$$(u^0,u^1)=0,$$

contradicting (2.78).

The inequality in (2.70) gives

$$F\subset H^1(\Omega)\times L^2(\Omega),\quad T>T_0.$$

It follows that

$$(H^1(\Omega))' \times L^2(\Omega) \subset F',$$

so that, if $y^0 \in L^2(\Omega)$, $y^1 \in (H^1(\Omega))'$, we can derive the system in (1.1) and (1.1b) to rest at T by the control function v given by

$$v = \begin{cases} (-\phi + \phi_{tt})|\upsilon_{\mathcal{A}}|_g \sqrt{\mathcal{G}(x)} & \text{on} \quad \Gamma_0 \times (0,T), \\ (-\phi + \Delta_\Gamma \phi)|\upsilon_{\mathcal{A}}|_g \sqrt{\mathcal{G}(x)} & \text{on} \quad \Gamma_1 \times (0,T). \end{cases}$$

This completes the proof of Theorem 1.2. □

Proof of Theorem 1.3. Let T_0 be given in (1.18) and $T > T_0$. Since $\Gamma_1 = \{\, x \,|\, x \in \Gamma,\ H(x) \cdot \upsilon(x) = 0 \,\}$, by starting with Lemma 2.4 and following the proof of Theorem 1.2, we can get the following inequality:

$$\int_0^T \int_\Gamma (\phi^2 + \phi_t^2)\, d\sigma dt \geq c_T E(0), \tag{2.81}$$

where $c_T > 0$ is a constant. This time we define a norm on initial data (φ^0, φ^1) by

$$\|(\varphi^0, \varphi^1)\|_F = \left(\int_0^T \int_\Gamma \phi^2 \, d\sigma dt \right)^{1/2}.$$

From (2.81), a technical argument similar to Lions [12] gives

$$\int_0^T \int_\Gamma \phi^2 \, d\sigma dt \geq c'_T E(0).$$

This completes our proof. □

Proof of Theorem 1.4. Let T_0 be given in (1.18) and $T > T_0$. Let ϕ solve problem (2.42). From Lemma 2.5 and by following the proof of Theorem 1.3, we can obtain $c_T > 0$, satisfying

$$\int_0^T \int_{\Gamma_0} (\phi^2 + \phi_t^2) \geq c_T E(0) \tag{2.82}$$

for all $(\varphi^0,\, \varphi^1) \in H^1_{\Gamma_1}(\Omega) \times L^2(\Omega)$, for which the left-hand side of (2.82) is finite.

On the initial data (φ^0, φ^1), we define the norm by

$$\|(\varphi^0, \varphi^1)\|_F = \left(\int_0^T \int_{\Gamma_0} (\phi^2 + \phi_t^2)\, d\sigma dt \right)^{1/2}.$$

From (2.82), we have

$$F \subset H^1_{\Gamma_1}(\Omega) \times L^2(\Omega).$$

Then

$$F' \supset (H^1_{\Gamma_1}(\Omega))' \times L^2(\Omega).$$

This completes our proof, where we take

$$v = -\phi + \phi_{tt} \quad \text{in} \quad \Gamma_0. \qquad \square$$

3. Examples.

Example 3.1. Let (a_{ij}) be a positive, symmetric constant matrix. Consider the operator $\mathcal{A}$ on $\mathbb{R}^n$, given by

$$\mathcal{A}u = -\sum_{ij=1}^{n} a_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j}.$$

It is easily verified that Riemannian manifold $(\mathbb{R}^n, g)$ is of zero sectional curvature, where

$$g = \sum_{ij}^{n} g_{ij}dx_i dx_j \quad \text{and} \quad (g_{ij}) = (a_{ij})^{-1}. \tag{3.1}$$

By Corollary 1.2,

$$a = 1. \tag{3.2}$$

Let $x^0 \in \mathbb{R}^n$. The distance function ρ satisfies

$$\rho^2(x) = d_g^2(x^0, x) = \sum_{ij=1}^{n} g_{ij}(x_i - x_i^0)(x_j - x_j^0) \quad \forall\ x \in \mathbb{R}^n. \tag{3.3}$$

We obtain

$$\frac{\partial h}{\partial x_k} = \sum_{j=1}^{n} g_{kj}(x_j - x_j^0), \tag{3.4}$$

where $h = \frac{1}{2}\rho^2(x)$. It follows from (3.4) and part 2 of Lemma 2.1 that

$$\nabla_g h = \sum_{i=1}^{n}\left(\sum_{k=1}^{n} a_{ik}\frac{\partial h}{\partial x_k}\right)\frac{\partial}{\partial x_i} = \sum_{i=1}^{n}(x_i - x_i^0)\frac{\partial}{\partial x_i}. \tag{3.5}$$

This gives

$$\sup_{x\in\Omega} |\nabla_g h|_g(x) = \sup_{x\in\Omega} \rho(x), \tag{3.6}$$

so that vector field $H = \nabla_g h$ meets condition (1.16) with $T_0 = 2\sup_{x\in\Omega}\rho(x)$ for any $\Omega \subset \mathbb{R}^n$. This coincides with the result in Komornik [6], where $a_{ij} = \delta_{ij}$. □

LEMMA 3.1. *Let $a_1, a_2, \ldots, a_n$ be positive numbers and M be the hypersurface of $\mathbb{R}^{n+1}$ given by*

$$M = \left\{ (x_1, x_2, \ldots, x_n, x_{n+1}) \;\Big|\; x_{n+1} = \frac{1}{2}\sum_{i=1}^{n} a_i x_i^2 \right\} \tag{3.7}$$

with the induced Riemannian metric in $\mathbb{R}^{n+1}$. Then M is of everywhere positive sectional curvature.

Proof. Let $p = (x_1, x_2, \ldots, x_{n+1}) \in M$. Then

$$M_p = \left\{ X \;\Big|\; X = \sum_{i=1}^{n+1} \alpha_i \frac{\partial}{\partial x_i},\ \alpha_{n+1} = \sum_{i=1}^{n} a_i x_i \alpha_i \right\}. \tag{3.8}$$

Set

$$Z=\sum_{i=1}^{n}a_ix_i\frac{\partial}{\partial x_i}-\frac{\partial}{\partial x_{n+1}}. \tag{3.9}$$

Then $v(p)=\frac{Z}{|Z|_0}$ is a unit normal vector field on M. Denote the connection of the usual flat metric of $\mathbb{R}^{n+1}$ by D^0. Let X, Y be vector fields on M. Then the second fundamental form of M is

$$S(X,Y)=-\left[(D^0_XY)\cdot v\right]v. \tag{3.10}$$

Set

$$X_i=\frac{\partial}{\partial x_i}+a_ix_i\frac{\partial}{\partial x_{n+1}},\quad i=1,2,\dots,n. \tag{3.11}$$

Then $X_1,X_2,\dots,X_n$ is a vector field basis of M. Since $D^0_{\frac{\partial}{\partial x_k}}\frac{\partial}{\partial x_l}=0$ for $1\le k,\,l\le n+1$, we obtain from (3.11)

$$D^0_{X_i}X_j=X_i(a_jx_j)\frac{\partial}{\partial x_{n+1}}=\frac{\partial}{\partial x_i}(a_jx_j)\frac{\partial}{\partial x_{n+1}}=\delta_{ij}a_j\frac{\partial}{\partial x_{n+1}},\quad 1\le i,\,j\le n. \tag{3.12}$$

It follows from (3.10) and (3.12) that

$$S(X_i,\,X_j)=\frac{\delta_{ij}a_j}{|Z|_0}v,\quad 1\le i,\,j\le n. \tag{3.13}$$

Denote the curvature tensor of the Riemannian metric of M by $\mathbb{R}$. Let $\pi\subset M_p$ be a two-dimensional subspace and X, Y be a basis of π, where

$$X=\sum_{i=1}^{n}\alpha_iX_i\quad\text{and}\quad Y=\sum_{i=1}^{n}\beta_iX_i.$$

From (3.13) and the Gauss equation (Wu, Shen, and Yu [19]), we have

$$\begin{aligned}\mathbb{R}(X,\,Y,\,X,\,Y)&=S(X,\,X)\cdot S(Y,\,Y)-|S(X,\,Y)|_0^2\\&=\frac{1}{|Z|_0}\left[\left(\sum_{i=1}^{n}a_i\alpha_i^2\right)\left(\sum_{i=1}^{n}a_i\beta_i^2\right)-\sum_{i=1}^{n}a_i\alpha_i\beta_i\right]>0,\end{aligned} \tag{3.14}$$

since X, Y are linearly independent. Equation (3.14) completes the proof. □

Example 3.2. Let $a_i>0$ be constants, with $i=1,2,\dots,n$. Consider the operator on $\mathbb{R}^n$:

$$\mathcal{A}u=-\sum_{i=1}^{n}\frac{\partial}{\partial x_i}\left(\frac{1+\sum\limits_{j\ne i}^{n}a_j^2x_j^2}{1+\sum\limits_{k=1}^{n}a_k^2x_k^2}\frac{\partial u}{\partial x_i}\right)+\sum_{i\ne j}\frac{\partial}{\partial x_i}\left(\frac{a_ia_jx_ix_j}{1+\sum\limits_{k=1}^{n}a_k^2x_k^2}\frac{\partial u}{\partial x_j}\right).$$

Set

$$A(x) = \frac{1}{1+\sum\limits_{k=1}^{n} a_k^2x_k^2}\begin{pmatrix} 1+\sum\limits_{i=2}^{n} a_i^2x_i^2 & -a_1a_2x_1x_2 & \cdots & -a_1a_nx_1x_n \\ -a_2a_nx_1x_n & 1+\sum\limits_{i\neq 2} a_i^2x_i^2 & \cdots & -a_2a_nx_2x_n \\ \cdots & \cdots & \cdots & \cdots \\ -a_na_1x_nx_1 & -a_na_2x_nx_2 & \cdots & 1+\sum\limits_{i=1}^{n-1} a_i^2x_i^2 \end{pmatrix}. \tag{3.15}$$

Then

$$G(x) = A^{-1}(x) = \begin{pmatrix} 1+a_1^2x_1^2 & a_1a_2x_1x_2 & \cdots & a_1a_nx_1x_n \\ a_2a_nx_1x_n & 1+a_2^2x_2^2 & \cdots & a_2a_nx_2x_n \\ \cdots & \cdots & \cdots & \cdots \\ a_na_1x_nx_1 & a_na_2x_nx_2 & \cdots & 1+a_n^2x_n^2 \end{pmatrix}. \tag{3.16}$$

Consider Riemannian manifold $(\mathbb{R}^n, g)$, where Riemannian metric g is determined in the natural coordinate system $x = (x_1, x_2, \ldots, x_n)$ by

$$g = \sum_{i,j=1}^{n} (\delta_{ij} + a_ia_jx_ix_j)\, dx_idx_j. \tag{3.17}$$

Then

$$\sum_{i,j=1}^{n} (\delta_{ij} + a_ia_jx_ix_j)\, \xi_i\xi_j \geq |\xi|_0^2 \quad \forall\, x,\, \xi = (\xi_1, \xi_2, \ldots, \xi_n) \in \mathbb{R}^n. \tag{3.18}$$

It is easily checked from (3.18) that $(\mathbb{R}^n, g)$ is a complete noncompact Riemannian manifold.

Let M be the hypersurface of $\mathbb{R}^{n+1}$ given by Lemma 3.1 with the induced Riemannian metric in $\mathbb{R}^{n+1}$. It is easily verified from (3.17) and (3.8) that map $\Phi : M \to (\mathbb{R}^n, g)$, defined by

$$\Phi(p) = x \quad \forall\, p = (x_1, x_2, \ldots, x_{n+1}) \in M \tag{3.19}$$

is an isometry. By Lemma 3.1, $(\mathbb{R}^n, g)$ is of everywhere positive sectional curvature. By Corollary 1.1, $H = Dh$ verifies condition (1.16) for any $\Omega \subset \mathbb{R}^n$, where h is a strictly convex function over $(\mathbb{R}^n, g)$. $\square$

Example 3.3. Let $\mathcal{A}$ be defined by

$$\begin{aligned} \mathcal{A}u = &-\frac{\partial}{\partial x}\left(\frac{1+y^6}{1+x^2+y^6}\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial x}\left(\frac{xy^3}{1+x^2+y^6}\frac{\partial u}{\partial y}\right) \\ &- \frac{\partial}{\partial y}\left(\frac{xy^3}{1+x^2+y^6}\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(\frac{1+x^2}{1+x^2+y^6}\frac{\partial u}{\partial y}\right). \end{aligned} \tag{3.20}$$

Set

$$A(x,y) = \begin{pmatrix} \dfrac{1+y^6}{1+x^2+y^6} & \dfrac{xy^3}{1+x^2+y^6} \\ \dfrac{xy^3}{1+x^2+y^6} & \dfrac{1+x^2}{1+x^2+y^6} \end{pmatrix}. \tag{3.21}$$

Then

$$G(x,y) = (g_{ij}) = A^{-1}(x,y) = \begin{pmatrix} 1+x^2 & -xy^3 \\ -xy^3 & 1+y^6 \end{pmatrix}. \tag{3.22}$$

Consider Riemannian manifold $(\mathbb{R}^2, g)$, where Riemannian metric g is defined in the natural coordinate system (x,y) by

$$g = (1+x^2)dxdx - xy^3dxdy - xy^3dydx + (1+y^6)dydy. \tag{3.23}$$

Let surface M in $\mathbb{R}^3$ be given by

$$M = \left\{ (x,y,z) \;\middle|\; z = \frac{1}{2}x^2 - \frac{1}{4}y^4 \right\}$$

with the induced Riemannian metric in $\mathbb{R}^3$. Then the map $\Phi(x,y,z) = (x,y)$, for any $(x,y,z) \in M$, determines an isometry from M to $(\mathbb{R}^2, g)$. We obtain the Gaussian curvature of $(\mathbb{R}^2, g)$ at (x,y):

$$\begin{aligned} k(x,y) &= \text{the Gaussian curvature of } M \text{ at } (x,y,z) \\ &= \frac{-3y^2}{(1+x^2+y^6)^2} \le 0 \quad \forall\, (x,y) \in \mathbb{R}^2. \end{aligned}$$

Let $(x^0, y^0) \in \mathbb{R}^2$. By Corollary 1.3, vector field $H = \rho D\rho$ satisfies condition (1.16) for any $\Omega \subset \mathbb{R}^2$ with $a = 1$, where ρ is the distance function from (x^0, y^0) to (x,y) in metric g. □

It is not an easy task to calculate the sectional curvature of a Riemannian manifold in general. For convenience, we give a lemma.

Lemma 3.2. *Let* $\mathbb{R}^2$ *have the metric*

$$g = g_1 dxdx + g_2 dydy,$$

where $g_1 > 0$, $g_2 > 0$ *are* C^∞ *functions on* $\mathbb{R}^2$. *Then the Gaussian curvature is*

$$\begin{aligned} k = \frac{1}{4g_1^2g_2^2}\Bigg[& g_2 \frac{\partial g_1}{\partial x}\frac{\partial g_2}{\partial x} + g_1 \frac{\partial g_1}{\partial y}\frac{\partial g_2}{\partial y} \\ & + g_1\left(\frac{\partial g_2}{\partial x}\right)^2 + g_2\left(\frac{\partial g_1}{\partial y}\right)^2 - 2g_1g_2\left(\frac{\partial^2 g_1}{\partial y^2} + \frac{\partial^2 g_2}{\partial x^2}\right)\Bigg]. \end{aligned} \tag{3.24}$$

Proof. Set

$$X = \frac{1}{\sqrt{g_1}}\frac{\partial}{\partial x}, \quad Y = \frac{1}{\sqrt{g_2}}\frac{\partial}{\partial y}.$$

Then X, Y is an orthonormal frame on $(\mathbb{R}^2, g)$. We thus have

$$k = \langle R_{XY}X, Y\rangle = \frac{1}{g_1g_2}\left\langle R_{\frac{\partial}{\partial x}\frac{\partial}{\partial y}}\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right\rangle, \tag{3.25}$$

where $R_{XY}X = -D_XD_YX + D_YD_XX + D_{[X,Y]}X$ and R is the curvature tensor. Let Γ_{ij}^k be the coefficients of connection D. Then

$$\Gamma_{12}^1 = \frac{1}{2g_1}\frac{\partial g_1}{\partial y}, \quad \Gamma_{11}^2 = -\frac{1}{2g_2}\frac{\partial g_1}{\partial y}, \quad \Gamma_{12}^2 = \frac{1}{2g_2}\frac{\partial g_2}{\partial x}.$$

From the properties of connection, we have

$$\begin{aligned} D_{\frac{\partial}{\partial x}} D_{\frac{\partial}{\partial y}} \frac{\partial}{\partial x} &= D_{\frac{\partial}{\partial x}} \left(\Gamma_{12}^1 \frac{\partial}{\partial x} + \Gamma_{12}^2 \frac{\partial}{\partial y} \right) \\ &= (\cdots) \frac{\partial}{\partial x} + \Gamma_{12}^1 D_{\frac{\partial}{\partial x}} \frac{\partial}{\partial x} + \frac{\partial \Gamma_{12}^2}{\partial x} \frac{\partial}{\partial y} + \Gamma_{12}^2 D_{\frac{\partial}{\partial x}} \frac{\partial}{\partial y} \\ &= (\cdots) \frac{\partial}{\partial x} + \left(\Gamma_{12}^1 \Gamma_{11}^2 + \frac{\partial \Gamma_{12}^2}{\partial x} + \Gamma_{12}^2 \Gamma_{12}^2 \right) \frac{\partial}{\partial y}. \end{aligned} \tag{3.26}$$

Thus

$$\begin{aligned} \left\langle D_{\frac{\partial}{\partial x}} D_{\frac{\partial}{\partial y}} \frac{\partial}{\partial x},\ \frac{\partial}{\partial y} \right\rangle_g &= \left(\Gamma_{12}^1 \Gamma_{11}^2 + \frac{\partial \Gamma_{12}^2}{\partial x} + \Gamma_{12}^2 \Gamma_{12}^2 \right) g_2 \\ &= \frac{1}{2} \frac{\partial^2 g_2}{\partial x^2} - \frac{1}{4 g_1} \left(\frac{\partial g_1}{\partial y} \right)^2 - \frac{1}{4 g_2} \left(\frac{\partial g_2}{\partial x} \right)^2, \end{aligned} \tag{3.27}$$

since $\langle \frac{\partial}{\partial x},\ \frac{\partial}{\partial y} \rangle_g = 0$. A calculation similar to (3.27) gives

$$\left\langle D_{\frac{\partial}{\partial y}} D_{\frac{\partial}{\partial x}} \frac{\partial}{\partial x},\ \frac{\partial}{\partial y} \right\rangle_g = -\frac{1}{2} \frac{\partial^2 g_1}{\partial y^2} + \frac{1}{4 g_1} \frac{\partial g_1}{\partial x} \frac{\partial g_2}{\partial x} + \frac{1}{4 g_2} \frac{\partial g_1}{\partial y} \frac{\partial g_2}{\partial y}. \tag{3.28}$$

Equations (3.27) and (3.28), together with (3.25), yield (3.24). □

Example 3.4. Let $\mathcal{A}$ be the operator

$$\mathcal{A} u = -\frac{\partial}{\partial x} \left(e^{x^3 + y^3} \frac{\partial u}{\partial x} \right) - \frac{\partial}{\partial y} \left(e^{x^3 + y^3} \frac{\partial u}{\partial y} \right). \tag{3.29}$$

Then Riemannian manifold $(\mathbb{R}^2, g)$ is with a metric

$$g = e^{-x^3 - y^3} (dx^2 + dy^2). \tag{3.30}$$

By Lemma 3.2, we have

$$k(x, y) = 3(x + y) e^{x^3 + y^3}, \quad (x, y) \in \mathbb{R}^2. \tag{3.31}$$

By (3.31), we obtain

$$\sup_{(x,y) \in \mathbb{R}^2} K(x, y) = \infty.$$

Take

$$H = -e^{-x} \frac{\partial}{\partial x} - e^{-y} \frac{\partial}{\partial y}, \tag{3.32}$$

and we obtain by calculation

$$\langle D_X H,\ X \rangle_g = \sum_{i,j=1}^{2} h_{ij} X_i X_j \quad \forall\, X = X_1 \frac{\partial}{\partial x} + X_2 \frac{\partial}{\partial y}, \tag{3.33}$$

where

$$(h_{ij}) = \begin{pmatrix} h e^{-x} + \frac{3}{2}(x^2 h e^{-x} + y^2 h e^{-y}) & \frac{3}{2}(x^2 e^{-y} - y^2 e^{-x}) h \\ \frac{3}{2}(y^2 e^{-x} - x^2 e^{-y}) h & h e^{-y} + \frac{3}{2}(x^2 h e^{-x} + y^2 h e^{-y}) \end{pmatrix}, \tag{3.34}$$

$$h = e^{-x^3-y^3}.$$

It follows from (3.33) and (3.34) that

$$\begin{aligned}\langle D_X H,\, X\rangle_g &= \left[he^{-x} + \frac{3}{2}(x^2he^{-x} + y^2he^{-y})\right]X_1^2 + \left[he^{-y} + \frac{3}{2}(x^2he^{-x} + y^2he^{-y})\right]X_2^2 \\ &\geq ah(X_1^2 + X_2^2) = a|X|_g^2, \quad X = X_1\frac{\partial}{\partial x} + X_2\frac{\partial}{\partial y} \in R^2_{(x,y)},\ (x,y) \in \overline{\Omega},\end{aligned}$$

where

$$a = \min_{x\in\overline{\Omega}} \left\{ e^{-x} + \frac{3}{2}(x^2e^{-x} + y^2e^{-y}),\ e^{-y} + \frac{3}{2}(x^2e^{-x} + y^2e^{-y}) \right\}.$$

Then vector field H, given by (3.32), meets condition (1.16) for any $\Omega \subset \mathbb{R}^2$.

Moreover, H is not the Hessian of any function on $\mathbb{R}^2$ in Riemannian metric g since matrix (h_{ij}), given by (3.34), is not symmetric. □

Example 3.5. Let $\mathcal{A}$ be given by

$$\mathcal{A}u = -\frac{\partial}{\partial x}\left(e^{x+y}\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(e^{x+y}\frac{\partial u}{\partial y}\right). \tag{3.35}$$

Riemannian manifold $(\mathbb{R}^2,\, g)$ has a metric

$$g = e^{-x-y}dxdx + e^{-x-y}dydy. \tag{3.36}$$

We obtain from Lemma 3.2 that $k(x,y) = 0$ for any $(x,y) \in \mathbb{R}^2$; that is, $(\mathbb{R}^2, g)$ is of zero curvature. By Corollary 1.3, vector field $H = \rho D\rho$ is satisfied with condition (1.16) with $a = 1$ for any $\Omega \subset \mathbb{R}^2$, where ρ is the distance function over Riemannian manifold $(\mathbb{R}^2, g)$. □

Example 3.6. Consider the operator

$$\mathcal{A}u = -\frac{\partial}{\partial x}\left(e^{x+y}\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(e^{-x-y}\frac{\partial u}{\partial y}\right).$$

Riemannian manifold $(\mathbb{R}^2,\, g)$ has the metric

$$g = e^{x+y}dxdx + e^{-x-y}dydy.$$

By Lemma 3.2 we obtain $k(x,y) = -\frac{1}{2}(e^{-x-y} + e^{x+y}) < 0$ for any $(x,y) \in \mathbb{R}^2$. $(\mathbb{R}^2, g)$ is of everywhere negative curvature. Corollary 1.3 can be applied. □

4. Necessary conditions for exact controllability and counterexamples. Given the linear partial differential operator P,

$$Pu = \frac{\partial^2 u}{\partial t^2} - \mathcal{A}u, \tag{4.1}$$

acting on functions, one sets $f_\alpha = e^{i\alpha(x\cdot\zeta + t\tau)}$ and

$$p(x,t,\zeta,\tau) = \lim_{\alpha\wr\infty}[\overline{f}_\alpha P(f_\alpha)](x,t) = \sum_{ij=1}^{n} a_{ij}(x)\zeta_i\zeta_j - \tau^2. \tag{4.2}$$

Function p is the principal symbol of P.

Given $(x_0, t_0, \zeta_0) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n/0$, $(x(s), t(s), \zeta(s), \tau(s))$ is a null bicharacteristic curve through (x_0, t_0, ζ_0) if it satisfies the (Hamiltonian) system of ordinary differential equations

$$x' = \frac{1}{2}\frac{\partial p}{\partial \zeta}, \quad t' = \frac{1}{2}\frac{\partial p}{\partial \tau}, \quad \zeta' = -\frac{1}{2}\frac{\partial p}{\partial x}, \quad \text{and} \quad \tau' = -\frac{1}{2}\frac{\partial p}{\partial t} \tag{4.3}$$

with $(x(0), t(0), \zeta(0)) = (x_0, t_0, \zeta_0)$ and $\tau(0)$ chosen so that $p(x_0, t_0, \zeta_0, \tau(0)) = 0$. To abbreviate the terminology, the projection of a bicharacteristic curve is called a ray.

Lemma 4.1. *Let $x(t)$ be a geodesic on Riemannian manifold $(\mathbb{R}^n, g)$ with $x(0) = x_0$ parameterized by arc length. Then*

$$(x(t), \pm t, \zeta(t), \mp 1) \tag{4.4}$$

are bicharacteristic curves through $(x_0, 0, \zeta_0)$, where

$$\zeta(t) = \sum_{i=1}^n \left[\sum_{j=1}^n g_{ij}(x(t))x_j'(t)\right] \frac{\partial}{\partial x_i}, \quad \zeta_0 = \zeta(0). \tag{4.5}$$

Proof. It will suffice to verify (4.3). From (4.5), we obtain

$$x_i'(t) = \sum_{j=1}^n a_{ij}(x(t))\zeta_j(t). \tag{4.6}$$

It follows from (4.2) and (4.6) that

$$x'(t) = \frac{1}{2}\frac{\partial p}{\zeta}. \tag{4.7}$$

In addition, we obtain

$$\sum_{l=1}^n \frac{\partial g_{lp}}{\partial x_i} a_{lk} = -\sum_{l=1}^n g_{lp}\frac{\partial a_{lk}}{\partial x_i}, \tag{4.8}$$

since $\sum_{l=1}^n g_{lp}a_{lk} = \delta_{pk}$.

Since $x(t)$ is a geodesic, we have

$$x_i'' + \sum_{lp=1}^n \Gamma_{lp}^i x_l' x_p' = 0, \tag{4.9}$$

where Γ_{lp}^i are the coefficients of the Levi–Civita connection, given by

$$\Gamma_{lp}^i = \frac{1}{2}\sum_{h=1}^n a_{ih}\left(\frac{\partial g_{hl}}{\partial x_p} + \frac{\partial g_{hp}}{\partial x_l} - \frac{\partial g_{lp}}{\partial x_h}\right). \tag{4.10}$$

It follows from (4.5), (4.9), (4.10), and (4.8) that

$$\zeta_i'(t) = \left(\sum_{j=1}^n g_{ij}(x(t))x_j'(t)\right)'$$

$$\begin{aligned}
&= \sum_{kj=1}^{n} \frac{\partial g_{ij}}{\partial x_k} x_k' x_j' + \sum_{j=1}^{n} g_{ij}(x(t)) x_j''(t) \\
&= \sum_{kj=1}^{n} \frac{\partial g_{ij}}{\partial x_k} x_k' x_j' - \sum_{jlp} g_{ij} \Gamma_{lp}^{j} x_l' x_p' \\
&= \frac{1}{2} \sum_{lp} \frac{\partial g_{lp}}{\partial x_i} x_l' x_p' \\
&= \frac{1}{2} \sum_{lpkj} g_{lp} \frac{\partial a_{lk}}{\partial x_i} a_{pj} \zeta_k \zeta_j = -\frac{1}{2} \sum_{kj} \frac{\partial a_{jk}}{\partial x_i} \zeta_k \zeta_j = -\frac{1}{2} \frac{\partial p}{\partial x_i};
\end{aligned}$$

that is,

$$\zeta'(t) = -\frac{1}{2} \frac{\partial p}{\partial x}.$$

Furthermore, (4.5) and (4.2) yield also

$$p(x(t), \pm t, \zeta(t), \mp 1) = \sum_{ijkl} a_{ij} g_{ik} g_{jl} x_k' x_l' - 1 = \sum_{kl} g_{kl} x_k' x_l' - 1 = 0,$$

since $x(t)$ is parameterized by arc length. This completes our proof. □

By Lemma 4.1 and Bardos, Lebeau, and Rauch [1, Thm. 3.2], the following result is immediate.

THEOREM 4.1. *If there is a closed geodesic that is contained in Ω, then the system in* (1.1), (1.1a) *has no exact controllability where $\Gamma_0 = \Gamma$; that is, for any $T > 0$,*

$$F_{T,\Gamma} \neq H_0^1(\Omega) \times L^2(\Omega).$$

Example 4.1. Consider the operator

$$\mathcal{A}u = -\frac{\partial}{\partial x}\left((1 + x^2 + y^2)^2 \frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left((1 + x^2 + y^2)^2 \frac{\partial u}{\partial y}\right). \tag{4.11}$$

The Riemannian metric g on $\mathbb{R}^2$ is

$$g = \frac{dxdx + dydy}{(1 + x^2 + y^2)^2}. \tag{4.12}$$

Set

$$B_1 = \{ (x, y) \,|\, x^2 + y^2 < 1 \}, \quad S = \{ (x, y) \,| x^2 + y^2 = 1 \}.$$

We have the following conclusions:

(1) If $\overline{\Omega} \subset B_1$ or $\overline{\Omega} \subset \mathbb{R}^2 - B_1 \bigcup S$, then Theorem 1.1 holds true.

(2) If $S \subset \Omega$, then the system in (1.1), (1.1a) has no exact controllability; that is for any $T > 0$,

$$F_{T,\Gamma} \neq H_0^1(\Omega) \times L^2(\Omega).$$

Proof. From Lemma 3.2, we obtain $k(x, y) = 4$ for any $(x, y) \in \mathbb{R}^2$. $(\mathbb{R}^2, g)$ is then a space form of constant curvature 4. Let M be the sphere of radius $\frac{1}{2}$ and center $(0, 0, \frac{1}{2})$ in $\mathbb{R}^3$, given by

$$M = \{ (x_1, x_2, x_3) \,|\, x_1^2 + x_2^2 + x_3^2 = x_3 \} \tag{4.13}$$

with the induced Riemannian metric in $\mathbb{R}^3$. Then $(\mathbb{R}^2 \cup \{\infty\},\, g)$ is isometric to M with isometry $\Phi : M \to (\mathbb{R}^2 \cup \{\infty\},\, g)$ defined by

$$\Phi(x_1, x_2, x_3) = \left(\frac{x_1}{1 - x_3},\ \frac{x_2}{1 - x_3}\right) \quad \forall \quad (x_1, x_2, x_3) \in M. \tag{4.14}$$

It is easily checked that

$$\sup_{(x,y)\in\mathbb{R}^2} d_g(0, (x, y)) = \frac{\pi}{2}, \qquad \sup_{x^2+y^2=1} d_g(0, (x, y)) = \frac{\pi}{4}.$$

This shows that we can obtain a geodesic ball $B((x^0, y^0), \gamma)$ in $(\mathbb{R}^2 \cup \{\infty\},\, g)$ such that

$$\overline{\Omega} \subset B((x^0, y^0), \gamma) \quad \text{and} \quad \gamma < \frac{\pi}{4}$$

when $\overline{\Omega} \subset B_1$ or $\overline{\Omega} \subset \mathbb{R}^2 - B_1 \bigcup S$. Part 1 then follows from Corollary 1.2.

In order to prove part 2 it will suffice from Theorem 4.1 to prove that the unit circle S is a closed geodesic on $(\mathbb{R}^2, g)$. It is well known that the big circle $C = \{\, (x_1, x_2, \frac{1}{2}) \mid x_1^2 + x_2^2 = \frac{1}{4}\,\}$ is a closed geodesic on the sphere M. In addition, it is easily checked that

$$S = \Phi(C),$$

where Φ is given by (4.14). This gives our desired result since Φ is an isometry from M to $(\mathbb{R}^2, g)$.

Appendix. A sketch of the proof of Theorem 1.1 for the case $n = 1$. Let $l > 0$ and $a(\cdot) \in C^1[0, 1]$, $a(x) > 0$ for any $x \in [0, l]$.

For system

$$\begin{cases} \phi_{tt} = (a(x)\phi_x)_x, & 0 < t < T, \quad 0 < x < l, \\ \phi(t, 0) = \phi(t, l) = 0, & 0 < t < T, \\ \phi(0, x) = \varphi_0(x), \quad \phi_t(0, x) = \varphi_1(x), & 0 < x < l, \end{cases} \tag{A.1}$$

we use multiplier $h\phi_x$, where h is a function on the interval $[0, l]$, and we obtain

$$\begin{aligned} \frac{1}{2}\int_0^T\int_0^l (ah\phi_x^2)_x dxdt = (\phi_t, h\phi_x)\Big|_0^T &+ \frac{1}{2}\int_0^T\int_0^l h_x(\phi_t^2 + a(x)\phi_x^2)\, dxdt \\ &- \frac{1}{2}\int_0^T\int_0^l a_x h\phi_x^2\, dxdt. \end{aligned} \tag{A.2}$$

Let h be the solution of the following problem:

$$\begin{cases} h_x = \dfrac{b}{a}h + 1, \\ h(0) = 0, \end{cases} \tag{A.3}$$

where $b(x) = \max(a_x, 0)$, e.g.,

$$h(x) = e^{\int_0^x \frac{b}{a} dx} \int_0^x e^{-\int_0^s \frac{b}{a} d\tau} ds, \qquad 0 \le x \le l. \tag{A.4}$$

We may obtain, when h, defined by (A.4), enters (A.2),

$$a(l)h(l)\int_0^T \phi_x^2(t, l)\, dt \ge 2TE(0) + 2(\phi_t, h\phi_x)\Big|_0^T, \qquad T > 0, \tag{A.5}$$

where $E(0) = \frac{1}{2}\int_0^l (\phi_t^2 + a\phi_x^2)\, dx$. Inequality (A.5) yields our desired result.

Acknowledgments. I would like to express my thanks to Professors D. X. Feng, I. Lasiecka, and J. L. Lions for their valuable advice. Special thanks go to Professor R. Triggiani for his valuable advice and kindly help.

REFERENCES

[1] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[2] J. Cheeger and D. Ebin, *Comparison Theorem in Riemannian Geometry*, North-Holland, Amsterdam, 1975.

[3] G. Chen, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249–274.

[4] R. E. Greene and H. Wu, *C^∞ convex functions and manifolds of positive curvature*, Acta Math., 137 (1976), pp. 209–245.

[5] R. E. Greene and H. Wu, *Function Theory on Manifolds Which Possess a Pole*, Lecture Notes in Math. 699, Springer-Verlag, New York, 1979.

[6] V. Komornik, *Exact Controllability and Stabilization*, RAM Res. Appl. Math., Masson/John Wiley, Paris, Chichester, 1994.

[7] J. Lagnese, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163–182.

[8] J. Lagnese, *The Hilbert uniqueness method: A retrospective*, in Optimal Conrol of Partial Differential Equations, Lecture Notes in Control and Inform. Sci. 148, M. Thoma and A. Wyner, eds., Springer-Verlag, New York, 1990, pp. 158–181.

[9] I. Lasiecka, J. L. Lions, and R. Triggiani, *Nonhomogeneous boundary value problems for second-order hyperbolic operators*, J. Math. Pures Appl., 65 (1981), pp. 149–192.

[10] I. Lasiecka and R. Triggiani, *Exact controllability of the wave equation with Neumann boundary control*, Appl. Math. Optim., 19 (1989), pp. 243–290.

[11] J. L. Lions, *Contrôlabilité exacte, stabilisation et perturbations des systémes distribués*, Vol. 1, Contrôlabilité exacte (E. Zuazua, compiler); Vol. 2, Perturbations; Vol. 3, Stabilisation, Masson, Paris, 1988.

[12] J. L. Lions, *Exact controllability, stabilization and perturbations for distributed systems*, J. von Neumann Lecture, SIAM Rev., 30 (1988), pp. 1–68.

[13] J. Ralston, *Gaussian beams and the propagation of singularities*, in Studies in Partial Differential Equations, MAA Stud. Math. 23, W. Littman, ed., Mathematical Association of America, Washington, DC, 1982, pp. 204–248.

[14] M. Spivak, *A Comprehensive Introduction to Differential Geometry* II, 2nd ed., Publish or Perish, Berkeley, CA, 1979.

[15] D. Tataru, *A-priori estimates of Carleman's type in domains with boundary*, J. Math. Pures Appl., 73 (1994), pp. 355–387.

[16] D. Tataru, *Boundary controllability for conservative PDEs*, Appl. Math. Optim., 31 (1995), pp. 257–295.

[17] D. Tataru, *Carleman estimates and unique continuation for solutions to boundary value problems*, J. Math. Pures Appl., 75 (1996), pp. 367–408.

[18] R. Triggiani, *Exact boundary controllability of $L_2(\Omega) \times H^{-1}(\Omega)$ of the wave equation with Dirichlet boundary control acting on a portion of the boundary and related problems*, Appl. Math. Optim., 18 (1988), pp. 241–277.

[19] H. Wu, C. L. Shen, and Y. L. Yu, *An Introduction to Riemannian Geometry*, University of Beijing, 1989 (in Chinese).

[20] P. F. Yao, *The observability inequalities for the Euler-Bernoulli equations with variable coefficients in a bounded region*, to appear.

SIAM J. Control Optim.
Vol. 37, No. 5, pp. 1600–1613

AVERAGING METHOD FOR ONE-SIDED LIPSCHITZ DIFFERENTIAL INCLUSIONS WITH GENERALIZED SOLUTIONS[*]

TZANKO DONCHEV[†] AND IORDAN SLAVOV[‡]

Abstract. We investigate the convergence of the solution set for singularly perturbed differential inclusions using an averaging technique. The limits of the fast solutions are considered as generalized solutions (i.e., Radon probability measures) of the degenerate system. Exploiting in addition the one-sided Lipschitz condition we prove the existence of a limit of the solution set in an appropriate topology.

Key words. differential inclusion, singular perturbation, one-sided Lipschitz, generalized solution, averaging method

AMS subject classifications. 34A60, 34E15, 34C29

PII. S0363012997321371

1. Introduction.

Consider the following singularly perturbed multivalued Cauchy problem:

$$\begin{pmatrix} \dot{x}(t) \\ \varepsilon\dot{y}(t) \end{pmatrix} \in H(t, x(t), y(t)), \quad x(0) = x^0, y(0) = y^0. \tag{1.1}$$

The real parameter $\varepsilon > 0$ represents the singular perturbation. Here $x \in \mathbf{R}^n, y \in \mathbf{R}^m$, while $H : I(:= [0,1]) \times \mathbf{R}^n \times \mathbf{R}^m \Longrightarrow \mathbf{R}^{n+m}$ is a multifunction. The variable x is commonly called a “slow” variable, while y is called a “fast” variable. The solution set $Z(\varepsilon), \varepsilon > 0$ of (1.1) consists of all absolutely continuous (AC) functions (x, y) satisfying (1.1) for almost everywhere (a.e.) $t \in I$. For $\varepsilon = 0^+$ it is natural to mean by $Z(0)$ the set of all pairs (x, y) with x-AC and y-integrable on I, satisfying for a.e. $t \in I$

$$\begin{pmatrix} \dot{x}(t) \\ 0 \end{pmatrix} \in H(t, x(t), y(t)), \quad x(0) = x^0. \tag{1.2}$$

The connection between the inclusions (1.1) and (1.2) has been investigated in many papers [5], [6], [7], [16], [17]. For example, in [6] the upper semicontinuity (USC) of the map $\varepsilon \to Z(\varepsilon)$ at $\varepsilon = 0^+$ in $C(I, \mathbf{R}^n) \times (L^2(I, \mathbf{R}^m)$-weak) topology is shown. The lower semicontinuity (LSC) is proved first in [16] and afterwards for more general systems in [4]. However, the question of the continuity (USC plus LSC in one and the same topology) of $Z(\varepsilon)$ at $\varepsilon = 0^+$ has no satisfactory answer. The limit of the solution set is evaluated in some simple cases and is not equal to $Z(0)$ (see [7]) which also makes the problem so complicated.

Our aim is to find conditions and “reasonable” topology in which $Z(\varepsilon)$ has a limit when $\varepsilon \to 0^+$. We consider reasonable any topology implying the continuity of an

[*] Received by the editors May 9, 1997; accepted for publication (in revised form) November 24, 1998; published electronically August 10, 1999. This research was partially supported by National Foundation for Scientific Research at the Bulgarian Ministry of Science and Education grants MM–701/97 and MM–807/98.

http://www.siam.org/journals/sicon/37-5/32137.html

[†] Department of Mathematics, University of Mining and Geology, 1100 Sofia, Bulgaria (donchev@or.math.bas.bg).

[‡] Institute of Applied Mathematics and Informatics, Technical University, 1000 Sofia, Bulgaria (iis@vmei.acad.bg).

integral cost minimized over $Z(\varepsilon)$, i.e., the trajectories of (1). Here we combine the averaging technique elaborated in [14], [12], [9], [10], [11] with the notion of generalized solutions (Radon probability measures) to obtain the existence of the limit of $Z(\varepsilon)$ at $\varepsilon = 0^+$. Let us note that our settings are more general and assumptions are weaker than those in [11] where a similar problem is investigated.

We consider the next particular case of (1.1):

$$\begin{aligned} &\dot{x}(t) \in F(t, x, y, u(t)), \quad x(0) = x^0, \\ &\varepsilon \dot{y}(t) \in G(x, y, u(t)), \quad y(0) = y^0. \end{aligned} \tag{1.3}$$

Here $u(t) \in U$, $u(\cdot)$ is measurable and U is a compact subset of a complete metric space. We will call all such functions *admissible controls.*

The choice of the above special form of (1.1) is dictated by the use of averaging techniques in our proofs. Note also that (1.3) is more general than a system of two differential inclusions (just fix the measurable function $u(\cdot)$ in (1.3)) and than a control system as well (in this case F and G are single-valued for every $u(\cdot)$). These are the cases studied many times using the method of averaging; see references [9], [10], [14]. For the general system (1.1) we do not know any relevant results. On the other hand, (1.3) also covers the case of singularly perturbed control systems with additional uncertainties or disturbances different than the control.

The fundamental theorem of Tikhonov [15] states that for single-valued H under appropriate conditions the unique solution of (1.1) converges as $\varepsilon \to 0$ to a special solution of (1.2) in $C(I, \mathbf{R}^n) \times C([\delta, 1], \mathbf{R}^m)$, for every $0 < \delta < 1$. While the LSC of the solution set of (1.1), (1.2) in Tikhonov's topology is not difficult to prove, the USC is. This problem is solved partially in [14], [17] under restrictive assumptions. Here, we suppose that the y-part of the solution set is embedded in the space of all Radon probability measures $\Re$, i.e., we consider y-variables as additional controls which are called, in optimal control theory, relaxed. Thus we pay special attention to the influence of the solutions of (1.1) with rapidly oscillating y-parts on the limit behavior of the whole solution set. First, we prove for (1.1) that $Z(\varepsilon)$ is USC at $\varepsilon = 0^+$ where in y we use the $[L^1(I, C(K))]^*$-weak* topology (K is the compact set defined after Lemma 3.1 which contains all y-parts of $Z(\varepsilon)$). This topology induces a convergence which is equivalent to the well-known weak convergence of measures in $\Re$.

This way we follow the ideas of [19] which very recently were developed for systems of singularly perturbed ODEs and control systems in [2], [18]. The class of measures considered in the cited papers is narrower, namely measures *invariant* with respect to the flow generated by the *associated* system (see below). Further attempt to apply the same approach is made in [1], where invariant measures for differential inclusions are introduced (see next section for details).

As mentioned above, the object we investigate is more general. Moreover, Example 1 in this paper tells us that the class of invariant measures still is not narrow enough to ensure LSC of $Z(\varepsilon)$ at $\varepsilon = 0^+$. Thus some additional conditions (e.g., additional restrictions on the class of measures used) should be imposed to describe the limit set.

We continue with a brief description of the averaging technique. Fix $t \in I$ and consider the following *associated* system:

$$\begin{aligned} &x = \text{const.}, \\ &\dot{y}(\tau) \in G(x, y(\tau), u(\tau)), \quad y(0) \in Q \subset \mathbf{R}^m, \quad u(\tau) \in U, \quad \tau \geq 0. \end{aligned} \tag{1.4}$$

For given x and t, denote (using Aumann's integral)

$$\bar{V}(t,x,S,Q) = \mathrm{cl}\left\{ \frac{1}{S}\int_0^S F(t,x,Y(\tau,x,S,Q),u(\tau))\,d\tau : u(\tau) \in U \right\},$$

where $Y(\tau,x,S,Q)$ is the solution set on the interval $[0,S]$ of (1.4) and **cl** denotes the closed hull. In the third section of the paper, following the scheme of [9], we show the existence of the nondepending on Q limit (in the Hausdorff metric)

$$\bar{V}(t,x) = \lim_{S\to\infty} \bar{V}(t,x,S,Q).$$

Then exploiting further the one-sided Lipschitz property (see section 3) we prove that the "slow part," that is, the projection of $Z(\varepsilon)$ on $\mathbf{R}^n$, converges in the C-topology to the solution set of the averaged inclusion

$$\dot{x}(t) \in \bar{V}(t,x), \quad x(0) = x^0, \quad t \in I. \tag{1.5}$$

The above-mentioned one-sided Lipschitz (OSL) condition plays a central role in our considerations. Postponing the exact definition until the third section we would like to illustrate the essential difference between OSL and Lipschitz conditions. Suppose $f(t,\cdot)$ (for a.e. $t \in I$) and $g(\cdot)$ are Lipschitz and consider the control system

$$\begin{aligned} \dot{x}(t) &= f(t,x,y,u(t)) + \phi(x), \quad x(0) = x^0,\\ \varepsilon\dot{y}(t) &= g(x,y,u(t)) + \psi(y), \quad y(0) = y^0. \end{aligned}$$

We assume $\phi(\cdot)$ and $\psi(\cdot)$ OSL are continuous (for example, $\phi(x) = -\sqrt[3]{x}$, $\psi(y) = -\sqrt[3]{y}$). Then the right-hand side of the system is OSL but it is not Lipschitz. So, adding very natural and simple functions to a Lipschitz right-hand side we could violate the conditions in [9]. Nevertheless we show here that the averaging technique will still work if the OSL condition is met (as in the system above).

Using the possibility to approximate $Z(\varepsilon)$ (for (1.3)) with the solution set of (1.5), we prove in Theorem 3.3 that $Z(\varepsilon)$ has a Kuratowski limit (the definition is given in the last section). Our proof is not constructive as in [7]. We just benefit from the fact that the set $Z(0)$ is extended with (generalized) functions equivalent in some sense to measures.

We finish the introduction with some notations and definitions. For $A \subset R^{n+m}$, we denote by $\hat{A}$ the projection of A on $\mathbf{R}^n$ and by $\tilde{A}$ the projection of A on $\mathbf{R}^m$. Throughout the paper, $\langle\cdot,\cdot\rangle$ is the scalar product, and $|.|$ is the norm. For a set A denote by $\sigma(x,A) := \sup_{y\in A}\langle x,y\rangle$ its support function and by $D_H(A,B)$ the Hausdorff distance between the sets A, B. Recall that $C(X,Y)$, resp., $L^1(X,Y)$, is the space of all continuous functions, resp., all Lebesgue integrable functions defined on X (equipped with the Lebesgue measure) with values in Y. The multifunction F from the topological space X into the topological space Y is said to be USC (resp., LSC) at $x \in X$ when to every open $V \supset F(x)$ ($V \bigcap F(x) \neq \emptyset$) there exists a neighborhood $W \ni x$ such that $V \supset F(y)$ ($V \bigcap F(y) \neq \emptyset$) for $y \in W$. When X and Y are metrizable (metric) spaces and F is compact valued then F is USC iff it admits a compact graph restricted to a compact subset of X. Furthermore F is LSC at x iff for every $x_i \to x$, $z \in F(x)$ there exist $z_i \in F(x_i)$ with $z_i \to z$ as $i \to +\infty$.

All the concepts not discussed in detail in the following can be found in [3] or in [19].

2. Generalized (relaxed) solutions. In this section we consider (1.1), (1.2) with respect to a topology in which USC at $\varepsilon = 0^+$ of $Z(\varepsilon)$ is easy to obtain but the LSC does not hold in general. Nevertheless in the next section we prove that $\lim_{\varepsilon\to 0^+} Z(\varepsilon)$ (in the Kuratowski sense) exists in the considered topology.

First, start with the following assumptions on the right-hand-side H of (1.1).

A1. H has nonempty, convex, compact values and is bounded on bounded sets. There exist $a, b, \mu > 0$ such that

$$\sigma(x, \hat{H}(t,x,y)) \le a(1+|x|^2+|y|^2),$$
$$\sigma(y, \tilde{H}(t,x,y)) \le b(1+|x|^2) - \mu|y|^2.$$

A2. $H(\cdot,\cdot,\cdot)$ is almost continuous, i.e., for every $\delta > 0$ there exists $I_\delta \subset I$ with measure greater than $1-\delta$ such that H is continuous on $I_\delta \times \mathbf{R}^{n+m}$.

The following statement is a consequence from a result proved in [4].

LEMMA 2.1. *Under* A1 *and* A2 *the solution set* $Z(\varepsilon)$ *of* (1.1) *is nonempty, every solution* $(x,y) \in Z(\varepsilon)$ *exists on the whole* I, *and there exist constants* $M, N > 0$ *such that* $|x(t)| + |y(t)| \le M$ *and* $|H(t,x(t),y(t))| \le N$ *for every* $t \in I$ *and* $\varepsilon > 0$.

This result allows us to consider the "fast" y-parts of $Z(\varepsilon)$ as measures over the *compact* set $K = \{y \in \mathbf{R}^m : |y| \le M\}$. To this end let $\Re(K)$ be the set of all Radon probability measures on K equipped with the weak convergence topology and define the set of functions

$$\wp := \{\nu : I \to \Re(K) \mid \nu(\cdot) \text{ is measurable}\}.$$

Then if every point $y \in K$ is considered as the Dirac measure δ_y concentrated at the point y (i.e., $\delta_y(\{y\}) = 1$) we can represent every measurable function $y : I \to K$ as $\bar{\nu}(\cdot) = \delta_{y(\cdot)}$ which is an element of $\wp$.

DEFINITION 2.1. *The pair* (x,ν) *of AC function* x *and* $\nu \in \wp$ *is said to be a generalized (relaxed) solution of the degenerate inclusion* (1.2) *when* $x(0) = x^0$ *and for a.e.* $t \in I$

$$\begin{pmatrix} \dot{x}(t) \\ 0 \end{pmatrix} \in \int_K H(t,x(t),y)\nu(t)\,(dy),$$

where dy *indicates that the integration is done with respect to* y. *Denote the set of the generalized solutions of* (1.2) *by* Z_G.

Let E^m be the space of all Carathéodory functions $f(\cdot,\cdot)$ on $I \times K$ with values in $\mathbf{R}^m$, i.e., $f(\cdot,y)$ is measurable, $f(t,\cdot)$ is continuous and integrally bounded. Then E^m is isometrically isomorphic to $L^1(I, C(K,\mathbf{R}^m))$ (see [19, Theorem I.5.25]). Moreover, from the Dunford–Pettis theorem ([19, Theorem IV.1.8]) we know that $\wp$ with the weak norm topology is isomorphic to the space $[L^1(I, C(K,\mathbf{R}^m))]^*$ equipped with the weak$*$ topology. Then $\nu_i \to \nu$ for $\nu_i, \nu \in \wp$ and $i = 1,2,\dots$ iff

$$\int_I \left(\int_K f(t,y)\nu_i(t)\,(dy)\right) dt \to \int_I \left(\int_K f(t,y)\nu(t)\,(dy)\right) dt, \quad \text{for every } f \in E^m,$$

which means that $y_i(\cdot) \in L^1(I,\mathbf{R}^m)$ converges to $\nu(\cdot)$ in $L^1(I, C(K,\mathbf{R}^m))^*$-weak* iff

$$\lim_{i\to\infty} \int_I f(t,y_i(t))\,dt = \int_I \left(\int_K f(t,y)\nu(t)\,(dy)\right) dt \tag{2.1}$$

for every $f \in E^m$.

From [19, Chapter IV], we know that $[L^1(I, C(K, \mathbf{R}^m))]^*$-weak* is separable and every closed bounded ball is metrizable. Furthermore, the set $\mathcal{B} = \{g \in L^1(I, C(K, \mathbf{R}^m)) : g(\cdot,\cdot) \text{ is Lipschitz}\}$ is a dense subset of $L^1(I, C(K, \mathbf{R}^m))$. Therefore to prove that $\{y_\varepsilon\}_{\varepsilon>0}$ converges to ν as $\varepsilon \to 0$ it is sufficient to show that $\{y_\varepsilon\}_{\varepsilon>0}$ is bounded and that (2.1) holds for any f_j, where $\mathcal{F} = \{f_j\}_{j=1}^\infty$ is dense in $\mathcal{B}$. We will use only the elements of $\mathcal{F}$ to check (2.1) in the proof of Theorem 3.3.

THEOREM 2.1. *Under the conditions* A1 *and* A2 *the map* $\varepsilon \to Z(\varepsilon)$ *is USC at* $\varepsilon = 0^+$ *in* $C(I, \mathbf{R}^n) \times [(L^1(I, C(K, \mathbf{R}^m))]^*$*-weak* topology.*

Proof. Suppose $\varepsilon_i \to 0$ and $(x_i, y_i) \in Z(\varepsilon_i)$ for every $i = 1, 2, \ldots$. The sequence $\{x_i(\cdot)\}_{i=1}^\infty$ is $C(I, \mathbf{R}^n)$ precompact due to Lemma 2.1 and to the Arzelà–Ascoli theorem. We know that $\{y_i(\cdot)\}_{i=1}^\infty$ is $[L^1(I, C(K, R^m))]^*$-weak* precompact (Theorem IV.2.1 of [19]). Therefore passing to subsequences if necessary (x_i, y_i) converges to (x_0, ν_0) and $(\dot{x}_i, \varepsilon_i \dot{y}_i)$ converges to $(\dot{x}_0, 0)$ in $L^1(I, \mathbf{R}^{m+n})$-weak. The second assertion is standard in singular perturbation theory; see, e.g., [6]. Now we will show that $(x_0, \nu_0) \in Z_G$.

Let $r \in \mathbf{R}^{n+m}$ be arbitrary and let $[s, t] \subset I$. For every i one has

$$\Big\langle r, (x_i(t) - x_i(s), \varepsilon_i(y_i(t) - y_i(s)))\Big\rangle \le \int_s^t \sigma(r, H(\tau, x_i(\tau), y_i(\tau)))\, d\tau.$$

By A2 and Lemma 2.1 we get $\sigma(r, H(\cdot, x, \cdot)) \in E^1$. Due to Theorem IV.2.9 of [19],

$$\lim_{i\to\infty} \int_s^t \sigma(r, H(\tau, x_i(\tau), y_i(\tau)))\, d\tau = \int_s^t \left\{ \int_K \sigma(r, H(\tau, x_0(\tau), y))\nu_0(\tau)\,(dy) \right\} d\tau.$$

Combining the above two inequalities we obtain

$$\Big\langle r, (x^0(t) - x^0(s), 0)\Big\rangle \le \int_s^t \left\{ \int_K \sigma(r, H(\tau, x_0(\tau), y))\nu_0(\tau)\,(dy) \right\} d\tau$$

for every $t \ge s \in I$. Consequently $x_0(0) = x^0$ and

$$\begin{pmatrix} \dot{x}(t) \\ 0 \end{pmatrix} \in \int_K H(t, x(t), y)\nu_0(t)\,(dy).$$

Hence $\varepsilon \to Z(\varepsilon)$ has $C(I, \mathbf{R}^n) \times [L^1(I, C(K, \mathbf{R}^m))]^*$ -weak* compact graph. $\square$

Remark 2.1. The result is also valid if instead of almost continuity we suppose almost USC, i.e., if we replace "continuous" by "USC" in A2.

Using the notion of invariant measures for ordinary differential inclusions introduced recently in [1] and the construction of the cited paper one can prove that Theorem 2.1 holds true if we restrict $\nu(t)$ to be an invariant measure for a.e. $t \in I$ of the associated differential inclusion

$$\dot{y}(\tau) \in \tilde{H}(\tau, x(t), y), \quad \tau \in [t, \infty).$$

Artstein defines in [1] invariant measures as limits of convex combinations of the so-called individual invariant measures. Not entering into details (exact definitions and the corresponding theory for flows can be found, e.g., in [13]), we will use the last characterization to show that restricting to invariant measures is still not sufficient to achieve LSC of $Z(\varepsilon)$ in the topology of Theorem 2.1. Consider the following example.

Example 2.1. The system

$$\begin{aligned}&\dot{x} = x + y + u(t), \quad x(0) = 0,\\ &\varepsilon\dot{y} = -y + \sqrt[3]{y}, \quad y(0) = 0, \quad u(t) \in [-1,1], \quad t \in I,\end{aligned}$$

satisfies conditions A1 and A2. Rewriting the second equation as $dy/d\tau = -y + \sqrt[3]{y}, \tau \in [0,\infty)$, it is not difficult to see that $y = -1$ and $y = 1$ are its stable resting points. Hence the Dirac measures δ_{-1} and δ_1 concentrated at -1 and 1, respectively, are *individual* invariant measures of the associated system. Then $\nu = (1/3)\delta_{-1} + (2/3)\delta_1$ is invariant measure of the associated system in the terminology of [1]. But there is no solution $(x_\varepsilon, y_\varepsilon)$ of the singularly perturbed system such that its "fast" part y_ε approaches ν in the topology considered.

3. Approximation of the solution set. In this section we consider the system (1.3) and show that $\bar{V}(t,x) = \lim_{S\to\infty} \bar{V}(t,x,S,Q)$ exists and does not depend on $Q \subset P$ where $P = \{y \in \mathbf{R}^m : |y| \le M\}$. The constant M will be determined in Lemma 3.1. Introduce the main condition.

B1. There exist positive constants A, B, D, μ such that

$$\begin{aligned}\sigma(x_1 - x_2, F(t,x_1,y_1,u)) - \sigma(x_1 - x_2, F(t,x_2,y_2,u)) &\le A|x_1 - x_2|^2 + B|y_1 - y_2|^2,\\ \sigma(y_1 - y_2, G(x_1,y_1,u)) - \sigma(y_1 - y_2, G(x_2,y_2,u)) &\le D|x_1 - x_2|^2 - \mu|y_1 - y_2|^2,\end{aligned}$$

uniformly in $u \in U$. F is almost continuous and G is continuous. They have nonempty, convex, and compact values and are bounded on bounded sets.

Condition B1 is the so-called OSL condition. Here we use the negative constant $-\mu$ in order to introduce some stability requirement on the "fast" variables y. But even with positive constants the above condition is "viable." For example, one can extend Theorem 2.1 of averaging of differential inclusions of [11] for system (1.3) replacing the Lipschitz condition used there by OSL plus continuity conditions.

Note also that

$$\begin{aligned}\sigma(x, F(t,x,y,u)) &\le \sigma(x, F(t,0,0,u)) + A|x|^2 + B|y|^2,\\ \sigma(y, G(x,y,u)) &\le \sigma(y, G(0,0,u)) + D|x|^2 - \mu|y|^2,\end{aligned}$$

i.e., if $H \equiv (F,G)$, then A1 and A2 follow from B1. Then we can prove the next lemma; see [4].

Lemma 3.1. *Under* B1 *the solution set* $Z(\varepsilon)$ *of* (1.3) *is nonempty, every solution* $(x,y) \in Z(\varepsilon)$ *exists on the whole* I*, and there exist constants* $M, N > 0$ *such that* $|x(t)| + |y(t)| \le M$ *and* $|F(t,x(t),y(t),u(t))| + |G(x(t),y(t),u(t))| \le N$ *for every* $t \in I, \varepsilon > 0$ *and* $u(\cdot)$*-admissible.*

Remark 3.1. Clearly M and N determined in Lemma 2.1 and in the last lemma are different. However, we do not distinguish them since this causes no confusion.

Lemma 3.2. *Given* $y^1, y^2 \in P$*, under* B1 *for every solution* $y_1(\cdot)$ *of* (1.4) *with* $y_1(0) = y^1$ *there exists a solution* $y_2(\cdot)$ *of* (1.4) *with* $y_2(0) = y^2$ *such that* $|y_1(\tau) - y_2(\tau)| \le \exp(-\mu\tau)|y^1 - y^2|, \tau \ge 0$.

Proof. Let $u(\cdot)$ correspond to $y_1(\cdot)$ in (1.4). Fix $u(\cdot)$ and consider the map

$$\Gamma(\tau, v) = \{w \in G(x, v, u(\tau)) : \langle y_1(\tau) - v, \dot{y}_1(\tau) - w\rangle \le -\mu|y_1(\tau) - v|^2\}.$$

Using the fashion of the proof of Theorem 1 in [16] (see also Theorem 3.2 in [4]) one can show that $\Gamma(\cdot,\cdot)$ is nonempty, convex, compact valued, and almost USC (for the last notion, see Remark 2.1). Therefore, there exists a solution $y_2(\cdot)$ of

$$\dot{z}(\tau) \in \Gamma(\tau, z(\tau)), \quad z(0) = y^2.$$

Obviously $\langle y_1(\tau)-y_2(\tau),\dot{y}_1(\tau)-\dot{y}_2(\tau)\rangle \le -\mu|y_1(\tau)-y_2(\tau)|^2$ and therefore $|y_1(\tau)-y_2(\tau)|\le m(\tau)$, where $\dot{m}\le -\mu m,\ \ m(0)=|y^1-y^2|$. □

The modulus of continuity of a multifunction is defined as follows.

Definition 3.1. *For* $\Gamma : I\times \mathbf{R}^p \Longrightarrow \mathbf{R}^q$*,* $\mathcal{A}\subset I, \mathcal{A}$ *compact, and* $\delta, a>0$ *we let*

$$\omega_\Gamma(\delta,\mathcal{A};a)=\sup\{D_H(\Gamma(t,z_1),\Gamma(s,z_2)) : |t-s|\le\delta,\ \ t,s\in\mathcal{A},\\ |z_1-z_2|\le a,\ \ |z_i|\le M\}.$$

Here M *is from Lemma* 3.1*. Clearly if* $\Gamma(\cdot,\cdot)$ *is continuous on* $\mathcal{A}\times\{z\in\mathbf{R}^p : |z|\le M\}$ *then* $\omega_\Gamma(\delta,\mathcal{A};a)\to 0$ *as* $\delta,a\to 0$*.*

In case we want to stress the dependence of the modulus on some parts of the phase variables we will put additional arguments, e.g., for $F(t,x,y,u)$ we denote the modulus by $\omega_F(\delta,\mathcal{A};a,b,c)$ for $b,c>0$.

B2. Let $f(\tau)=\omega_F(0,I;0,M\exp(-\mu\tau),0)$ and suppose that

$$\int_0^\infty f(\tau)\,d\tau = L_1<\infty.$$

Remark 3.2. The above condition holds, for example, if $F(t,x,\cdot,u)$ is Hölderian, i.e., there exist C and α such that $D_H(F(t,x,v,u),F(t,x,w,u))\le C(|v-w|^\alpha+|v-w|^{1+\alpha})$. It is easy to see that in this case $L_1=C(M^\alpha/\mu\alpha+M^{1+\alpha}/(1+\alpha)\mu)$. Furthermore, if $g(\cdot,x,\cdot)\in\mathcal{B}$, then g satisfies B2.

The following statement is an obvious consequence of Lemma 3.2.

Lemma 3.3. *Given* $y^1,y^2\in P$*, under* B1 *and* B2 *for every solution* $y_1(\cdot)$ *of* (1.4) *with* $y_1(0)=y^1$*, there exists a solution* $y_2(\cdot)$ *of* (1.4) *with* $y_2(0)=y^2$ *such that*

$$\frac{1}{S}\int_0^S D_H(F(t,x,y_1(\tau),u(\tau)),F(t,x,y_2(\tau),u(\tau)))\,d\tau\le\frac{L_1}{S}$$

for every $S>0$ *and every* $t\in I$*. Here* L_1 *is the constant from* B2*.*

Repeating with minor modifcations the reasons (not so short) in [8] we can derive the following.

Theorem 3.1. *Under the assumptions* B1 *and* B2 $\lim_{S\to\infty}\bar{V}(t,x,S,Q)=\bar{V}(t,x)$ *exists and does not depend on* $Q\subset P$*, and* $\bar{V}(t,x)$ *is closed and convex. Moreover,* $D_H(\bar{V}(t,x,S,Q),\bar{V}(t,x))\le L_2/\sqrt{S}; L_2=\text{const.}$*, for* S *sufficiently large.*

Here we deal with (1.3) under the OSL condition B1 in contrast to [8], [9], [10] where control systems under Lipschitz conditions are considered. Although B1 is weaker it enables us to derive more directly the closeness of $Z(\varepsilon)$ to the solution set of (1.5).

Lemma 3.4. $\bar{V}(\cdot,\cdot)$ *is almost continuous and OSL, i.e., there exists a constant* $\tilde{L}$ *such that*

$$\sigma(x_1-x_2,\bar{V}(t,x_1))-\sigma(x_1-x_2,\bar{V}(t,x_2))\le\tilde{L}|x_1-x_2|^2.$$

Proof. The almost continuity of $\bar{V}$ follows from the almost continuity of F and the continuity of G. Now, let $S>0$ and denote by $y(\cdot,x,u)$ any solution of (1.4) corresponding to x and admissible control $u(\cdot)$. Then fix one particular $y(\cdot,x_1,u)$ and consider

$$K=\sigma\left(x_1-x_2,\frac{1}{S}\int_0^S F(t,x_1,y(\tau,x_1,u),u(\tau))\,d\tau\right)\\ -\sigma\left(x_1-x_2,\frac{1}{S}\int_0^S F(t,x_2,y(\tau,x_2,u),u(\tau))\,d\tau\right),$$

where $y(\cdot, x_2, u)$ will be chosen later. By virtue of B1 one has

$$\begin{aligned}K &= \frac{1}{S}\int_0^S \big\{\sigma(x_1 - x_2, F(t, x_1, y(\tau, x_1, u), u(\tau)))\\ &\quad -\sigma(x_1 - x_2, F(t, x_2, y(\tau, x_2, u), u(\tau)))\big\}\, d\tau\\ &\le \frac{1}{S}\int_0^S \big(A|x_1 - x_2|^2 + B|y(\tau, x_1, u) - y(\tau, x_2, u)|^2\big)\, d\tau\\ &= A|x_1 - x_2|^2 + \frac{B}{S}\int_0^S |y(\tau, x_1, u) - y(\tau, x_2, u)|^2\, d\tau.\end{aligned}$$

Consider the map

$$\begin{aligned}&\Gamma(\tau, v) = \big\{w \in G(x_2, v, u(\tau)) : \ \langle y(\tau, x_1, u) - v, \dot{y}(\tau, x_1, u) - w\rangle\\ &\le \sigma(y(\tau, x_1, u) - v, G(x_1, y(\tau, x_1, u), u(\tau))) - \sigma(y(\tau, x_1, u) - v, G(x_2, v, u(\tau)))\big\}.\end{aligned}$$

By B1 we have

$$\langle y(\tau, x_1, u) - v, \dot{y}(\tau, x_1, u) - w\rangle \le D|x_1 - x_2|^2 - \mu|y(\tau, x_1, u) - v|^2.$$

Then Γ has nonempty, convex, and compact values and is almost USC (see the proof of Theorem 3.3 in [4]). Therefore the differential inclusion

$$\dot{v}(\tau) \in \Gamma(\tau, v(\tau)), \quad v(0) = y^0,$$

has a solution $y(\tau, x_2, u)$ for which

$$|y(\tau, x_1, u) - y(\tau, x_2, u)|^2 \le r(\tau) \text{ where } \dot{r} \le 2D|x_1 - x_2|^2 - 2\mu r, \ \ r(0) = 0.$$

Consequently

$$r(\tau) \le \exp(-2\mu\tau)\int_0^\tau 2D|x_1 - x_2|^2 \exp(2\mu s)\, ds \le \frac{D}{\mu}|x_1 - x_2|^2. \tag{3.1}$$

Hence $K \le (A + \frac{BD}{\mu})|x_1 - x_2|^2$. Now it is straightforward to prove that

$$\sigma\big(x_1 - x_2, \bar{V}(t, x_1)\big) - \sigma\big(x_1 - x_2, \bar{V}(t, x_2)\big) \le \tilde{L}|x_1 - x_2|^2,$$

where $\tilde{L} = A + BD/\mu$. □

COROLLARY 3.1. *Let $x(\cdot)$ be AC-function such that dist$(\dot{x}(t), \bar{V}(t, x(t))) = q(t)$, $t \in [0, 1]$ and $x(0) = x^0$. Then there exists a solution $z(\cdot)$ of* (1.5) *such that $|x(t) - z(t)| \le \exp(\tilde{L})\int_0^t q(s)\, ds$.*

Proof. We can argue as in the proof of Lemma 3.2, introducing the map

$$\hat{\Gamma}(t, v) = \big\{w \in \bar{V}(t, v) : \ \langle x(t) - v, \dot{x}(t) - w\rangle \le \tilde{L}|x(t) - v|^2 + |x(t) - v|q(t)\big\}. \qquad \square$$

The following two lemmas actually are the proof of the theorem of averaging of (1.3). Their proofs are similar to the corresponding results in [10], [11] but are modified to reflect the fact that some of our conditions are weaker.

LEMMA 3.5. *There exists a function $\alpha(\cdot)$ with $\lim_{\varepsilon\to 0}\alpha(\varepsilon) = 0$ such that for every solution $(x_\varepsilon, y_\varepsilon)$ of* (1.3) *there exists a solution z_ε of* (1.5) *such that $|x_\varepsilon(t) - z_\varepsilon(t)| \le \alpha(\varepsilon)$ on* $[0, 1]$.

Proof. Suppose that $\lim_{\varepsilon\to 0} S_\varepsilon = +\infty$ and $\lim_{\varepsilon\to 0} \varepsilon S_\varepsilon = 0$. Then substitute $t_j = j\varepsilon S_\varepsilon, \tau_j = jS_\varepsilon, x_j = x_\varepsilon(t_j)$ for $j = 0, 1, \dots, N_\varepsilon$ where $N_\varepsilon = [(\varepsilon S_\varepsilon)^{-1}]$ and $[s]$ is the integer part of s. Set also $t_{N_\varepsilon+1} = 1$.

Define the map

$$\begin{aligned}&P_\varepsilon(t, v, w) = \{g \in G(v, w, u_\varepsilon(t)) :\\ &\langle y_\varepsilon(t) - w, \varepsilon\dot{y}_\varepsilon(t) - g\rangle \le D|x_j - v|^2 - \mu|y_\varepsilon(t) - w|^2 + \eta(t)|y_\varepsilon(t) - w|\}.\end{aligned}$$

Here $y_\varepsilon, u_\varepsilon$ correspond to x_ε in (1.3) and $\eta(t) = dist(\varepsilon\dot{y}_\varepsilon(t), G(x_j, y_\varepsilon(t), u_\varepsilon(t)))$. It is easy to show using B1 that P_ε is nonempty, convex, and compact valued and almost USC. Define $y_z(t)$ as a solution on $[t_j, t_{j+1}], j = 0, 1, \dots, N_\varepsilon$ of the differential inclusion

$$\varepsilon\dot{y}(t) \in P_\varepsilon(t, x_j, y(t)), \quad y(t_j) = y_j \overset{def}{=} \lim_{t\to t_j^-} y(t)$$

and satisfying $y_z(0) = y^0$. Obviously $\bar{y}_z(\tau) = y_z(\varepsilon(\tau + \tau_j))$ is a solution of the associated system (1.4) with $\bar{y}_z(0) = y_j$ on the interval $[0, S_\varepsilon]$. It is easy to see that $|y_z(t) - y_\varepsilon(t)| \le r(t)$, where

$$\varepsilon\dot{r}(t) = -\mu r(t) + dist(\varepsilon\dot{y}_\varepsilon(t), G(x_j, y_\varepsilon(t), u_\varepsilon(t))), \quad r(t_j) = |y_z(t_j) - y_\varepsilon(t_j)|.$$

Thus

$$r(t) \le \exp\left(-\frac{\mu t}{\varepsilon}\right)\left\{r(t_j) + \frac{1}{\varepsilon}\int_{t_j}^{t} \exp\left(\frac{\mu s}{\varepsilon}\right) dist(\varepsilon\dot{y}_\varepsilon(s), G(x_j, y_\varepsilon(s), u_\varepsilon(s)))\, ds\right\}.$$

Due to Lemma 3.1, $x_\varepsilon(\cdot)$ is N-Lipschitz. Thus $|x_\varepsilon(t + \varepsilon S_\varepsilon) - x_\varepsilon(t)| \le \varepsilon S_\varepsilon N$. Denote $\lambda(\varepsilon) = \sup\{D_H(G(x_j, y_\varepsilon(t), u_\varepsilon(t)), G(x_\varepsilon(t), y_\varepsilon(t), u_\varepsilon(t))) : t \in [t_j, t_{j+1}]\}$. Then $\lambda(\varepsilon) \to 0$ as $\varepsilon \to 0$ since G is continuous. For $t \in [t_j, t_{j+1}]$, we have

$$r(t) \le \exp\left(-\frac{\mu t}{\varepsilon}\right)\left(\exp\left(\frac{\mu t}{\varepsilon}\right) - \exp\left(\frac{\mu t_j}{\varepsilon}\right)\right)\frac{\lambda(\varepsilon)}{\mu} + \exp\left(-\frac{\mu t}{\varepsilon}\right) r(t_j).$$

Then

$$\begin{aligned}&r(t)\\ &\le \frac{\lambda(\varepsilon)}{\mu}\exp\left(-\frac{\mu t}{\varepsilon}\right)\left(\sum_{i=0}^{j-1}\left(\exp\left(\frac{\mu t}{\varepsilon}\right) - \exp\left(\frac{\mu t_j}{\varepsilon}\right)\right) + \left(\exp\left(\frac{\mu t}{\varepsilon}\right) - \exp\left(\frac{\mu t_j}{\varepsilon}\right)\right)\right)\\ &= \frac{\lambda(\varepsilon)}{\mu}\left(1 - \exp\left(-\frac{\mu t}{\varepsilon}\right)\right) \le \frac{\lambda(\varepsilon)}{\mu}.\end{aligned}$$

By B1 for $\delta > 0$ there exists a compact set $I_\delta \subset I$ with measure greater than $1 - \delta$ such that F is continuous on $I_\delta \times \mathbf{R}^{n+m}$. Let $\delta > 0$ be so small that

$$\int_{A_\delta}\left(\sup_{|x|+|y|\le M}(|F(t, x, y)| + |\bar{V}(t, x)|)\right)^2 dt \le \frac{\varepsilon}{2},$$

where $A_\delta = I \setminus I_\delta$ and M is the constant from Lemma 3.1. Introduce the function

$$\rho_\delta(t) = \begin{cases} \sup_{|x|+|y|\le M}(|F(t, x, y)| + |\bar{V}(t, x)|), & t \in A_\delta,\\ 0 & \text{elsewhere.}\end{cases}$$

It is obvious that $\int_I \rho_\delta^2(t)\,dt < \varepsilon/2$ and $\int_I \rho_\delta(t)\,dt < (\delta\varepsilon)/2$.

For $t \in [t_j, t_{j+1}], j = 0, 1, \dots, N_\varepsilon$, define the following multifunction:

$$R_\varepsilon(t, v) = \{w \in F(t_j, x_j, y_z(t), u_\varepsilon(t)) : \langle x_\varepsilon(t) - v, \dot{x}_\varepsilon(t) - w\rangle \le A|x_\varepsilon(t) - v|^2 \\ +B|y_\varepsilon(t) - y_z(t)|^2 + \rho_\delta^2(t) + \omega_F^2(\varepsilon S_\varepsilon + \delta, I_\delta; 0)\}.$$

It is not difficult to show that $R_\varepsilon(\cdot,\cdot)$ is almost USC with nonempty, convex, and compact values. Hence there exists AC-function $\xi(\cdot)$ with

$$\dot{\xi}(t) \in R_\varepsilon(t, \xi(t)), \quad \xi(t_j) = \lim_{t \to t_j^-} \xi(t),$$

and $\xi(0) = x^0$. Therefore

$$\langle \xi(t) - x_\varepsilon(t), \dot{\xi}(t) - \dot{x}_\varepsilon(t)\rangle \le A|x_\varepsilon(t) - \xi(t)|^2 + B|y_\varepsilon(t) - y_z(t)|^2 \\ +\rho_\delta^2(t) + \omega_F^2(\varepsilon S_\varepsilon + \delta, I_\delta; 0).$$

Since $|y_\varepsilon(t) - y_z(t)| = r(t) \le \lambda(\varepsilon)/\mu$, after some standard calculations we get

$$|\xi(t) - x_\varepsilon(t)| \le C\left(\bar{\lambda}(\varepsilon) + \int_I \rho_\delta^2(t)\,dt\right) \le C(\bar{\lambda}(\varepsilon) + \varepsilon). \tag{3.2}$$

Here C is an appropriate constant and $\bar{\lambda}(\varepsilon) \to 0$ as $\varepsilon \to 0$. From Theorem 3.1 one can conclude that there exists $v_j \in \bar{V}(t_j, x_j)$, $j = 1, 2, \dots, N_\varepsilon - 1$, such that

$$\left|\frac{1}{S_\varepsilon}\int_{\tau_j}^{\tau_{j+1}} \dot{\xi}(\varepsilon\tau)\,d\tau - v_j\right| \le \frac{L_2}{\sqrt{S_\varepsilon}}.$$

Set $\psi(t) = \psi(t_j) + v_j(t - t_j)$ for $t \in [t_j, t_{j+1}], j = 0, 1, \dots, N_\varepsilon$, where $\psi(0) = x^0$. Therefore

$$|\psi(t) - \xi(t)| \le \frac{L_2}{\sqrt{S_\varepsilon}} \tag{3.3}$$

for every $t \in [0, 1]$. We have that

$$\mathit{dist}(\dot{\psi}(t), \bar{V}(t, \psi(t))) \le \mathit{dist}(\dot{\psi}(t), \bar{V}(t_j, x_j)) + D_H(\bar{V}(t_j, x_j), \bar{V}(t_j, \psi(t_j))) \\ + D_H(\bar{V}(t_j, \psi(t_j)), \bar{V}(t_j, \psi(t))) + D_H(\bar{V}(t_j, \psi(t)), \bar{V}(t, \psi(t))) \\ \le D_H(\bar{V}(t_j, x_j), \bar{V}(t_j, \psi(t_j))) + \omega_{\bar{V}}(0, I_\delta; N\varepsilon S_\varepsilon) \\ + \omega_{\bar{V}}(\varepsilon S_\varepsilon, I_\delta; 0) + 2\rho_\delta(t).$$

By (3.2) we get $|\psi(t_j) - x_j| \le |\psi(t_j) - \xi(t_j)| + |\xi(t_j) - x_\varepsilon(t_j)| \le 2M\varepsilon S_\varepsilon + 2C(\lambda(\varepsilon) + \rho_\delta(t))$. Therefore $\mathit{dist}(\dot{\psi}(t), \bar{V}(t, \psi(t))) \le \nu(\varepsilon)$ where by Definition 3.1 $\lim_{\varepsilon\to 0} \nu(\varepsilon) = 0$. Thus from Corollary 3.1 there exists a solution $z_\varepsilon(\cdot)$ of (1.5), such that $|\psi(t) - z_\varepsilon(t)| \le \exp(\tilde{L})\nu(\varepsilon)$. Using the triangle inequality, (3.2), and (3.3) we obtain the proof. □

Lemma 3.6. *There exists a function* $\alpha(\cdot)$ *tending to zero as* $\varepsilon \to 0$ *such that for every solution* z *of* (1.5) *there exists a solution* $(x_\varepsilon, y_\varepsilon)$ *of* (1.3) *such that* $|x_\varepsilon(t) - z(t)| \le \alpha(\varepsilon)$ *on* $[0, 1]$.

Proof. We will use the same partition of I and notations as in the previous proof. Let $z(\cdot)$ be a solution of (1.5). Then for $t \in [t_j, t_{j+1}], j = 0, 1, \dots, N_\varepsilon$, we have

$$\dot{z}(t) \in \bar{V}(t_j, z(t_j)) + \theta(\varepsilon)B_1(0),$$

where $\theta(\varepsilon, t)\overset{def}{=}\omega_{\bar{V}}(\varepsilon S_\varepsilon, I_\delta; C\varepsilon S_\varepsilon)+\rho_\delta(t)$. By Definition 3.1 and the definition of $\rho_\delta(t)$, we have $\int_I \theta^2(\varepsilon, t)dt \to 0$ as $\varepsilon \to 0$.

For $[t_j, t_{j+1}], j = 0, 1, \dots, N_\varepsilon$, define successively the function $w(\cdot)$ and the map

$$W(t,p) = \{v \in \bar{V}(t_j, w(t_j)) : \langle z(t) - p, \dot{z}(t) - v\rangle \le (\tilde{L} + 1)|z(t) - w(t_j)|^2 + \theta^2(\varepsilon)\},$$

where $\tilde{L} \le A + BD/\mu$ is determined in Lemma 3.3. For $j = 0$ we set $w(t_0) = w(0) = x^0$. Then $W(t,p) \ne \emptyset$ thanks to Lemma 3.1 and it is clear that $W(\cdot,\cdot)$ is compact, convex valued, and almost USC. Therefore the differential inclusion

$$\dot{w}(t) \in W(t, w), \quad w(t_j) = \lim_{t\to t_j^-} w(t)$$

admits a solution $w(t)$ on $[t_j, t_{j+1}]$. Then $|w(t) - z(t)|^2 \le r(t)$, where

$$\dot{r}(t) = (\tilde{L} + 1)r + \theta^2(\varepsilon, t), \quad r(0) = 0,$$

i.e., $r(t) \le \exp(C(t - t_j)) \int_I \theta^2(\varepsilon, s)\, ds$ for $t \in [t_j, t_{j+1})$. Here and below C is an arbitrary and sufficiently large constant.

Let $\Delta = \varepsilon S_\varepsilon$ and z_j be the projection of $\Delta^{-1} \int_{t_j}^{t_{j+1}} \dot{w}(t)\, dt$ on $\bar{V}(t_j, z(t_j))$, i.e., $z_j \in \bar{V}(t_j, z(t_j))$ and $|z_j - \Delta^{-1} \int_{t_j}^{t_{j+1}} \dot{w}(t)\, dt| = dist(\Delta^{-1} \int_{t_j}^{t_{j+1}} \dot{w}(t)\, dt, \bar{V}(t_j, z(t_j)))$. On the interval $(t_j, t_{j+1}]$ we define the control $u(t)$ to satisfy the inequality

$$dist\left(z_j, \frac{1}{\Delta}\int_{t_j}^{t_{j+1}} F(t_j, w(t_j), y_w(t), u(t))\, dt\right) \le \frac{2}{S_\varepsilon}.$$

Here $y_w(t) = \tilde{y}_w(t/\varepsilon)$ where $\tilde{y}_w(\tau)$ is a solution on $[\tau_j, \tau_{j+1}] = [jS_\varepsilon, (j + 1)S_\varepsilon]$ of

$$\dot{y}(\tau) \in G(v(\tau_j), y(\tau), u(\tau)), \quad y(\tau_j) = y_\varepsilon(\tau_j).$$

Such control exists thanks to Theorem 3.1 and the definition of $\bar{V}$. Let $(x_\varepsilon(\cdot), y_\varepsilon(\cdot))$ be a solution of

$$\begin{aligned} &\dot{x}(t) \in \tilde{F}(t, x(t), y(t), u(t)), \quad x(0) = x^0,\\ &\varepsilon\dot{y}(t) \in \tilde{G}(t, x(t), y(t), u(t)), \quad y(0) = y^0, \end{aligned}$$

where $\tilde{F}$ and $\tilde{G}$ are defined on $[t_j, t_{j+1}], j = 0, 1, \dots, N_\varepsilon$, as follows:

$$\begin{aligned} &\tilde{F}(t, \alpha, \beta, u(t)) = \{p \in F(t, \alpha, \beta, u(t)) :\\ &\langle p - z_j, \alpha - w(t)\rangle \le A|\alpha - w(t)|^2 + B|\beta - y_w(t)|^2 + 4\Delta^2 + |z_j - w(t)|^2\},\\ &\tilde{G}(t, \alpha, \beta, u(t)) = \{q \in G(\alpha, \beta, u(t)) : \langle q - \varepsilon\dot{y}_w(t), \beta - y_w(t)\rangle\\ &\le D|\alpha - w(t)|^2 - \mu|\beta - y_w(t)|^2 + 4\Delta^2 + |z_j - w(t)|^2\}. \end{aligned}$$

Thus $|x_\varepsilon(t) - z(t)|^2 \le m(t),\ |y_\varepsilon(t) - y_v(t)|^2 \le s(t)$, where

$$\begin{aligned} &\dot{m}(t) \le Am + Bs + 4\Delta^2 + r(t), \quad m(0) = 0,\\ &\varepsilon\dot{s}(t) \le Dm - \mu s + 4\Delta^2 + r(t), \quad s(0) = 0. \end{aligned}$$

Therefore, integrating by parts one obtains

$$s(t) \le \exp\left(-\frac{\mu t}{\varepsilon}\right)\int_0^t (Dm(\lambda)+4\Delta^2+r(\lambda)) \exp\left(\frac{\mu\lambda}{\varepsilon}\right) d\left(\frac{\lambda}{\varepsilon}\right) \le \frac{D}{\mu}m(t)+\frac{4\Delta^2 t}{\mu}, +\frac{r(t)}{\mu},$$

since $\dot{m}(t) \geq 0, \dot{r}(t) \geq 0$. Consequently

$$\dot{m}(t) \leq Am + \frac{BD}{\mu}m + 4\frac{\Delta^2 B}{\mu} + r(t)\frac{B}{\mu} + 4\Delta^2 + r(t).$$

That is,

$$m(t) \leq \exp\left\{\left(A + \frac{BD}{\mu}\right)t\right\}\left(4\frac{\Delta^2 B}{\mu}t + \left(\frac{B}{\mu} + 1\right)\int_0^t r(\lambda)d\,\lambda + 4\Delta^2 t\right),$$

where $t \in [0, \Delta]$. Thus

$$m(t) \leq \exp(C)\left(4\frac{\Delta^2 B}{\mu} + 4\Delta^2 + \Delta\right) \leq C\varepsilon S_\varepsilon. \qquad \square$$

Now the first principal result in this section follows from Lemma 3.5 and Lemma 3.6. Namely, the projection $X(\varepsilon) = \hat{Z}(\varepsilon)$ can be approximated by the solution set of the averaged inclusion (1.5).

THEOREM 3.2. *Under the conditions* B1 *and* B2 *there exists* $\alpha(\cdot)$ *such that* $\lim_{\varepsilon\to 0^+} \alpha(\varepsilon) = 0$ *and* $D_H(X(\varepsilon), X_0) = \alpha(\varepsilon)$, *where* $X(\varepsilon)$ *is the "slow" part of the solution set of* (1.3) *and* X_0 *is the solution set of* (1.5).

Now we are able to prove the main result in the paper, namely that the solution set of (1.3) (denote it again by $Z(\varepsilon)$) possesses a limit in topology $\mathcal{T} = C(I, \mathbf{R}^n) \times [L^1(I, C(K, \mathbf{R}^m))]^*$-weak*.

DEFINITION 3.2. *We will say that* $Z_0 = \lim_{\varepsilon\to 0^+} Z(\varepsilon)$ *in* $\mathcal{T}$ *iff*

(a) $Z_0 \subset \liminf_{\varepsilon\to 0^+} Z(\varepsilon)$, *i.e., for every* $(x_0, \nu_0) \in Z_0$ *there exists* $(x_\varepsilon, y_\varepsilon) \in Z(\varepsilon), \varepsilon > 0$ *such that* $(x_\varepsilon, y_\varepsilon) \to (x_0, y_0)$ *in* $\mathcal{T}$;

(b) $\limsup_{\varepsilon\to 0^+} Z(\varepsilon) \subset Z_0$, *i.e., all cluster points in* $\mathcal{T}$ *of every* $\{(x_\varepsilon, y_\varepsilon)\}_{\varepsilon>0}$ *with* $(x_\varepsilon, y_\varepsilon) \in Z(\varepsilon)$ *are contained in* Z_0.

Since $\mathcal{T}$ is metrizable we consider only countable subsequences.

THEOREM 3.3. *Suppose* B1 *and* B2 *are fulfilled. Then there exists the limit* $Z_0 = \lim_{\varepsilon\to 0^+} Z(\varepsilon) \subset Z_G$ *in* $C(I, \mathbf{R}^n) \times [L^1(I, C(K, \mathbf{R}^m))]^*$*-weak* topology.*

Proof. By Theorem 2.1 it follows that $\limsup_{\varepsilon\to 0^+} Z(\varepsilon) \neq \emptyset$ (in $\mathcal{T}$) and is included in Z_G. To finish the proof we have to show that $\limsup_{\varepsilon\to 0^+} Z(\varepsilon) \subset \liminf_{\varepsilon\to 0^+} Z(\varepsilon)$ (the opposite inclusion is always satisfied).

Let $(x_0, \nu_0) \in \limsup_{\varepsilon\to 0^+} Z(\varepsilon)$. Then there are $(x_\varepsilon, y_\varepsilon) \in Z(\varepsilon), \varepsilon > 0$ such that (x_0, ν_0) is a cluster point of $\{(x_\varepsilon, y_\varepsilon)\}_{\varepsilon>0}$ in $\mathcal{T}$. Denote

$$\begin{aligned} w_j^\varepsilon(t) &= \int_0^t f_j(s, y_\varepsilon(s))\,ds, \\ w_j^0 &= \int_I \left(\int_K f_j(t, y)\nu_0(t)\,(dy)\right) dt, j = 1, 2, \ldots, \end{aligned} \tag{3.4}$$

where $f_j \in \mathcal{F}$. Recall that the set of Lipschitz functions $\mathcal{F} = \{f_j(\cdot, \cdot)\}_{j=1}^\infty$ is dense in $L^1(I, C(K, \mathbf{R}^m))$.

There exist subsequences $x_i = x_{\varepsilon_i}, y_i = y_{\varepsilon_i}, \varepsilon_i \to 0, i = 1, 2, \ldots$, such that

$$\begin{aligned} &x_i \to x_0 \text{ in } C(I, \mathbf{R}^n), \\ &w_j^i(1) = \int_I f_j(t, y_i(t))\,dt \to w_j^0 \text{ for } j = 1, 2, \ldots. \end{aligned} \tag{3.5}$$

We have that $(x_i, w_i, y_i) = (x_i, w_1^i, w_2^i, \dots, w_k^i, y_i)$ is a solution of the following system:

$$\begin{aligned} \dot{x}(t) &\in F(t, x(t), y(t), u(t)), \quad x(0) = x^0, \\ \dot{w}_1(t) &= f_1(t, y(t)), \quad w_1(0) = 0, \\ &\vdots \\ \dot{w}_k(t) &= f_k(t, y(t)), \quad w_k(0) = 0, \\ \varepsilon \dot{y}(t) &\in G(x(t), y(t), u(t)), \quad y(0) = y^0, \end{aligned} \tag{3.6}$$

where $\varepsilon = \varepsilon_i, i = 1, 2, \dots$. On the other hand, (x_i, w_i) has a cluster point $(x_0, \tilde{w}_0)$ in $C(I, \mathbf{R}^{n+k})$ due to Lemma 2.1 and the Arzelà–Ascoli theorem, and by (3.5) $w_i(1) \to \tilde{w}_0(1) = (w_1^0, w_2^0, \dots, w_k^0)$.

So $(x_0, \tilde{w}_0) \in \limsup_{\varepsilon \to 0^+} S(\varepsilon)$ where $S(\varepsilon)$ is the "slow" part of the solution set of (3.6). Since $f_j \in \mathcal{F}$, system (3.6) satisfies the conditions of Theorem 3.2, hence $(x_0, \tilde{w}_0) \in \liminf_{\varepsilon \to 0^+} S(\varepsilon)$. This means that for every $\varepsilon > 0$ there is a solution $(\tilde{x}_\varepsilon, \tilde{w}_\varepsilon, \tilde{y}_\varepsilon)$ of (3.6) such that $(\tilde{x}_\varepsilon, \tilde{w}_\varepsilon) \to (x_0, \tilde{w}_0)$ in $C(I, \mathbf{R}^{n+k})$. Consequently

$$\tilde{x}_\varepsilon \to x_0 \text{ in } C(I, \mathbf{R}^n), \tag{3.7}$$

$$\int_I f_j(t, \tilde{y}_\varepsilon(t))\, dt \to \int_I \left(\int_K f_j(t, y) \nu_0(t)\, (dy) \right) dt \text{ for } f_j \in \mathcal{F}, j = 1, 2, \dots, k. \tag{3.8}$$

The second relation follows by $\tilde{w}_\varepsilon(1) \to \tilde{w}_0(1) = (w_1^0, w_2^0, \dots, w_k^0)$ and (3.4).

Fix the natural number k and denote for every $\varepsilon > 0$

$$Z_\varepsilon^k = \{(\tilde{x}_\varepsilon, \tilde{y}_\varepsilon) \in Z(\varepsilon) : \text{ (3.7) and (3.8) are fulfilled}\}.$$

For every $0 < \varepsilon < 1$ we have that $Z_\varepsilon^{k+1} \subset Z_\varepsilon^k \neq \emptyset$ and Z_ε^k is $C(I, \mathbf{R}^{n+m})$ compact. Therefore $Z_\varepsilon^\infty = \bigcap_{k=1}^\infty Z_\varepsilon^k$ is not empty. Then if $(\tilde{x}_\varepsilon, \tilde{y}_\varepsilon) \in Z_\varepsilon^\infty, \varepsilon > 0$, relation (3.8) holds for any $f_j \in \mathcal{F}$; consequently $(\tilde{x}_\varepsilon, \tilde{y}_\varepsilon) \to (x_0, \nu_0)$ in $\mathcal{T}$. □

COROLLARY 3.2. *Consider the optimization problem*

$$\int_I h(t, x_\varepsilon(t), y_\varepsilon(t))\, dt \to \inf,$$

where the functional is minimized over the solutions of (1.3). *Suppose* $h(\cdot, x, y)$ *is measurable on* I, $h(t, \cdot, \cdot)$ *is continuous on* $\mathbf{R}^{n+m}$, *and* B1, B2 *are fulfilled. Denote by* $J(\varepsilon), \varepsilon > 0$, *the corresponding optimal value. Then* $\lim_{\varepsilon \to 0} J(\varepsilon)$ *exists.*

Proof. In system (3.6) we add the equation

$$\dot{w}_0(t) = h(t, x(t), y(t)), \quad w_0(0) = 0.$$

All the arguments in the proof of Theorem 3.3 will be valid having in mind that

$$\int_I h(t, x_\varepsilon(t), y_\varepsilon(t))\, dt \to \int_I \left(\int_K h(t, x_0(t), y) \nu_0(t)\, (dy) \right) dt, \varepsilon \to 0$$

when $(x_\varepsilon, y_\varepsilon) \to (y_0, \nu_0)$ in the considered topology (see [19]). □

Acknowledgments. We wish to thank the anonymous referees for their valuable comments and suggestions, which contributed much to improve the paper.

REFERENCES

[1] Z. Artstein, *Invariant measures of differential inclusions applied to singular perturbations*, J. Differential Equations, 152 (1999), pp. 289–307.

[2] Z. Artstein and A. Vigodner, *Singularly perturbed ordinary differential equations with dynamic limits*, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 541–569.

[3] K. Deimling, *Multivalued Differential Equations*, de Gruyter, Berlin, 1992.

[4] T. Donchev and I. Slavov, *Singularly perturbed functional differential inclusions*, Set Valued Anal., 3 (1995), pp. 113–128.

[5] A. Dontchev, T. Donchev, and I. Slavov, *A Tikhonov-type theorem for singularly perturbed differential inclusions*, Nonlinear Anal., 26 (1996), pp. 1547–1554.

[6] A. Dontchev and I. Slavov, *Upper semicontinuity of solutions of singularly perturbed differential inclusions*, in System Modelling and Optimization, H.-J. Sebastian and K. Tammer, eds., Lecture Notes in Control and Inform. Sci. 143, Springer, New York, 1991, pp. 273–280.

[7] A. V. Dontchev and V. M. Veliov, *Singular perturbation in Mayer's problem for linear systems*, SIAM J. Control Optim., 21 (1983), pp. 566–581.

[8] V. Gaitsgory, *Use of averaging method in control problems*, Differential Equations, 22 (1986), pp. 1290–1299 (English translation).

[9] V. Gaitsgory, *Control of Systems with Fast and Slow Motions*, Nauka, Moscow, 1991 (in Russian).

[10] V. Gaitsgory, *Suboptimization of singularly perturbed control systems*, SIAM J. Control Optim., 30 (1992), pp. 1228–1249.

[11] G. Grammel, *Singularly perturbed differential inclusions: An averaging approach*, Set-Valued Anal., 4 (1996), pp. 361–374.

[12] M. Happaev and O. Filatov, *Averaging of differential inclusions with slow and fast Variables*, Math. Zametky, 47 (1990), pp. 102–109.

[13] V. V. Nemytskii and V. V. Stepanov, *Qualitative Theory of Differential Equations*, Princeton University Press, Princeton, NJ, 1960.

[14] V. Plotnikov, *Averaging Methods in Control Problems*, Libid Kiev, Odessa, 1992 (in Russian).

[15] A. Tikhonov, *Systems of differential equations containing a small parameter in the derivatives*, Mat. Sb., 31(73) (1952), pp. 575–586 (in Russian).

[16] V. Veliov, *Differential inclusions with stable subinclusions*, Nonlinear Anal. TMA, 23 (1994), pp. 1027–1038.

[17] V. Veliov, *A generalization of the Tikhonov theorem for singularly perturbed differential inclusions*, J. Dynam. Control Systems, 3 (1997), pp. 291–319.

[18] A. Vigodner, *Limits of singularly perturbed control problems with statistical dynamics of fast motions*, SIAM J. Control Optim., 35 (1997), pp. 1–28.

[19] R. Warga, *Optimal Control of Differential and Functional Differential Equations*, Academic Press, New York, 1973.

SIAM J. Control Optim.
Vol. 37, No. 5, pp. 1614–1647

FINITE-DIMENSIONAL COMPENSATORS FOR THE H^∞-OPTIMAL CONTROL OF INFINITE-DIMENSIONAL SYSTEMS VIA A GALERKIN-TYPE APPROXIMATION*

MINGQING XIAO† AND TAMER BAŞAR‡

Abstract. We study the existence of general finite-dimensional compensators in connection with the H^∞-optimal control of linear time-invariant systems on a Hilbert space with noisy output feedback. The approach adopted uses a Galerkin-type approximation, where there is no requirement for the system operator to have a complete set of eigenvectors. We show that if there exists an infinite-dimensional compensator delivering a specific level of attenuation, then a finite-dimensional compensator exists and achieves the same level of disturbance attenuation. In this connection, we provide a complete analysis of the approximation of infinite-dimensional generalized Riccati equations by a sequence of finite-dimensional Riccati equations. As an illustration of the theory developed here, we provide a general procedure for constructing finite-dimensional compensators for robust control of flexible structures.

Key words. H^∞-optimal control, finite-dimensional compensators, infinite-dimensional systems, Riccati equations, flexible structures

AMS subject classifications. 49J20, 49N10, 93C20

PII. S0363012998333505

1. Introduction.

Let X, U, W be real, separable Hilbert spaces, and let $\mathcal{L}(X;U)$, $\mathcal{L}(X;W)$ be the class of linear bounded operators from, respectively, X to U and X to W. Consider the uncertain evolution equation on X:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) + Dw(t), \\ x(0) &= x_0, \end{aligned} \tag{1.1}$$

where $B \in \mathcal{L}(X;U)$, $D \in \mathcal{L}(X;W)$; u is the control and w is an unknown deterministic disturbance, with $u(t) \in U$ and $w(t) \in W$. The operator A is the infinitesimal generator of a strongly continuous semigroup $T(t)$ on X, which we will henceforth refer to as the structure operator for system (1.1). The *partial observation* (also called noise-corrupted measurement) is given by

$$y(t) = Cx(t) + \eta(t), \tag{1.2}$$

where $C \in \mathcal{L}(X;Y)$, Y being a Hilbert space, called the space of measurements, and $\eta \in W_1$ being another disturbance, modeling the measurement error. The controller is allowed to be only a causal function of the observation y. For symmetry purposes, we write the performance index in terms of a second output

$$y_1(t) = Hx(t) + D_{12}u(t), \tag{1.3}$$

[*] Received by the editors February 2, 1998; accepted for publication (in revised form) October 1, 1998; published electronically August 10, 1999. Research supported in part by the U.S. Department of Energy under grant DOE-DEFG-02-94ER1 3939, in part by the National Science Foundation under grant ECS-93-12807, and in part by a University of Illinois fellowship.

http://www.siam.org/journals/sicon/37-5/33350.html

[†] Department of Mathematics, University of California, Davis, CA 95616 (xiao@math.ucdavis.edu).

[‡] Coordinated Science Laboratory, University of Illinois, 1308 West Main Street, Urbana, IL 61801 (tbasar@decision.csl.uiuc.edu).

where $y_1 \in Y_1$, with Y_1 a real Hilbert space, $H \in \mathcal{L}(X;Y_1)$, and $D_{12} \in \mathcal{L}(U;Y_1)$ satisfying the following standard hypotheses:

$$D_{12}^* D_{12} = I, \qquad D_{12}^* H = 0. \tag{1.4}$$

Let the cost function corresponding to the output y_1 be

$$K_{x_0}(u,w) = \int_0^\infty \|y_1(t)\|_{Y_1}^2 dt \equiv \int_0^\infty (\|Hx(t)\|_{Y_1}^2 + \|u(t)\|_U^2) dt. \tag{1.5}$$

A natural class of infinite-dimensional compensators for (1.1) and (1.2), characterized by three maps $L \in \mathcal{L}(X;U)$, $M \in \mathcal{L}(X;X)$, $G \in \mathcal{L}(Y;X)$, is the following one:

$$\begin{aligned} u(t) &= Lz(t), \\ \dot{z}(t) &= (A+M)z(t) + Gy(t). \end{aligned} \tag{1.6}$$

Here we do not lump A and M into a single operator, because A is unbounded while M is bounded. Let $w \in L^2(0,\infty;W)$ and $\eta \in L^2(0,\infty;W_1)$, where $L^2(0,\infty;W)$ is the Hilbert space of Lebesgue square integrable functions on $(0,\infty)$, with values in W. The H^∞-optimal design problem for (1.1), under (1.2) and (1.6), and with the cost function (1.5) is the following: Given $\gamma > 0$, find L, M, and G such that

$$\sup_{w,\eta} \frac{K_0(u,w)}{\displaystyle\int_0^\infty (\|w(t)\|_W^2 + \|\eta(t)\|_{W_1}^2) dt} < \gamma^2. \tag{1.7}$$

More precisely, we have the following definition.

Definition 1.1. *The γ-attenuation level is attained for system* (1.1), *with observation* (1.2) *and cost function* (1.5), *if there exist L, M, and G in* (1.6) *such that* (1.7) *holds.*

It is well known [1] that in the finite-dimensional case, and when the dimension of the compensator is the same as the dimension of the state (i.e., $\dim(z) = \dim(x)$), under some stabilizability and detectability conditions an explicit solution can be obtained for the triple L, M, G, in terms of two Riccati equations coupled through a spectral radius condition. This result was subsequently generalized to infinite-dimensional spaces by Bensoussan and Bernhard [3]. Before stating this extension, let us introduce the two associated infinite-dimensional Riccati equations:

$$\begin{aligned} &\Pi A + A^*\Pi - \Pi\left(BB^* - \frac{1}{\gamma^2}DD^*\right)\Pi + H^*H = 0, \\ &\Pi = \Pi^* \geq 0, \\ &A - \left(BB^* - \frac{1}{\gamma^2}DD^*\right)\Pi \text{ exponentially stable,} \end{aligned} \tag{1.8}$$

and

$$\begin{aligned} &\Sigma A^* + A\Sigma - \Sigma\left(C^*C - \frac{1}{\gamma^2}H^*H\right)\Sigma + DD^* = 0, \\ &\Sigma = \Sigma^* \geq 0, \\ &A^* - \left(C^*C - \frac{1}{\gamma^2}H^*H\right)\Sigma \text{ exponentially stable.} \end{aligned} \tag{1.9}$$

Then we have [3] as follows.

THEOREM 1.1 (Bensoussan and Bernhard [3]). *Assume that the pair* (A, B) *is stabilizable and the pair* (A, H) *is detectable. Then the* γ*-attenuation level is attained for system* (1.1), *with observation* (1.2) *and under cost function* $K_0(u, w)$*, iff the Riccati equation* (1.8) *admits a solution* Π*, and* (1.9) *admits a solution* Σ*, and furthermore*

$$(1.10) \qquad I - \frac{1}{\gamma^2}\Pi\Sigma \text{ is invertible, or equivalently } \Sigma\left(I - \frac{1}{\gamma^2}\Pi\Sigma\right)^{-1} \geq 0.$$

An infinite-dimensional compensator that achieves the bound γ^2 *in* (1.7) *is given by* (1.6) *where*

$$(1.11) \qquad L = -B^*\Pi, \quad M = -\left(BB^* - \frac{1}{\gamma^2}DD^*\right)\Pi - \Gamma C^* C, \quad G = \Gamma C^*$$

and

$$\Gamma = \Sigma\left(I - \frac{1}{\gamma^2}\Pi\Sigma\right)^{-1}.$$

Even though this is a complete solution for the problem at hand, the fact that the compensator is infinite dimensional renders it impractical. There is therefore a need to develop a theory that will deliver finite-dimensional stabilizing compensators, which achieve a given level of disturbance attenuation. We will address this problem here when the measurement y is finite-dimensional.

The existence of finite-dimensional compensators for infinite-dimensional linear quadratic regulator (LQR) problems or for linear quadratic Gaussian (LQG) problems has been established in [8], [25], and [26] provided that the structure operator of the underlying system has a complete set of generalized eigenvectors; their construction is based on the eigenvectors of either the structure operator or the closed-loop structure operator obtained under state feedback. Gibson [12], [13], Gibson and Adamian [14], and Ito [15] have established the existence of finite-dimensional compensators for LQR problems by using Galerkin-type approximations in which independent basis elements were used instead of the complete set of eigenvectors, making it possible for the results to be applied to general parabolic systems, hereditary differential systems, and flexible structures.

The existence and design of general finite-dimensional compensators for infinite-dimensional systems in an H^∞ framework was one of the open problems in H^∞-optimal control, as posed by Curtain and Salamon in 1990 [7]. Özbay [20], Özbay and Tannenbaum [21], [22], and Curtain and Zhou [10] have obtained a number of results on finite-dimensional H^∞ controllers by using frequency domain techniques under appropriate assumptions. Banks, Demetriou, and Smith [2] studied a two-dimensional structural acoustic model by using H^∞ periodic control, which was approximated by a finite-dimensional compensator: piezoceramic actuators. Ito and Morris [16] developed an approximation scheme when the *full output measurement* is available. However, the problem of finding finite-dimensional H^∞ controllers from state-space approximations in complete generality was an unsolved problem. Thus one of the objectives of this paper is to fill this gap. In the present paper, we establish a general existence result for finite-dimensional H^∞-optimal compensators (1.6) for (1.1) with finite-dimensional measurement (1.2), such that the γ-attenuation level (1.7) is achieved. More precisely, we study the evolution system

$$\dot{x}(t) = Ax(t) + Bu(t) + Dw(t), \qquad x(0) = x_0 \in X \tag{1.12}$$

along with the partial observation

$$y(t) = Cx(t) + \eta(t) \tag{1.13}$$

with finite-dimensional compensator

$$\begin{aligned} u(t) &= L_c z(t), \\ \dot{z}(t) &= (A_c + M_c)z(t) + G_c y(t), \end{aligned} \tag{1.14}$$

where $u(t)$ is an $\mathbb{R}^m$-valued control function, the measurement space is $\mathbb{R}^p$, and $C \in \mathcal{L}(X, \mathbb{R}^p)$, $z(t) \in Z = \mathbb{R}^{n_c}$, and A_c, M_c, G_c, L_c are matrices belonging to appropriate spaces. The cost function is still expressed as

$$K_{x_0}(u, w) = \int_0^\infty (\|Hx(t)\|_{Y_1}^2 + \|u(t)\|_{\mathbb{R}^m}^2)dt. \tag{1.15}$$

We will show that there exists a finite-dimensional compensator (1.14) which stabilizes the infinite-dimensional system (1.12) with noise measurement output (1.13), and a γ-attenuation level is achieved, i.e.,

$$\sup_{w,\eta} \frac{K_0(u,w)}{\displaystyle\int_0^\infty (\|w(t)\|_W^2 + \|\eta(t)\|_{W_1}^2)dt} < \gamma^2 \tag{1.16}$$

provided that this level of attenuation level is achieved under the infinite-dimensional compensator (1.6).

The rest of this paper is organized as follows. Section 2 provides assumptions and some preliminary results on asymptotic behavior and approximation of Riccati equations, proofs of which have been included in Appendix II. Results obtained in Appendix II can be viewed as independent from the main body of the paper and can be read as such. The existence of finite-dimensional compensators for the infinite-dimensional problem is proven in section 3. Section 4 presents an important application of this theory to the control of flexible structures. Section 5 includes some concluding remarks, and Appendix I contains a duality theorem which is needed in section 3.

2. Assumptions and some preliminaries.

Let $\{X^N\}$ be a sequence of finite-dimensional subspaces of X, and let P^N denote the orthogonal projection[1] of X onto X^N. Throughout this paper, the following notation and conventions will be adopted, unless otherwise indicated: $\langle \cdot, \cdot \rangle$ denotes the inner-product on X, $\mathcal{L}(X)$ denotes the space of all bounded linear operators on X, and A^* denotes the adjoint of the linear operator A,

$$\Sigma(X) = \{T \in \mathcal{L}(X) : \ T \text{ is Hermitian}\}$$

$$\Sigma^+(X) = \{T \in \Sigma(X) : \quad \langle Tx, x \rangle \ \geq 0 \ \forall x \in X\}.$$

[1] Orthogonal projection means that the range and the null spaces of the projection are orthogonal.

We next introduce a sequence of operators, which will be used in the main body of the paper. Let $A^N \in \mathcal{L}(X^N)$, $B^N = P^N B \in \mathcal{L}(U, X^N)$, $D^N = P^N D \in \mathcal{L}(W, X^N)$, and C^N and H^N represent the restrictions of C and H onto X^N, respectively, and $\Pi^N, \Sigma^N \in \mathcal{L}(X^N)$.

Definition 2.1. *A sequence $\{x_N \in X^N\}$ converges to $x \in X$ if*

$$\|P^N x - x_N\|_{X^N} \to 0 \textit{ as } N \to \infty.$$

Definition 2.2. *Let X, Y be two Banach spaces. A sequence of linear operators $A^N : X^N \to Y^N$ converges to an operator $A : D(A) \subset X \to Y$ if*

$$\mathcal{D} = \{x : P^N x \in D(A^N), A^N P^N x \textit{ converges in } Y^N\}$$

and

$$Ax = \lim_{N\to\infty} A^N P^N x \quad \textit{for} \quad x \in \mathcal{D} \cap D(A).$$

We will denote this type of convergence by $A^N \to\!\!\to A$.

Definition 2.3. *We say that the Riccati equation*

$$\Pi A + A^*\Pi - \Pi\left(BB^* - \frac{1}{\gamma^2}DD^*\right)\Pi + H^*H = 0$$

admits a solution Π, if $\Pi \in \Sigma^+(X)$ and $\forall x, y \in D(A)$ the following holds:

$$\langle \Pi x, Ay\rangle + \langle Ax, \Pi y\rangle - \left\langle \Pi\left(BB^* - \frac{1}{\gamma^2}DD^*\right)\Pi x, y\right\rangle + \langle H^*Hx, y\rangle = 0.$$

A similar definition applies to the Riccati equation

$$\Sigma A^* + A\Sigma - \Sigma\left(C^*C - \frac{1}{\gamma^2}H^*H\right)\Sigma + DD^* = 0.$$

Definition 2.4. *The pair (A, B) in* (1.1) *is stabilizable if there exists an operator $F \in \mathcal{L}(X, U)$ such that $A + BF$ is exponentially stable.*

Definition 2.5. *The pair (A, H) (A in* (1.1), *H in* (1.15)*) is detectable if there exists an operator $G \in \mathcal{L}(Y_1, X)$ such that $A + GH$ is exponentially stable.*

Assumptions.

(A1) For every $x \in X$, $e^{A^N t}P^N x$ converges strongly to $T(t)x$, $e^{A^{N*} t}P^N x$ converges strongly to $T^*(t)x$, and the convergence is uniform in t on bounded intervals of $[0, \infty[$. Here $T(t)$ is the C_0 semigroup generated by A, and P^N is the orthogonal projection of X onto a finite-dimensional space X^N.

(A2) For each integer $N > 0$, $A^N - (B^N B^{N*} - \frac{1}{\gamma^2}D^N D^{N*})\Pi^N$ generates an exponentially stable semigroup on X^N, and

$$\|e^{(A^N - (B^N B^{N*} - \frac{1}{\gamma^2}D^N D^{N*})\Pi^N)t}P^N\|_{\mathcal{L}(X,X^N)} \le M_1 e^{-\omega_1 t}, \qquad t \in [0, \infty[,$$

for some $M_1 \ge 1$ and $\omega_1 > 0$, independent of N.

(A3) For each integer $N > 0$, $A^{N*} - (C^{N*}C^N - \frac{1}{\gamma^2}H^{N*}H^N)\Sigma^N$ generates an exponentially stable C_0 semigroup on X^N, and

$$\|e^{(A^{N*} - (C^{N*}C^N - \frac{1}{\gamma^2}H^{N*}H^N)\Sigma^N)t}P^N\|_{\mathcal{L}(X,X^N)} \le M_2 e^{-\omega_2 t}, \qquad t \in [0, \infty[,$$

for some $M_2 \ge 1$ and $\omega_2 > 0$, independent of N.

(A4) For each $N > 0$, the operator matrix

$$\mathcal{A}_p^N = \begin{bmatrix} A + \frac{1}{\gamma^2} DD^*\Pi & BL^N \\ G^N C & A^N + M^N \end{bmatrix}$$

satisfies the spectrum-determined growth condition,[2] where

$$L^N = -B^{N*}\Pi^N, \quad G^N = \Gamma^N C^{N^*}, \quad \Gamma^N = \Sigma^N \left(I - \frac{1}{\gamma^2} \Pi^N \Sigma^N \right)^{-1}$$

and

$$M^N = -\left(B^N B^{N^*} - \frac{1}{\gamma^2} D^N D^{N*} \right) \Pi^N - \Gamma^N C^{N*} C^N.$$

Remark 1. (1) In (A1), $e^{A^N t} P^N x$ converges strongly to $T(t)x$ means that

$$\| P^N T(t) x - e^{A^N t} P^N x \|_{X^N} \to 0 \quad \text{as } N \to \infty.$$

For $t = 0$, assumption (A1) implies that $P^N x \to x$ for each $x \in X$; thus we have the result that the subspace X^N approximates X.

(2) According to (A1) we have for each $u \in U$, $B^N u \to Bu$ and for each $x \in X$, $C^N x \to Cx$ as $N \to \infty$. Moreover, since both B and C are compact operators, principle of uniform boundedness [27] implies that

$$B^N \to B, \qquad C^N \to C \qquad \text{as} \quad N \to \infty$$

where the convergences are in the corresponding operator norms.

(3) From assumption (A2), we know that the resolvent set of matrix

$$A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*} \right) \Pi^N$$

contains the ray (ω_1, ∞) and

$$\left\| \left[\lambda I - \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*} \right) \Pi^N \right) \right]^{-1} \right\|_{\mathcal{L}(X^N)} \leq M_1 / (\lambda + \omega_1)$$
$$\text{for } \mathrm{Re}\lambda > -\omega_1.$$

A similar corresponding statement follows from assumption (A3).

The proof of the following two lemmas can be found in Appendix II.

LEMMA 2.1. *Suppose that for a given $\gamma > 0$ the Riccati equation*

$$\Pi A + A^* \Pi - \Pi \left(BB^* - \frac{1}{\gamma^2} DD^* \right) \Pi + H^* H = 0 \tag{2.1}$$

[2] Spectrum-determined growth condition means that

$$\sup\{ \mathrm{Re}\lambda : \lambda \in \sigma(\mathcal{A}_p) \} = \lim_{t \to \infty} \frac{1}{t} \ln \| e^{\mathcal{A}_p t} \|,$$

where $\sigma(A)$ denotes the spectrum of A.

admits a solution $\Pi \in \Sigma^+(X)$ *such that* $A - BB^* - \frac{1}{\gamma^2}DD^*\Pi$ *is exponentially stable, and for each* N

$$\Pi^N A^N + A^{N*}\Pi^N - \Pi^N \left(B^N B^{N*} - \frac{1}{\gamma^2}D^N D^{N*}\right)\Pi^N + H^{N*}H^N = 0 \tag{2.2}$$

admits a solution $\Pi^N \in \Sigma^+(X^N)$ *which satisfies assumption* (A2). *Then, under assumption* (A1), *we have* $\Pi^N \to\!\!\to \Pi$ *as* $N \to \infty$.

Lemma 2.2. *Suppose that for a given* $\gamma > 0$, *the Riccati equation*

$$\Sigma A^* + A\Sigma - \Sigma\left(C^*C - \frac{1}{\gamma^2}H^*H\right)\Sigma + DD^* = 0 \tag{2.3}$$

admits a solution $\Sigma \in \Sigma^+(X)$ *such that* $A^* - (C^*C - \frac{1}{\gamma^2}H^*H)\Sigma$ *is exponentially stable, and*

$$\Sigma^N A^{N*} + A^N\Sigma^N - \Sigma^N\left(C^{N*}C^N - \frac{1}{\gamma^2}H^{N*}H^N\right)\Sigma^N + D^N D^{N*} = 0 \tag{2.4}$$

admits a solution $\Sigma^N \in \Sigma^+(X)$ *which satisfies assumption* (A3). *Then, under assumption* (A1), *we have* $\Sigma^N \to\!\!\to \Sigma$ *as* $N \to \infty$.

The following lemma now provides some properties of Riccati equations provided that for the given $\gamma > 0$, the γ-attenuation level is achieved. The proof can be found in [3].

Lemma 2.3. *Assume that* (A, B) *is stabilizable and that* (A, H) *is detectable, and the* γ*-attenuation level holds for system* (1.12) *with measurement* (1.13) *and cost function* K_0. *Then, we have*

(1) $A - BB^*\Pi$ *is exponentially stable, where* Π *is a solution of* (1.8);

(2) $A^* + \frac{1}{\gamma^2}\Pi DD^* - (C^*C - \frac{1}{\gamma^2}\Pi BB^*\Pi)\Gamma$ *is exponentially stable, where*

$$\Gamma = \Sigma\left(I - \frac{1}{\gamma^2}\Pi\Sigma\right)^{-1}$$

and Σ *is the solution of* (1.9);

(3) $A^* + \frac{1}{\gamma^2}\Pi DD^* - C^*C\Gamma$ *is exponentially stable;*

(4) Γ *is the unique solution of the following Riccati equation*

$$\Gamma\left(A^* + \frac{1}{\gamma^2}\Pi DD^*\right) + \left(A + \frac{1}{\gamma^2}DD^*\Pi\right)\Gamma - \Gamma\left(C^*C - \frac{1}{\gamma^2}\Pi BB^*\Pi\right)\Gamma + DD^* = 0; \tag{2.5}$$

(5) *the system*

$$\begin{aligned} \dot{x} &= \left(A^* + \frac{1}{\gamma^2}\Pi DD^*\right)x + C^*v + \Pi B\mu, \qquad x(0) = 0, \\ z &= D^*x \end{aligned} \tag{2.6}$$

with cost function

$$K_0(v,\mu) = \int_0^\infty (\|z\|_W^2 + \|v\|_Y^2)dt \tag{2.7}$$

achieves the γ-attenuation level; i.e., there exists $\hat{v} \in Y$ such that

$$\sup_{\mu} \frac{\int_0^\infty (\|z\|^2 + \|\hat{v}\|^2)dt}{\int_0^\infty \|\mu\|^2 dt} < \gamma^2. \tag{2.8}$$

3. Main existence result and its proof.

MAIN THEOREM. *Consider the uncertain evolution system* (1.1) *along with finite-dimensional measurement* (1.13) *and cost function K_{x_0} given by* (1.15). *Assume that (A, B) is stabilizable and (A, H) is detectable, and there exists an infinite-dimensional compensator for* (1.12) *such that the γ-attenuation level is achieved with the measurement* (1.13). *Then, under assumptions* (A1)–(A4), *there exists a finite-dimensional compensator-based controller*

$$\begin{aligned} u(t) &= L^N z(t), \\ \dot{z}(t) &= (A^N + M^N)z(t) + G^N y(t) \end{aligned}$$

which stabilizes the uncertain system (1.12), *and the γ-attenuation level is achieved under the same measurement* (1.13) *and cost function* (1.15).

We prove the result using a sequence of lemmas and theorems. First let us introduce

$$\mathcal{A}^N := \begin{bmatrix} A & BL^N \\ G^N C & A^N + M^N \end{bmatrix}$$

where

$$L^N = -B^{N*}\Pi^N, \quad G^N = \Gamma^N C^{N^*}, \quad \Gamma^N = \Sigma^N \left(I - \frac{1}{\gamma^2}\Pi^N\Sigma^N \right)^{-1}$$

and

$$M^N = -\left(B^N B^{N^*} - \frac{1}{\gamma^2} D^N D^{N*} \right) \Pi^N - \Gamma^N C^{N*} C^N.$$

One goal here is to show that there exists a sufficiently large N such that $\mathcal{A}^N$ is exponentially stable on $X \times X^N$.

LEMMA 3.1. *Assume that assumptions* (A1) *and* (A2) *are satisfied. Then for each N, $L^N = -B^{N*}\Pi^N \in \mathcal{L}(X^N, \mathbb{R}^m)$ and $L^N P^N$ converges $L = -B^*\Pi$ in its operator norm as $N \to \infty$. Furthermore, there exists an integer N_1 such that if $N \geq N_1$ and $Re\lambda > -\omega_3$*

$$\left(\lambda I - \left(A + \frac{1}{\gamma^2} DD^*\Pi + BL^N P^N \right) \right)^{-1} \in \mathcal{L}(X)$$

and

$$e^{(\lambda I - (A + \frac{1}{\gamma^2} DD^*\Pi + BL^N P^N))t} \to e^{(\lambda I - (A + \frac{1}{\gamma^2} DD^*\Pi + BL))t} \quad \textit{as} \quad N \to \infty$$

where the convergence is uniform on bounded t-intervals.

Proof. Since B is a compact operator, so is B^*. By assumption (A1) we know that $B^{N*} \to B^*$ in norm, and by Lemma 2.1 we have $\Pi^N \to\to \Pi$. Thus the first conclusion follows. For the second assertion, let $S^N(t)$ be the C_0 semigroup generated by the operator $A+\frac{1}{\gamma^2}DD^*\Pi+BL^NP^N$, and let $S(t)$ be the C_0 semigroup generated by the operator $A+\frac{1}{\gamma^2}DD^*\Pi+BL$. Since

$$A+\frac{1}{\gamma^2}DD^*\Pi+BL^NP^N=A+\frac{1}{\gamma^2}DD^*\Pi+BL+B(L^NP^N-L),$$

in view of the "Perturbation Theorem" of [9], we have

$$S^N(t)=S(t)+\int_0^t S(t-s)B(L^NP^N-L)S^N(s)ds. \tag{3.1}$$

Since $A+\frac{1}{\gamma^2}DD^*\Pi+BL$ is exponentially stable, there exist $M_3\geq 1$ and $\omega\geq 0$ such that

$$\|S(t)\|\leq M_3e^{-\omega t}.$$

Thus we have

$$\|S^N(t)\|\leq M_3e^{-\omega t}+\int_0^t M_3e^{-\omega(t-s)}\|B\|\|L-L^NP^N\|\|S^N(s)\|ds, \tag{3.2}$$

and Gronwall's lemma implies that

$$\|S^N(t)\|\leq M_3e^{(-\omega+M_3\|B\|\|L-L^NP^N\|)t},\qquad t\geq 0. \tag{3.3}$$

Since $L^NP^N\to L$, by the Hille–Yosida theorem (see Theorem 1.5.3 of [24]) we know that there exists $N_1>0$ such that when $N>N_1$, $(\lambda I-(A+\frac{1}{\gamma^2}DD^*\Pi+BL^N))^{-1}$ is bounded in $\mathcal{L}(X)$. This yields the second statement. To conclude the proof, note that

$$\begin{aligned}&\left(\lambda I-\left(A+\frac{1}{\gamma^2}DD^*\Pi+BL^NP^N\right)\right)^{-1}-\left(\lambda I-\left(A+\frac{1}{\gamma^2}DD^*\Pi+BL\right)\right)^{-1}\\&\quad=\left(\lambda I-\left(A+\frac{1}{\gamma^2}DD^*\Pi+BL^NP^N\right)\right)^{-1}\\&\qquad\times B(L^NP^N-L)\left(\lambda I-\left(A+\frac{1}{\gamma^2}DD^*\Pi+BL\right)\right)^{-1}\end{aligned}$$

and $L^N\to\to L$, which lead to

$$\left(\lambda I-\left(A+\frac{1}{\gamma^2}DD^*\Pi+BL^NP^N\right)\right)^{-1}\to\left(\lambda I-\left(A+\frac{1}{\gamma^2}DD^*\Pi+BL\right)\right)^{-1}\quad\text{as } N\to\infty.$$

By the Trotter–Kato theorem (see Theorem 3.4.4 of [24]), the proof of Lemma 3.1 is complete. □

Remark 2. Clearly, from (3.3) there exist $\omega_3\geq 0$ and $N_3>0$ such that when $N>N_3$, we have

$$\|S^N(t)\|\leq M_3e^{-\omega_3 t},\qquad t\geq 0,$$

where both ω_3 and M_3 are independent of N.

Lemma 3.2. *Let A^N, B^N, C^N, D^N, and Π^N satisfy* (A1), (A2), *and* (A3) *for each $N > 0$. Then for every $x \in X$,*

$$e^{(A^N - (B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*})\Pi^N)t} P^N x \to T_1(t)x$$

uniformly on bounded intervals with respect to t, where $T_1(t)$ is the C_0 semigroup generated by $A - (BB^* - \frac{1}{\gamma^2} DD^*)\Pi$.

Proof. According to the Trotter–Kato theorem and Remark 1, part 3, we need to show that

$$\begin{aligned}&\left[\lambda I - \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N\right)\right]^{-1} P^N x\\ &\qquad \to \left[\lambda I - \left(A - \left(BB^* - \frac{1}{\gamma^2} DD^*\right)\Pi\right)\right]^{-1} x\end{aligned}$$

holds for any $x \in X$. Since

$$\begin{aligned}&\left[I - \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N\right)\right]^{-1} P^N - (\lambda I - A^N)^{-1}\\ &\quad = -\left[I - \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N\right)\right]^{-1}\\ &\qquad \times \left[\left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N\right](\lambda I - A^N)^{-1}\end{aligned}$$

and

$$\begin{aligned}&\left[\lambda I - \left(A - \left(BB^* - \frac{1}{\gamma^2} DD^*\right)\Pi\right)\right]^{-1} - (\lambda I - A)^{-1}\\ &\quad = -\left[\lambda I - \left(A - \left(BB^* - \frac{1}{\gamma^2} DD^*\right)\Pi\right)\right]^{-1}\left[\left(BB^* - \frac{1}{\gamma^2} DD^*\right)\Pi\right](\lambda I - A)^{-1},\end{aligned}$$

it can then be verified that

$$\begin{aligned}&\left[\lambda I - \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N\right)\right]^{-1} P^N\\ &\quad - \left[\lambda I - \left(A - \left(BB^* - \frac{1}{\gamma^2} DD^*\right)\Pi\right)\right]^{-1}\\ &= \Bigg\{\Bigg[I - \left[\lambda I - \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N\right)\right]^{-1}\\ &\quad \times \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N P^N\Bigg]\left[(\lambda I - A^N)^{-1} P^N - (\lambda I - A)^{-1}\right]\\ &\quad + \left[\lambda I - \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N\right)\right]^{-1}\\ &\quad \times \left[\left(BB^* - \frac{1}{\gamma^2} DD^*\right)\Pi - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N P^N\right](\lambda I - A)^{-1}\Bigg\}(\lambda I - A)\\ &\quad \times \left[\lambda I - \left(A - \left(BB^* - \frac{1}{\gamma^2} DD^*\right)\Pi\right)\right]^{-1}.\end{aligned}$$

According to assumptions (A1), (A2), and Lemma 2.1, we have

$$(\lambda I - A^N)^{-1} \to\to (\lambda I - A)^{-1}, \ \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right) \Pi^N \to\to \left(BB^* - \frac{1}{\gamma^2} DD^*\right) \Pi$$

as $N \to \infty$. Note that

$$\left[\lambda I - \left(A - \left(BB^* - \frac{1}{\gamma^2} DD^*\right) \Pi\right)\right]^{-1} : X \to D(A),$$

and hence the closed graph theorem implies that the operator

$$(\lambda I - A) \left[\lambda I - \left(A - \left(BB^* - \frac{1}{\gamma^2} DD^*\right) \Pi\right)\right]^{-1}$$

is bounded. Thus we have the desired result. □

THEOREM 3.3. *Let* A^N, B^N, C^N, D^N, *and* Π^N *satisfy* (A1) *and* (A2), *and introduce*

$$\mathcal{A}_p^N := \begin{bmatrix} A + \frac{1}{\gamma^2} DD^* \Pi & BL^N \\ G^N C & A^N + M^N \end{bmatrix},$$

where Π *is the solution of* (1.8). *Then, for sufficiently large* N, *there exists* $\epsilon > 0$ *which is independent of* N *such that*

$$\sigma(\mathcal{A}_p^N) \subseteq (-\infty, -\epsilon]$$

where $\sigma(\mathcal{A}_p)$ *is the spectrum*[3] *of* $\mathcal{A}_p$.

Proof. For a given $(h_1, h_2) \in X \times X^N$, consider the equation

$$(\lambda I - \mathcal{A}_p^N) \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}, \tag{3.4}$$

which is equivalent to

$$\lambda v_1 - \left(A v_1 + \frac{1}{\gamma^2} DD^* \Pi v_1 + BL^N v_2\right) = h_1 \tag{3.5}$$

and

$$\lambda v_2 - [G^N C v_1 + (A^N + M^N) v_2] = h_2. \tag{3.6}$$

What we now show is that for any given $(h_1, h_2) \in X \times X^N$, when N is sufficiently large, (3.5)–(3.6) admits a solution (v_1, v_2) which is also in $X \times X^N$.

According to Lemma 3.1, there exists an integer $N_1 > 0$ such that when $\mathrm{Re}\lambda > -\omega_3$, $(\lambda I - (A + \frac{1}{\gamma^2} DD^* \Pi + BL^N))^{-1}$ is well defined. Hence, (3.5) can be written as

$$v_1 = \left[\lambda I - \left(A + \frac{1}{\gamma^2} DD^* \Pi + BL^N\right)\right]^{-1} [BL^N (v_2 - v_1) + h_1]. \tag{3.7}$$

[3] The spectrum of an operator consists of three mutually exclusive parts: the point spectrum, the continuous spectrum, and the residual spectrum; for more details, see Kato [18].

Similarly, since $A^N - (B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*})\Pi^N$ generates an exponentially stable semigroup by the hypothesis of the theorem and in view of Lemma 2.1, there exists λ, with $\mathrm{Re}\lambda > -\omega_1$, for which (3.6) becomes

$$(3.8)\quad v_2 = \left[\lambda I - \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\right)\Pi^N\right]^{-1} [G^N C(v_1 - v_2) + h_2].$$

Thus, if $\mathrm{Re}\lambda > -\min(\omega_1, \omega_3)$ and $N > N_1$ we have

$$\begin{aligned} v_1 - v_2 &= \left[\lambda I - \left(A + \frac{1}{\gamma^2} DD^*\Pi + BL^N\right)\right]^{-1} [BL^N(v_2 - v_1) + h_1] \\ &\quad - \left[\lambda I - \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\right)\Pi^N\right]^{-1} [G^N C(v_1 - v_2) + h_2]. \end{aligned}$$

We next show that the above equation admits a unique solution. By some manipulations, we arrive at

$$\begin{aligned} (3.9)\qquad & \left[\lambda I - \left(A^N + \frac{1}{\gamma^2} D^N D^{N*}\Pi^N - G^N C^N\right)\right](v_1 - v_2) \\ &= \left[\lambda I - \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N\right)\right](v_1 - v_2) + (B^N L^N + G^N C^N)(v_1 - v_2) \\ &= \left[\lambda I - \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N\right)\right] \\ &\quad \times \left(-\left[\lambda I - \left(A + \frac{1}{\gamma^2} DD^*\Pi + BL^N\right)\right]^{-1} BL^N\right. \\ &\quad\quad \left. + \left[\lambda I - \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N\right)\right]^{-1} B^N L^N\right)(v_1 - v_2) \\ &+ \left[\lambda I - \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N\right)\right]\left(\left[\lambda I - \left(A + \frac{1}{\gamma^2} DD^*\Pi + BL^N\right)\right]^{-1} h_1\right. \\ &\left. - \left[\lambda I - \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N\right)\right]^{-1} P^N h_1\right) \\ &+ P^N h_1 - h_2 + G^N(C^N - C)(v_1 - v_2). \end{aligned}$$

By Lemma 2.3, part 3, for all λ such that $\mathrm{Re}\lambda > -\omega_2$, the matrix $[\lambda I - (A^N + \frac{1}{\gamma^2} D^N D^{N*}\Pi^N - G^N C^N)]$ is invertible, in view of which the following equality holds:

$$\begin{aligned} &\left[\lambda I - \left(A^N + \frac{1}{\gamma^2} D^N D^{N*}\Pi^N - G^N C^N\right)\right]^{-1} \\ &\quad \times \left[\lambda I - \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N\right)\right] \\ &= I + \left[\lambda I - \left(A^N + \frac{1}{\gamma^2} D^N D^{N*}\Pi^N - G^N C^N\right)\right]^{-1} (B^N B^{N*}\Pi^N - G^N C^N). \end{aligned}$$

Thus for $\mathrm{Re}\lambda > -\min(\omega_1, \omega_2, \omega_3) := -\omega$ and $N > N_1$, (3.9) can be written as

$$\begin{aligned}(3.10)\qquad &(v_1 - v_2) - \varphi^N BL^N(v_1 - v_2)\\ &\quad + \left[\lambda I - \left(A^N + \frac{1}{\gamma^2} D^N D^{N*}\Pi^N - G^N C^N\right)\right]^{-1} G^N(C - C^N)(v_1 - v_2)\\ &\quad = \left[\lambda I - \left(A^N + \frac{1}{\gamma^2} D^N D^{N*}\Pi^N - G^N C^N\right)\right]^{-1} (P^N h_1 - h_2) - \varphi^N h_1,\end{aligned}$$

where φ^N is

$$\begin{aligned}\varphi^N = &\left\{ I + \left[\lambda I - \left(A^N + \frac{1}{\gamma^2} D^N D^{N*}\Pi^N - G^N C^N\right)\right]^{-1} (B^N B^{N*}\Pi^N - G^N C^N) \right\}\\ &\times \left\{ - \left[\lambda I - \left(A + \frac{1}{\gamma^2} DD^*\Pi + BL^N\right)\right]^{-1} \right.\\ &\qquad \left. + \left[\lambda I - \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\right)\Pi^N\right]^{-1} P^N \right\}.\end{aligned}$$

We show next that there exists N' such that if $N > N'$,

$$(3.11)\quad \left\| \varphi^N BL^N + \left[\lambda I - \left(A^N + \frac{1}{\gamma^2} D^N D^{N*}\Pi^N - G^N C^N\right)\right]^{-1} G^N(C - C^N) \right\|_{\mathcal{L}(X)} < 1,$$

which then says that (3.5), (3.6) are solvable for $\mathrm{Re}\lambda \geq 0$ according to (3.7) and (3.8).

Again we denote by S^N the C_0 semigroup generated by operator $A + \frac{1}{\gamma^2} DD^*\Pi + BL^N$. Since both

$$\left[\lambda I - \left(A + \frac{1}{\gamma^2} DD^*\Pi + BL^N\right)\right]^{-1} B = \int_0^\infty e^{-\lambda t} S^N(t) B dt$$

and

$$\begin{aligned}&\left[\lambda I - \left(A^N + \frac{1}{\gamma^2} D^N D^{N*}\Pi^N - B^N B^{N*}\Pi^N\right)\right]^{-1} P^N B\\ &\qquad = \int_0^\infty e^{-\lambda t} e^{(A^N + \frac{1}{\gamma^2} D^N D^{N*}\Pi^N - B^N B^{N*}\Pi^N)t} P^N B dt\end{aligned}$$

hold for $\mathrm{Re}\lambda \geq -\omega$, for any $T > 0$ we have

$$\begin{aligned}&\left\| \left[\lambda I - \left(A + \frac{1}{\gamma^2} DD^*\Pi + BL^N\right)\right]^{-1} B \right.\\ &\qquad \left. - \left[\lambda I - \left(A^N + \frac{1}{\gamma^2} D^N D^{N*}\Pi^N - B^N B^{N*}\Pi^N\right)\right]^{-1} P^N B \right\|\\ &\qquad \leq \int_0^T e^{-\mathrm{Re}\lambda t} \| e^{(A^N + \frac{1}{\gamma^2} D^N D^{N*}\Pi^N - B^N B^{N*}\Pi^N)t} P^N B - S^N B \| dt\\ &\qquad\quad + \left(M_1 \frac{e^{-(\mathrm{Re}\lambda + \omega_1)T}}{\mathrm{Re}\lambda + \omega_1} + M_3 \frac{e^{-(\mathrm{Re}\lambda + \omega_3)T}}{\mathrm{Re}\lambda + \omega_3} \right) \|B\|.\end{aligned}$$

By Lemma 2.3, part 3, and the principle of uniform boundedness [27], for $\mathrm{Re}\lambda \geq -\omega$ there is a constant $c > 0$ independent of N such that

$$\left\| I + \left[\lambda I - \left(A^N + \frac{1}{\gamma^2} D^N D^{N*}\Pi^N - G^N C^N\right)\right]^{-1} (B^N B^{N*} P^N - G^N C^N)\right\| \leq c.$$

Let $c_L := \max_N \|L^N\|$ and choose $T > 0$ such that for $\mathrm{Re}\lambda > -\omega$

$$c_L c \left(M_1 \frac{e^{-(\mathrm{Re}\lambda + \omega_1)T}}{\mathrm{Re}\lambda + \omega_1} + M_3 \frac{e^{-(\mathrm{Re}\lambda + \omega_3)T}}{\mathrm{Re}\lambda + \omega_3}\right) \|B\| \leq \frac{1}{4}.$$

On the other hand, as $N \to \infty$ we have

$$\begin{aligned} &e^{(A^N + \frac{1}{\gamma^2} D^N D^{N*}\Pi^N - B^N B^{N*}\Pi^N)t} P^N B \to S(t)B \qquad \text{strongly},\\ &S^N(t)B \rightarrow\rightarrow S(t)B \end{aligned}$$

uniformly on $[0, T]$. Thus there exists $N'' > 0$ such that when $N \geq N''$, the following holds:

$$\int_0^T e^{-\mathrm{Re}\lambda t} \| e^{(A^N + \frac{1}{\gamma^2} D^N D^{N*}\Pi^N - B^N B^{N*}\Pi^N)t} P^N B - S^N B\| dt \leq \frac{1}{4 c c_L}$$

by noting that B is compact. It then follows that for $N \geq \max(N', N'') := \hat{N}$, $\|\varphi^N B L^N\| \leq \frac{1}{2}$. Note that $C^N \to C$ as $N \to \infty$, and there exists $\overline{N} > 0$ such that when $N > \overline{N}$

$$\left\| \left[\lambda I - \left(A^N + \frac{1}{\gamma^2} D^N D^{N*} P^N - G^N C^N\right)\right]^{-1} G^N (C - C^N)\right\|_{\mathcal{L}(X)} < \frac{1}{2}.$$

Therefore, for $N > \max(\hat{N}, \overline{N})$, (3.11) holds, and (3.5), (3.6) are solvable for sufficiently large N when $\mathrm{Re}\lambda \geq -\omega$, which yields the conclusion. □

THEOREM 3.4. *Under the assumptions of Theorem* 3.3, *let* $\mathcal{A}_p^N$ *be given. Let*

$$\mathcal{A} = \begin{bmatrix} A + \frac{1}{\gamma^2} DD^*\Pi & -BB^*\Pi \\ \Gamma C^* C & A - (BB^* - \frac{1}{\gamma^2} DD^*)\Pi - \Gamma C^* C \end{bmatrix}$$

and $\mathcal{S}(t)$ *be the* C_0 *semigroup generated by* $\mathcal{A}$ *on* $X \times X$ *with* $D(\mathcal{A}) = D(A) \times D(A)$. *If there exists a* λ *such that the operator* $[\lambda I - \mathcal{A}_p^N]^{-1}$ *is uniformly bounded for all N, then*

$$\mathcal{S}^N(t) \rightarrow\rightarrow \mathcal{S}(t) \qquad \textit{uniformly on bounded } t\textit{-intervals as } N \to \infty,$$

where $\mathcal{S}^N(t)$ *is the* C_0 *semigroup generated by* $\mathcal{A}_p^N$ *on* $X \times X$ *with* $D(\mathcal{A}_p^N) = D(A) \times X^N$.

Proof. Introduce

$$Q^N = \begin{bmatrix} \frac{1}{\gamma^2} DD^*\Pi & -BB^{N*}\Pi^N \\ \Gamma^N C^{N*} C & -(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*})\Pi^N - \Gamma^N C^{N*} C^N \end{bmatrix}$$

and

$$Q = \begin{bmatrix} \frac{1}{\gamma^2} DD^*\Pi & -BB^*\Pi \\ \Gamma C^* C & -(BB^* - \frac{1}{\gamma^2} DD^*)\Pi - \Gamma C^* C \end{bmatrix}.$$

Let $\mathcal{A} = \mathcal{A}_1 + Q$ and $\mathcal{A}_p^N = \mathcal{A}_{p1}^N + Q^N$. To show $S^N(t) \to\to S(t)$ for $t \geq 0$ is equivalent to proving that for $\mathrm{Re}\lambda > -\omega$

$$R(\lambda : \mathcal{A}_p^N) \to\to R(\lambda : \mathcal{A}), \tag{3.12}$$

i.e.,

$$R(\lambda : \mathcal{A}_p^N)\overline{P}^N x \to R(\lambda : \mathcal{A})x \qquad \forall x \in X,$$

where $\overline{P}^N : X \times X \to X \times X^N$ is given by

$$\overline{P}^N = \begin{bmatrix} I & 0 \\ 0 & P^N \end{bmatrix}.$$

Hence we next claim (3.12). First, it is not difficult to verify that

$$\begin{aligned}[\lambda I - (\mathcal{A}_{p1}^N + Q^N)]^{-1} - &[\lambda I - \mathcal{A}_{p1}^N]^{-1} \\ &= [\lambda I - (\mathcal{A}_{p1}^N + Q^N)]^{-1} Q^N [\lambda I - \mathcal{A}_{p1}^N]^{-1}\end{aligned}$$

and

$$[\lambda I - (\mathcal{A}_1 + Q)]^{-1} - [\lambda I - \mathcal{A}_1]^{-1} = [\lambda I - (\mathcal{A}_1 + Q)]^{-1} Q [\lambda I - \mathcal{A}_1]^{-1}.$$

Hence we have

$$\begin{aligned}[\lambda I - &(\mathcal{A}_{p1}^N + Q^N)]^{-1}\overline{P}^N - [\lambda I - (\mathcal{A}_1 + Q)]^{-1} \\ &= [\lambda I - (\mathcal{A}_{p1}^N + Q^N)]^{-1} Q^N \big([\lambda I - \mathcal{A}_{p1}^N]^{-1}\overline{P}^N - [\lambda I - \mathcal{A}_1]^{-1}\big) \\ &\qquad - [\lambda I - (A_{p1}^N + Q^N)]^{-1}(\overline{P}^N Q - Q^N)[\lambda I - \mathcal{A}_1]^{-1} \\ &\qquad + \big([\lambda I - (\mathcal{A}_{p1}^N + Q^N)]^{-1}\overline{P}^N - [\lambda I - \mathcal{A}_1 + Q)]^{-1}\big) Q [\lambda I - \mathcal{A}_1]^{-1} \\ &\qquad + [\lambda I - \mathcal{A}_{p1}^N]^{-1}\overline{P}^N - [\lambda I - \mathcal{A}_1]^{-1}.\end{aligned} \tag{3.13}$$

Reorganizing (3.13) yields

$$\begin{aligned}\Big([\lambda I - &(\mathcal{A}_{p1}^N + Q^N)]^{-1}\overline{P}^N - [\lambda I - (\mathcal{A}_1 + Q)]^{-1}\Big)(I - Q[\lambda I - \mathcal{A}_1]^{-1}) \\ &= \Big(I + [\lambda I - (\mathcal{A}_{p1}^N + Q^N)]^{-1} Q^N\Big)\Big([\lambda I - \mathcal{A}_{p1}^N]^{-1}\overline{P}^N - [\lambda I - \mathcal{A}_1]^{-1}\Big) \\ &\qquad - [\lambda I - (\mathcal{A}_{p1}^N + Q^N)]^{-1}(\overline{P}^N Q - Q^N)[\lambda I - \mathcal{A}_1]^{-1}.\end{aligned}$$

Since

$$[I - Q(\lambda I - \mathcal{A}_1)^{-1}]^{-1} = [\lambda I - \mathcal{A}_1][\lambda I - (\mathcal{A}_1 + Q)]^{-1},$$

we know that $[I + Q(\lambda I - \mathcal{A}_1)^{-1}]^{-1}$ is bounded by the closed graph theorem (see [27]). Therefore, we arrive at

$$\begin{aligned}&[\lambda I - (\mathcal{A}_{p1}^N + Q^N)]^{-1}\overline{P}^N - [\lambda I - (\mathcal{A}_1 + Q)]^{-1} \\ &= \Big\{\Big(I - [\lambda I - (\mathcal{A}_{p1}^N + Q^N)]^{-1} Q^N\Big)\Big([\lambda I - \mathcal{A}_{p1}^N]^{-1}\overline{P}^N - [\lambda I - \mathcal{A}_1]^{-1}\Big) \\ &\qquad - [\lambda I - (\mathcal{A}_{p1}^N + Q^N)]^{-1}(\overline{P}^N Q - Q^N)[\lambda I - \mathcal{A}_1]^{-1}\Big\}[\lambda I - \mathcal{A}_1][\lambda I - (\mathcal{A}_1 + Q)]^{-1}.\end{aligned}$$

By assumption (A1), we have

$$[\lambda I - \mathcal{A}_{p1}^N]^{-1}\overline{P}^N \to [\lambda I - \mathcal{A}_1]^{-1}, \qquad \overline{P}^N Q - Q^N \to 0 \qquad \text{strongly as } N \to \infty.$$

Therefore, (3.12) holds as $N \to \infty$, thus completing the proof. □

Remark 3. One sufficient condition for $[\lambda I - \mathcal{A}_p^N]^{-1}$ to be uniformly bounded in N is assumption (A4). Clearly if A is an analytic C_0 semigroup, $[\lambda I - \mathcal{A}_p^N]^{-1}$ is uniformly bounded in N according to Theorem 3.3.

THEOREM 3.5. *Under the assumptions of Theorem* 3.3, *let $\mathcal{A}_p^N$ be given. Recall that*

$$\mathcal{A}^N = \begin{bmatrix} A & BL^N \\ G^N C & A^N + M^N \end{bmatrix}.$$

If $\mathcal{A}_p^N$ generates a stable semigroup, then $\mathcal{A}^N$ also generates a stable semigroup on $X \times X^N$.

Proof. *Step* 1. Consider the system

$$\begin{aligned}\dot{x_h} &= \left(A + \frac{1}{\gamma^2}DD^*\Pi\right)x_h - BB^{N*}\Pi^N p_k^N + Dw,\\ \dot{p_k^N} &= \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2}D^N D^{N*}\right)\Pi^N - \Gamma^N C^{N*}C^N\right)p_k^N\\ &\qquad + \Gamma^N C^{N*}Cx_h + \Gamma^N C^{N*}\eta,\\ x_h(0) &= h,\\ p_k^N(0) &= k.\end{aligned} \tag{3.14}$$

We claim that for a given $\delta > 0$, $0 < \delta < \gamma$, there exists $N_\delta > 0$ such that when $N \geq N_\delta$ the following estimate holds:

$$\begin{aligned}&\int_0^\infty \left(\|Hx_h\|^2 + \|B^{N*}\Pi^N p_k^N\|^2 - \gamma^2\left(\left\|w + \frac{1}{\gamma^2}D^*\Pi x_h\right\|^2 + \|\eta\|^2\right)\right)dt\\ &\qquad \leq -\delta^2\int_0^\infty (\|w\|^2 + \|\eta\|^2)dt + C_0(\|h\|^2 + \|k\|^2),\end{aligned} \tag{3.15}$$

where C_0 is some positive constant. First we denote the solution of (3.14) with initial condition $(0,0)$ by (x_0, p_0^N). It is easy to see that by evaluating $\frac{d}{dt}(\Pi x_0, x_0)$ and integrating from 0 to ∞, we have the following identity:

$$\begin{aligned}&\int_0^\infty \Bigg(\|Hx_0\|^2 + \|B^{N*}\Pi^N p_0^N\|^2 - \|B^*\Pi x_0 - B^{N*}\Pi^N p_0^N\|^2\\ &\qquad + \gamma^2\left(\|w\|^2 - \left\|w + \frac{1}{\gamma^2}D^*\Pi x_0\right\|^2\right)\Bigg)dt = 0.\end{aligned} \tag{3.16}$$

We then consider the following coupled (optimal) system:

$$\begin{aligned}\dot{\overline{x}} &= \left(A + \frac{1}{\gamma^2}DD^*\Pi\right)\overline{x} - BB^*\Pi\overline{p} + Dw,\\ \dot{\overline{p}} &= \left(A - \left(BB^* - \frac{1}{\gamma^2}DD^*\right)\Pi - \Gamma C^*C\right)\overline{p} + \Gamma C^*C\overline{x} + \Gamma C^*\eta,\\ \overline{x}(0) &= 0,\\ \overline{p}(0) &= 0.\end{aligned} \tag{3.17}$$

Letting $\xi = \overline{x} - \overline{p}$, we have

$$\dot{\xi} = \left(A + \frac{1}{\gamma^2} DD^*\Pi - \Gamma C^* C\right)\xi + Dw - \Gamma C^*\eta, \qquad \xi(0) = 0. \tag{3.18}$$

Note that (3.17) is in fact the dual system of (2.6) in Lemma 2.3, and thus by the duality theorem (see Appendix I) we have

$$\gamma_1^2 := \sup_{w,\eta} \frac{\int_0^\infty \|B^*\xi\|^2 dt}{\int_0^\infty (\|w\|^2 + \|\eta\|^2)dt} = \sup_{w,\eta} \frac{\int_0^\infty \|B^*\Pi(\overline{x}(t) - \overline{p}(t))\|^2 dt}{\int_0^\infty (\|w\|^2 + \|\eta\|^2)dt} < \gamma^2.$$

Since $(x_h, p_k^N) \in L^2(0,\infty; X) \times L^2(0,\infty; X^N)$ and $(x,p) \in L^2(0,\infty;X) \times L^2(0,\infty;X)$, according to Theorem 3.4, letting $\varepsilon = (\gamma^2 - \gamma_1^2)/2$, there exists $N' > 0$ such that when $N > N'$

$$\int_0^\infty \|B^*\Pi x_0 - B^*\Pi\overline{x}\|^2 dt < \varepsilon, \qquad \int_0^\infty \|B^*\Pi p - B^{N*}\Pi^N p^N\|^2 dt < \varepsilon,$$

which implies that when $N > N'$,

$$\sup_{w,\eta} \frac{\int_0^\infty \|B^*\Pi x_0(t) - B^{N*}\Pi^N p_0^N(t)\|^2 dt}{\int_0^\infty (\|w\|^2 + \|\eta\|^2)dt} < \gamma^2.$$

Hence we can find $\delta < \gamma$ and $N_\delta > N'$ such that for $N \geq N_\delta$ we have

$$\sup_{w,\eta} \frac{\int_0^\infty \|B^*\Pi x_0(t) - B^{N*}\Pi^N p_0^N(t)\|^2 dt}{\int_0^\infty (\|w\|^2 + \|\eta\|^2)dt} \leq \gamma^2 - \delta^2. \tag{3.19}$$

Hence (3.15), (3.19) yield

$$\begin{aligned}\int_0^\infty \left(\|Hx_0\|^2 + \|B^{N*}\Pi^N p_0^N\|^2 - \gamma^2\left(\left\|w + \frac{1}{\gamma^2}D^*\Pi x_0\right\|^2 + \|\eta\|^2\right)\right) \\ \leq -\delta^2 \int_0^\infty (\|w\|^2 + \|\eta\|^2)dt.\end{aligned} \tag{3.20}$$

Let $x_1 := x_h - x_0$, $p_1^N := p_k^N - p_0^N$. Note that (x_1, p_1^N) depends only on h, k and not on w, η, and clearly for some positive constants c_1 and c_2:

$$\|x_1\| \leq c_1\|h\|, \qquad \|p_1^N\| \leq c_2\|k\|.$$

Therefore, replacing (x_0, p_0^N) by $(x_h - x_1, p_k^N - p_1^N)$ in (3.20), we arrive at (3.15).

Step 2. Introduce the system

(3.21)

$$\begin{aligned}
&\dot{\hat{x}}_h = A\hat{x}_h - BB^{N*}\Pi^N\hat{p}_k^N,\\
&\dot{\hat{p}}_k^N = \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2}D^N D^{N*}\right)\Pi^N - \Gamma^N C^{N*}C^N\right)\hat{p}_k^N + \Gamma^N C^{N*} C\hat{x}_h,\\
&\hat{x}_h(0) = h,\\
&\hat{p}_k^N(0) = k.
\end{aligned}$$

We shall prove that there exists $N_1 > 0$ such that for $N \geq N_1$ we have

$$\hat{x}_h, \hat{p}_k^N \in L^2(0, +\infty; X).$$

For any $T > 0$, we define

$$x_T(t) = \begin{cases} \hat{x}_h(t) & \text{if } t \leq T, \\ x_{\hat{x}_T, \hat{p}_T^N}(t - T) & \text{if } t > T \end{cases}$$

and

$$p_T^N(t) = \begin{cases} \hat{p}_k^N(t) & \text{if } t \leq T, \\ p^N_{\hat{x}_T, \hat{p}_T^N}(t - T) & \text{if } t > T, \end{cases}$$

where $(x_{\hat{x}_T, \hat{p}_T^N}, p^N_{\hat{x}_T, \hat{p}_T^N})$ is the solution of

(3.22)
$$\begin{aligned} &\dot{x} = \left(A + \frac{1}{\gamma^2} DD^*\Pi\right) x - BB^{N*}\Pi^N p^N, \\ &\dot{p^N} = \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N - \Gamma^N C^{N*} C^N\right) p^N + \Gamma^N C^{N*} Cx, \\ &x(T) = \hat{x}_h(T), \\ &p^N(T) = \hat{p}_k^N(T). \end{aligned}$$

If we let

$$w_T(t) = \begin{cases} -\gamma^{-2} D^*\Pi\hat{x}_h(t) & \text{if } t \leq T, \\ 0 & \text{if } t > T, \end{cases}$$

we can see that the pair $(x_T(\cdot), p_T^N(\cdot))$ is the solution of the system (3.14) corresponding to the perturbation $w(\cdot) = w_T(\cdot)$ and $\eta = 0$. Hence for $N \geq N_\delta$, by applying (3.15) we have

$$\begin{aligned} &\int_0^\infty \left(\|Hx_T\|^2 + \|B^{N*}\Pi^N p_T^N\|^2 - \gamma^2 \left\| w_T + \frac{1}{\gamma^2} D^*\Pi x_T \right\|^2 \right) dt \\ &\qquad \leq -\delta^2 \int_0^\infty \|w\|^2 dt + C_0(\|h\|^2 + \|k\|^2). \end{aligned} \tag{3.23}$$

Note that by the definition of w_T we have

$$\begin{aligned} &\int_T^\infty \left(\|Hx_T\|^2 + \|B^{N*}\Pi^N p_T^N\|^2 - \gamma^2 \left\| w_T + \frac{1}{\gamma^2} D^*\Pi x_T \right\|^2 \right) dt \\ &\qquad = \int_T^\infty \left(\|Hx_{\hat{x}_T, \hat{p}_T^N}\|^2 + \|B^{N*}\Pi^N p^N_{\hat{x}_T, \hat{p}_T^N}\|^2 - \gamma^2 \left\| \frac{1}{\gamma^2} D^* p^N_{\hat{x}_T, \hat{p}_T^N} \right\|^2 \right) dt. \end{aligned}$$

On the other hand, making use of Riccati equation (1.8) yields

$$\begin{aligned} &\int_T^\infty \left(\|Hx_T\|^2 + \|B^{N*}\Pi^N p_T^N\|^2 - \gamma^2 \left\| w_T + \frac{1}{\gamma^2} D^*\Pi x_T \right\|^2 \right) dt \\ &\qquad = (\Pi\hat{x}_T, \hat{x}_T) + \int_T^\infty \|B^*\Pi x_{\hat{x}_T, \hat{p}_T^N} - B^{N*}\Pi^N p^N_{\hat{x}_T, \hat{p}_T^N}\|^2 dt. \end{aligned} \tag{3.24}$$

Therefore we deduce from (3.23) and (3.24) the estimate

$$\begin{aligned}
&\int_0^\infty \left(\|Hx_T\|^2 + \|B^{N*}\Pi^N p_T^N\|^2 - \gamma^2 \left\| w_T + \frac{1}{\gamma^2} D^*\Pi x_T \right\|^2 \right) dt \\
&= \int_0^T (\|Hx_T\|^2 + \|B^{N*}\Pi^N p_T^N\|^2) dt + (\Pi \hat{x}_T, \hat{x}_T) \\
&\qquad + \int_T^\infty \|B^* \Pi x_{\hat{x}_T, \hat{p}_T^N} - B^{N*}\Pi^N p^N_{\hat{x}_T, \hat{p}_T^N}\|^2 dt \\
&< -\delta^2 \int_0^\infty \|w_T\|^2 dt + C_0(\|h\|^2 + \|k\|^2),
\end{aligned} \tag{3.25}$$

which yields (with c_1 some positive constant):

$$\int_0^T \|D^*\Pi \hat{x}_h\|^2 dt \leq C_1(\|h\|^2 + \|k\|^2).$$

By letting $T \to \infty$, we have $D^*\Pi\hat{x}_h \in L^2(0, +\infty; W)$. Note that $(\hat{x}_h, \hat{p}_k^N)$ appears as the solution of (3.14) with

$$w = -\frac{1}{\gamma^2} D^*\Pi\hat{x}_h \text{ and } \eta = 0.$$

Thus we have $\hat{x}_h, \hat{p}_k^N \in L^2(0, +\infty; X)$, and according to a theorem due to Datko (see [11]) $\mathcal{A}^N$ is exponentially stable. This completes the proof of Theorem 3.5. □

COROLLARY 3.6. *For any choice of the matrices A^N, L^N, G^N, M^N, if $\mathcal{A}_p^N$ defined in Theorem* 3.3 *satisfies the spectrum-determined growth condition, then there exists $N' > 0$ such that when $N > N'$, $\mathcal{A}^N$ which is defined in Theorem* 3.5 *generates an exponentially stable semigroup.*

Proof. The proof is quite straightforward. Note that according to Theorem 3.3, there exists an $N' > 0$ such that when $N > N'$ we have

$$\sigma(\mathcal{A}_p^N) \subseteq (-\infty, -\varepsilon]$$

for some $\varepsilon > 0$ which is independent of N, and $\mathcal{A}_p^N$ satisfies the spectrum-determined growth condition. Thus the growth constant ω_p,

$$\omega_p = \lim_{t\to\infty} \frac{1}{t} \ln \|e^{\mathcal{A}_p^N t}\|,$$

is negative, and hence $\mathcal{A}_p^N$ generates an exponentially stable C_0 semigroup. By Theorem 3.5, we have that when $N > N'$, $\mathcal{A}^N$ is exponentially stable. □

Equipped with all the results above, we are now in a position to prove the main theorem stated at the beginning of this section.

Proof of the main theorem. Let us view the system and its finite-dimensional compensator together as a coupled system, i.e.,

$$\begin{aligned}
&(3.26) \\
&\dot{x} = Ax - BB^{N*}\Pi^N p + Dw, \\
&\dot{p}^N = \left(A^N - \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N - \Gamma^N C^{N*} C^N\right) p^N + \Gamma^N C^{N*} C x + \Gamma^N C^{N*}\eta, \\
&x(0) = 0, \\
&p^N(0) = 0.
\end{aligned}$$

By Corollary 3.6, there exists $N_1 > 0$ such that when $N > N_1$, the matrix operator

$$\mathcal{A}^N = \begin{bmatrix} A & BL^N \\ G^N C & A^N + M^N \end{bmatrix}$$

generates an exponentially stable C_0 semigroup; thus the coupled system (3.26) is an exponentially stable system. Let $\hat{w} = w - \frac{1}{\gamma^2} D^* \Pi x$, $w \in L^2(0, \infty; W)$; then there exists N_2 such that (x, p^N) satisfies (3.15) with $h = 0, k = 0$ for $N > N_2$, i.e.,

$$\begin{aligned} \int_0^\infty \Big(\|Hx\|^2 + \|B^{N*}\Pi^N p^N\|^2 - \gamma^2(\|\hat{w}\|^2 + \|\eta\|^2) \Big) dt \\ \leq -\delta^2 \int_0^\infty \left(\left\| \hat{w} + \frac{1}{\gamma^2} D^* \Pi x \right\|^2 + \|\eta\|^2 \right) dt, \end{aligned} \tag{3.27}$$

which is equivalent to

$$\begin{aligned} \sup_{\hat{w},\eta} \frac{\int_0^\infty (\|Hx\|^2 + \|B^{N*}\Pi^N p^N\|^2) dt}{\int_0^\infty (\|\hat{w}\|^2 + \|\eta\|^2) dt} \\ \leq \gamma^2 - \delta^2 \inf_{\hat{w},\eta} \frac{\int_0^\infty (\|\eta\|^2 + \|\hat{w} + \frac{1}{\gamma^2} D^* \Pi x\|^2) dt}{\int_0^\infty (\|\hat{w}\|^2 + \|\eta\|^2) dt}. \end{aligned}$$

Since $\mathcal{A}_p^N$ is exponentially stable, we immediately have

$$\inf_{w,\eta} \frac{\int_0^\infty (\|\eta\|^2 + \|w - \frac{1}{\gamma^2} D^* \Pi x\|^2) dt}{\int_0^\infty (\|w\|^2 + \|\eta\|^2) dt} > 0.$$

Therefore (1.16) holds and the proof is complete. □

4. An application: Robust control of flexible structures. As an illustration of the theory developed in the previous section, we consider here the problem of robust control of flexible structures and provide a general computational scheme for a finite-dimensional compensator design, which can be used for numerical implementation of the full-order controller for particular structures such as the Euler–Bernoulli beam with the Kelvin–Voigt damping.

A typical class of flexible structures can be generically described by a system of partial differential equations:

$$M(\xi) \frac{\partial^2}{\partial t^2} z(\xi, t) + d_0 \frac{\partial}{\partial t} z(\xi, t) + A_0 z(\xi, t) = B_0 u(\xi, t) + D_0 w(\xi, t), \qquad t > 0, \tag{4.1}$$

where $z(\xi, t)$ is a vector of displacements of the structure (denoted by Ω) off its equilibrium position, as a function of space variable ξ and time t; $M(\xi)$ is the mass density; A_0 is a linear, time-invariant, self-adjoint, and positive differential operator; the domain $D(A_0)$ consists of all smooth functions satisfying (4.1) with appropriate boundary conditions and is dense in $L^2(\Omega)$; d_0 represents the inherent damping operators; $B_0 u$ denotes the control forces, while $D_0 w$ represents the external disturbance forces on the structure.

In most cases, the operator A_0, attributed to stiffness, is assumed to have a compact resolvent and thus the eigenvalues $\{\lambda_i\}$ form an infinitely increasing sequence of nonnegative real numbers. This can be expressed as the eigenproblem: $A_0 \phi_i =$

$\lambda_i \phi_i$, where $\{\sqrt{\lambda_i}\}$ are called the *vibration mode frequencies*, and $\{\phi_i(\xi)\}$ are the corresponding *vibration mode shapes*. Possible disturbances result from atmospheric effects, meteorite collisions, as well as pumps and motors. The objective of designing a (finite-dimensional) H^∞-optimal compensator for flexible structures is to maintain the attitude of each mode as close as possible to the desired attitude in the presence of disturbances acting on the structure (system) and corrupting the measurements.

Under the assumption that M is a bounded, self-adjoint, and coercive linear operator on $L^2(\Omega)$, without loss of any generality, (4.1) can be transformed into an evolution equation on $L^2(\Omega)$ in the form

$$\ddot{z}(t) + \tilde{d}_0 \dot{z}(t) + \tilde{A}_0 z(t) = \tilde{B}_0 u(t) + \tilde{D}_0 w(t), \qquad t > 0.$$

Let $H_0 = L^2(\Omega)$, and let V be the Hilbert space which is the completion of $D(\tilde{A}_0)$ with respect to the inner product $\langle v_1, v_2 \rangle_V := \langle \tilde{A}_0 v_1, v_2 \rangle_{H_0}$, that is, $V = D(\tilde{A}_0^{1/2})$. In this case the following embedding holds:

$$V \subset H_0 = H_0' \subset V'$$

where H_0' and V' represent the dual spaces of H_0 and V, respectively, with the injection from V into H_0 and from H_0 into V' being continuous with dense ranges. The total energy space is defined as $X = V \times H_0$ with inner product given by

$$\langle [v_1, h_1], [v_2, h_2] \rangle_X = \langle v_1, v_2 \rangle_V + \langle h_1, h_2 \rangle_{H_0}.$$

The state-space formulation of (4.1) is then

$$\dot{x}(t) = Ax(t) + Bu(t) + Dw(t),$$

where

$$A = \begin{bmatrix} 0 & I \\ -\tilde{A}_0 & -\tilde{d}_0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ \tilde{B}_0 \end{bmatrix}, \qquad D = \begin{bmatrix} 0 \\ \tilde{D}_0 \end{bmatrix}, \qquad x = \begin{bmatrix} z \\ \dot{z} \end{bmatrix}.$$

Let $\mathcal{V} := V \times V$ and $\mathcal{H} := H_0 \times H_0$. In view of the nature of flexible structures and the above discussion, we can assume that $\tilde{d}_0 \in \mathcal{L}(V)$ and the operator $A \in \mathcal{L}(\mathcal{V}, \mathcal{V}')$ is generated by the bilinear continuous form on $\mathcal{V}$,

$$\langle -Ax, v \rangle_{\mathcal{V}' \times \mathcal{V}} = a(x, v) \qquad \text{for } x, v \in \mathcal{V},$$

where $a(\cdot, \cdot)$ is symmetric, and V-H_0 coercive:

$$\exists \alpha > 0, \quad \exists \lambda \in \mathbb{R}, \quad a(x, x) + \lambda \|x\|_{\mathcal{V}}^2 \geq \alpha \|x\|_{\mathcal{H}}^2.$$

Thus A generates an *analytic* semigroup e^{At} on $\mathcal{H}$ (see Theorem 2.12 of [6]). Moreover, if $0 \in \rho(A)$, then e^{At} is a compact C_0 semigroup of contraction. Let V^N be the span of N linearly independent vectors e_j, which is the orthogonal projection of V by operator P_{V^N}. Furthermore, we define the Hilbert space $X^N = V^N \times V^N$, whose inner product is the same as X. Let

$$A^N = \begin{bmatrix} 0 & I \\ -\mathcal{M}^{-\mathcal{N}} \mathcal{K}^{\mathcal{N}} & -\mathcal{M}^{-\mathcal{N}} \lceil_\infty^{\mathcal{N}} \end{bmatrix}_{2N \times 2N},$$

where $\mathcal{M}^N = [\langle e_i, e_j \rangle_{H_0}]_{N \times N}$, $\mathcal{K}^N = [\langle \tilde{A}_0^{1/2} e_i, \tilde{A}_0^{1/2} e_j \rangle_{H_0}]_{N \times N}$, $d_1^N = [\langle d_0 e_i, e_j \rangle]_{N \times N}$. Then the semigroups $e^{A^N t}$ and $e^{A^{N*} t}$ satisfy assumption (A1) of section 2.

Proposition 4.1. *Assume that the projection operator $P_{V^N} : V \to V^N$ is the V-projection onto V^N, which is a sequence of finite-dimensional subspaces, and $v_N = P_{V^N} v \in V^N$ converges to v for any $v \in V$. Let*

$$P^N = \begin{bmatrix} P_{V^N} & 0 \\ 0 & P_{H^N} \end{bmatrix},$$

where P_{H^N} is the H-projection onto V^N. Then, we have the following:

(a) *$e^{A^N t} P^N$ converges strongly to e^{At}, uniformly in t for t in bounded intervals;*

(b) *$e^{A^{N*} t} P^N$ converges strongly to $e^{A^* t}$, uniformly in t for t in bounded intervals.*

Outline of the proof. Since the embedding $V \subset H_0$ is compact, a direct calculation leads to the result that $(\lambda I - A^N)^{-1}$ strongly converges to $(\lambda I - A)^{-1}$ when $\lambda > 0$. Then application of the Trotter–Kato theorem yields (a), and the proof of (b) is similar. □

We assume that the control forces are implemented through m actuators. Thus there exist m vectors $b_i \in H_0, 1 \leq i \leq m$, such that

$$\tilde{B}_0 u = \sum_{i=1}^{m} b_i u_i, \quad \text{where} \quad u = [u_1, u_2, \ldots, u_m]^T \in \mathbb{R}^m,$$

where u_i are called the *actuator amplitudes* and b_i are referred to as the *actuator influence functions* in H_0. The adjoint of $\tilde{B}_0$ is thus given by

$$\tilde{B}_0^* h = [\langle b_1, h\rangle H_0, \langle b_2, h\rangle_{H_0}, \ldots, \langle b_m, h\rangle_{H_0}]^T, \quad \text{where} \quad h \in H_0.$$

Recall that the performance index is given by (1.15). Here we define $Q := H^* H$. Since $Q = Q^* \in \mathcal{L}(X)$ and $X = V \times H_0$, Q can be written as

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12}^* & Q_{22} \end{bmatrix},$$

where $Q_{11} = Q_{11}^* \in \mathcal{L}(V), Q_{12} \in \mathcal{L}(H_0, V)$, and $Q_{22} = Q_{22}^* \in \mathcal{L}(H_0)$.

Let the measurement be produced by p sensors

$$y(t) = C_0 z(t) + E_0 \dot{z}(t) + \eta(t),$$

where $C_0, E_0 \in \mathcal{L}(H_0, \mathbb{R}^p), y_i(t) = (c_i, z)_{H_0} + (v_i, \dot{z})_{H_0} + \eta_i, c_i, v_i \in H_0, \eta_i \in W_1, i = 1, 2, \ldots, p$. Functions $\{c_i\}$ are the position sensors, and functions $\{v_i\}$ are the velocity sensors in H_0. If we introduce $C = [C_0, E_0]$, then the partial observation can be written as $y(t) = Cx(t) + \eta(t)$.

Next we let

$$\hat{Q}^N = \begin{bmatrix} [\langle e_i, Q_{11} e_j\rangle]_V & [\langle e_i, Q_{12} e_j\rangle]_V \\ [\langle e_i, Q_{12} e_j\rangle]_V^T & [\langle e_i, Q_{22} e_j\rangle]_{H_0} \end{bmatrix}_{2N \times 2N}, \quad C^N = \begin{bmatrix} [\langle e_i, c_j\rangle]_{H_0} \\ [\langle e_i, v_j\rangle]_{H_0} \end{bmatrix}_{2N \times p}^T,$$

$$B^N = \begin{bmatrix} 0 \\ \mathcal{M}^{-N} \mathcal{B}^N \end{bmatrix}_{2N \times m}, \quad d^N = \begin{bmatrix} 0 \\ \mathcal{M}^{-N} d_1^N \end{bmatrix}_{2N \times N}, \quad D^N = \begin{bmatrix} 0 \\ \mathcal{M}^{-N} \mathcal{D}^N \end{bmatrix}_{2N \times 1},$$

and

$$\mathcal{B}^N = [\langle e_i, b_j\rangle]_{N \times m}, \quad \mathcal{D}^N = [\langle e_i, P_{H^N} D\cdot\rangle_H]_{N \times 1}.$$

Straightforward calculations lead to the following two propositions.

PROPOSITION 4.2. *Let P^N, B^N, C^N, and D^N be given as above. Then we have*

$$(1)\ B^N = P^N B; \quad (2)\ C^N = P^N C\big|_{X^N}; \quad (3)\ D^N = P^N D; \quad (4)\ Q^N = P^N Q\big|_{X^N};$$

$$(5)\ D^N D^{N*} = \begin{bmatrix} 0 & 0 \\ 0 & [\langle P_{H_0^N} D D^* P^*_{H^N} e_i, e_j \rangle_{H_0}]_{N\times N} \end{bmatrix}_{2N\times 2N},$$

where $C\big|_{X^N}$ represents the restriction of operator C onto X^N, and so does $Q\big|_{X^N}$.

PROPOSITION 4.3. *Let $\mathcal{G}^N$ be the Gramian matrix defined as*

$$\mathcal{G}^N = \begin{bmatrix} [(e_i, e_j)]_V & 0 \\ 0 & [(e_i, e_j)]_{H_0} \end{bmatrix}_{2N\times 2N}.$$

Then we have (1) $A^{N*} = \mathcal{G}^{-N}(A^N)^T \mathcal{G}^N$; (2) $B^{N*} = (B^N)^T \mathcal{G}^N$; (3) $Q^N = \mathcal{G}^{-N}\hat{Q}^N$.

Hence we have provided the complete matrix representations of the Riccati equations:

$$\Pi^N A^N + A^{N*}\Pi^N - \Pi^N \left(B^N B^{N*} - \frac{1}{\gamma^2} D^N D^{N*} \right) \Pi^N + Q^N = 0 \tag{4.2}$$

and

$$\Sigma A^{N*} + A^N \Sigma^N - \Sigma^N \left(C^{N*} C^N - \frac{1}{\gamma^2} Q^N \right) \Sigma^N + D^N D^{N*} = 0 \tag{4.3}$$

through Propositions 4.2 and 4.3. Let γ^{Π}_{Nc} (respectively, γ^{Σ}_{Nc}) be the optimum attenuation level associated with (4.2) (respectively, (4.3)), that is, for $\gamma > \gamma^{\Pi}_{Nc}$ (respectively, $\gamma > \gamma^{\Sigma}_{Nc}$) (4.2) (respectively, (4.3)) admits a nonnegative solution Π^N (respectively, Σ^N), which satisfies assumption (A2) (respectively, (A3)), and for $\gamma < \gamma^{\Pi}_{Nc}$ (respectively, $\gamma < \gamma^{\Sigma}_{Nc}$) there will be no such solution. We use similar notations γ^{Π}_c and γ^{Σ}_c to represent the optimum attenuation level associated with Riccati equations (1.8), (1.9), respectively. Now we are interested in the question of whether we can have

$$\gamma^{\Pi}_{Nc} \to \gamma^{\Pi}_c, \qquad \gamma^{\Sigma}_{Nc} \to \gamma^{\Sigma}_c \qquad \text{as } N \to \infty$$

which is equivalent to the question of whether the performance of the system can be made arbitrarily close to that obtained under the infinite-dimensional controller. According to [30], γ^{Π}_{Nc} has a closed-form representation in terms of the norm of an operator, denoted by $\mathcal{S}_{N\Pi} \in \mathcal{L}(X^N)$. That is, $\gamma^{\Pi}_{Nc} = \|\mathcal{S}_{\mathcal{N}\Pi}\|_{\mathcal{L}(\mathcal{X}^{\mathcal{N}})}$, and so does $\gamma^{\Pi}_c = \|\mathcal{S}_{\Pi}\|_{\mathcal{L}(X)}$. Thus to require $\gamma^{\Pi}_{Nc} \to \gamma^{\Pi}_c$ is equivalent to requiring

$$\|\mathcal{S}_{N\Pi}\|_{\mathcal{L}(X^N)} \to \|\mathcal{S}_{\Pi}\|_{\mathcal{L}(X)}. \tag{4.4}$$

However, we generally have only

$$\mathcal{S}_{N\Pi} \to\to \mathcal{S}_{\Pi}.$$

Hence it will be necessary to impose additional restrictions on the system for the approximation. One can verify that a sufficient condition for (4.4), by the closed-form of γ^{Π}_c, is that either the semigroup e^{At} or operator D be compact, since in this case operator $\mathcal{S}_{\Pi}$ would be a compact operator. Similar results hold for the Riccati equation (4.3). Let the matrix Π^N

$$\Pi^N = \begin{bmatrix} \Pi^N_{11} & \Pi^N_{12} \\ \Pi^{N*}_{12} & \Pi^N_{22} \end{bmatrix}_{2N\times 2N}$$

be the solution of (4.2). Then the finite-dimensional compensator in X^N can be constructed as

$$\begin{aligned} u_c(t) &= -\tilde{B}_0^*[\Pi_{11}^{N*} z_{N1}(t) + \Pi_{22}^N \dot{z}_{2N}(t)], \\ \dot{z}_N(t) &= \left(A^N - \left(B^N B^{N*} + \frac{1}{\gamma^2} D^N D^{N*}\right)\Pi^N\right) z_N(t) \\ &\quad + \Sigma^N \left(I - \frac{1}{\gamma^2}\Pi^N \Sigma^N\right)^{-1} C^{N*} y(t), \end{aligned} \tag{4.5}$$

where $z_N = (z_{N1}, z_{2N}) \in X^N$. We summarize the above discussion in the following theorem.

THEOREM 4.4. *Assume that* (A, B) *is stabilizable,* (A, H) *is detectable, and either* $0 \in \rho(A)$ *or both* D *and* H *are compact. If a* γ*-attenuation level is achieved by an infinite-dimensional compensator, then*

(1) *for sufficiently large* N*, Riccati equation* (4.2) *admits a nonnegative solution* $\Pi^N \in \Sigma^+(X^N)$ *which satisfies assumption* (A2);

(2) *for sufficiently large* N*, Riccati equation* (4.3) *admits a nonnegative solution* $\Sigma^N \in \Sigma^+(X^N)$ *which satisfies assumption* (A3);

(3) *the finite-dimensional compensator given in* (4.5) *achieves the same attenuation level by choosing an appropriate* N*; that is, when* $x_0 = 0$ *we have*

$$\sup_{w,\eta} \frac{\displaystyle\int_0^\infty (\|Hx(t)\|_{Y_1}^2 + \|u_c(t)\|_{\mathbb{R}^m}^2)dt}{\displaystyle\int_0^\infty (\|w(t)\|_W^2 + \|\eta(t)\|_{W_1}^2)dt} < \gamma^2.$$

5. Concluding remarks. In this paper we have established the existence of finite-dimensional compensators in the H^∞-optimal control of infinite-dimensional systems by using a Galerkin-type approximation. This result readily covers the case when the system structure operator A has a complete set of generalized eigenvectors, such as A being a Riesz-spectral operator.[4] In order to construct such a finite-dimensional compensator, we required that there exist an infinite-dimensional compensator achieving a γ-attenuation level. This is quite natural because one cannot expect a finite-dimensional compensator to be able to achieve the γ-attenuation level if there is no infinite-dimensional compensator to do so. From the proofs we can see that the finite-dimensional controller in fact converges to the infinite-dimensional controller as its order increases. The order of the finite-dimensional controller depends on how much error the system can tolerate, which is measured by the difference between γ and γ_c, where γ_c is the optimum level of disturbance attenuation for the system, which can be defined in a similar way as in the finite-dimensional case (see [30]).

The approach used in this paper can be extended to the case when the finite-dimensional controller is only allowed to act on the boundary. Such an extension as well as results of some numerical experiments will be reported elsewhere.

[4] A is a Riesz-spectral operator means that it has simple eigenvalues and the corresponding eigenvectors form a Riesz basis: a sequence of vectors $\{\phi_n, n \geq 1\}$ in a Hilbert space X such that (1) $\overline{span\{\phi_n\}} = X$, (2) there exist constants m, M such that for all N

$$m \sum_{n=1}^N |\alpha_n|^2 \leq \left\| \sum_{n=1}^N \phi_n \right\|^2 \leq M \sum_{n=1}^N |\alpha_n|^2.$$

Appendix I. The following theorem is from [3].

DUALITY THEOREM. *Consider Hilbert spaces X, Y, Z, identified with their duals. Let $\mathcal{A}$ be the infinitesimal generator of a C_0 semigroup in X, and $\mathcal{B} \in \mathcal{L}(Y, X), \mathcal{C} \in \mathcal{L}(X, Z)$. With the triple $(\mathcal{A}, \mathcal{B}, \mathcal{C})$, associate a linear system*

$$\dot{x} = \mathcal{A}x + \mathcal{B}u, \quad x(0) = 0$$

and a corresponding observation $\mathcal{C}x$. The system dual to this is defined as

$$\dot{\xi} = \mathcal{A}^*\xi + \mathcal{C}^*v, \quad \xi(0) = 0,$$

with a corresponding observation $\mathcal{B}^*\xi$. Assume that $\mathcal{A}$ is exponentially stable. Then we have

$$\sup_u \frac{\int_0^\infty |\mathcal{C}x|^2 dt}{\int_0^\infty |u|^2 dt} = \sup_v \frac{\int_0^\infty |\mathcal{B}^*\xi|^2 dt}{\int_0^\infty |v|^2 dt}.$$

Appendix II. We provide here a proof for Lemma 2.1, which is in the form of a sequence of propositions.

PROPOSITION A.1. *The Riccati equation*

$$\Pi A + A^*\Pi - \Pi\left(BB^* - \frac{1}{\gamma^2}DD^*\right)\Pi + H^*H = 0 \tag{A2.1}$$

has at most one solution $\Pi \in \Sigma^+(X)$ such that $A - (BB^* - \frac{1}{\gamma^2}DD^*)\Pi$ is exponentially stable.

Proof. Let $\Pi_1, \Pi_2 \in \Sigma^+(X)$ be two solutions of (A2.1) such that both $A - (BB^* - \frac{1}{\gamma^2}DD^*)\Pi_1$ and $A - (BB^* - \frac{1}{\gamma^2}DD^*)\Pi_2$ are exponentially stable. Consider the system

$$\dot{x} = Ax + Bu + Dw, \qquad x(0) = x_0.$$

Since Π_1, Π_2 satisfy (A2.1), we have

$$\begin{aligned}&\frac{d}{dt}\langle(\Pi_1 - \Pi_2)x, x\rangle \\ &= \gamma^2\left(\left\|w - \frac{1}{\gamma^2}D^*\Pi_2 x\right\|^2 - \left\|w - \frac{1}{\gamma^2}D^*\Pi_1 x\right\|^2\right) + \|u + B^*\Pi_2 x\|^2 - \|u + B^*\Pi_1 x\|^2.\end{aligned}$$

By integrating this between 0 and T, we obtain

$$\begin{aligned}&\langle(\Pi_1 - \Pi_2)x(T), x(T)\rangle - \langle(\Pi_1 - \Pi_2)x_0, x_0\rangle \\ &\quad = \int_0^T \left\{\gamma^2\left(\left\|w - \frac{1}{\gamma^2}D^*\Pi_1 x\right\|^2 - \left\|w - \frac{1}{\gamma^2}D^*\Pi_2 x\right\|^2\right)\right. \\ &\qquad\qquad \left. + \|u + B^*\Pi_2 x\|^2 - \|u + B^*\Pi_1 x\|^2\right\} dt.\end{aligned}$$

Setting $u = -B^*(\frac{\Pi_1+\Pi_2}{2})x, w = \gamma^{-2}D^*(\frac{\Pi_1+\Pi_2}{2})x$, we have

$$\langle(\Pi_1 - \Pi_2)x_0, x_0\rangle = \langle(\Pi_1 - \Pi_2)x(T), x(T)\rangle. \tag{A2.2}$$

Next we claim that the system

$$\dot{x} = Ax - BB^*\left(\frac{\Pi_1+\Pi_2}{2}\right)x + \frac{1}{\gamma^2}DD^*\left(\frac{\Pi_1+\Pi_2}{2}\right)x, \qquad x(0)=x_0 \tag{A2.3}$$

is exponentially stable. Let $\mathcal{A}_1 = \frac{1}{2}(A - BB^*\Pi_1 + \frac{1}{\gamma^2}DD^*\Pi_1), \mathcal{A}_2 = \frac{1}{2}(A - BB^*\Pi_2 + \frac{1}{\gamma^2}DD^*\Pi_2)$. Note that both $\mathcal{A}_1$ and $\mathcal{A}_2$ are exponentially stable according to the assumption at the beginning. Thus by the generalized Lyapunov's theorem [11] there exist two (unique) self-adjoint, positive definite, bounded operators $P_1, P_2 \in \mathcal{L}(X)$ such that

$$\langle P_1\mathcal{A}_1x, x\rangle + \langle x, P_1\mathcal{A}_1x\rangle = -\langle x, x\rangle \qquad \forall x \in D(A)$$

and

$$\langle P_2\mathcal{A}_1x, x\rangle + \langle x, P_2\mathcal{A}_1x\rangle = -\langle x, x\rangle \qquad \forall x \in D(A).$$

It is known that $\langle x, y\rangle_{e_1} := \langle x, P_1y\rangle, \langle x, y\rangle_{e_2} := \langle x, P_2y\rangle$ define two equivalent inner products on X such that

$$\begin{aligned} \alpha_1\|x\|_X^2 \le \|x\|_{e_1}^2 \le \beta_1\|x\|_X^2, \qquad \alpha_1 > 0,\\ \alpha_2\|x\|_X^2 \le \|x\|_{e_2}^2 \le \beta_2\|x\|_X^2, \qquad \alpha_2 > 0. \end{aligned}$$

Thus from (A2.3), for $x_0 \in D(A)$, we have

$$\begin{aligned} &\frac{d}{dt}\|x\|^2 = \langle \dot{x}, x\rangle + \langle x, \dot{x}\rangle \\ &= \langle \mathcal{A}_1x, x\rangle + \langle x, \mathcal{A}_1x\rangle + \langle \mathcal{A}_2x, x\rangle + \langle x, \mathcal{A}_2x\rangle \\ &\le \alpha_1^{-1}(\langle \mathcal{A}_1x, x\rangle_{e_1} + \langle x, \mathcal{A}_1x\rangle_{e_1}) + \alpha_2^{-1}(\langle \mathcal{A}_2x, x\rangle_{e_2} + \langle x, \mathcal{A}_2x\rangle_{e_2}) \\ &= -(\alpha_1^{-1} + \alpha_2^{-1})\langle x, x\rangle, \end{aligned}$$

which implies that $x \in L^2(0,\infty; X)$. Hence (A2.2) is exponentially stable.

Now in (A2.2), letting $T \to \infty$ yields

$$\langle (\Pi_1 - \Pi_2)x_0, x_0\rangle = 0 \qquad \forall x_0 \in X,$$

which implies $\Pi_1 = \Pi_2$. □

Reasoning similar to that given above leads to the next proposition.

PROPOSITION A.2. *The Riccati equation*

$$\Sigma A^* + A\Sigma - \Sigma\left(C^*C - \frac{1}{\gamma^2}H^*H\right)\Sigma + DD^* = 0$$

has at most one positive solution $\Sigma \in \Sigma^+(X)$ such that $A^* - (C^*C - \frac{1}{\gamma^2}H^*H)\Sigma$ is exponentially stable.

PROPOSITION A.3. *Suppose that the Riccati equation*

$$\Pi A + A^*\Pi - \Pi\left(BB^* - \frac{1}{\gamma^2}DD^*\right)\Pi + H^*H = 0 \tag{A2.4}$$

has a solution $\Pi \in \Sigma^+(X)$ *such that* $A - (BB^* - \frac{1}{\gamma^2}DD^*)\Pi$ *is exponentially stable and that* (A, H) *is detectable. Then, for any* $t_f > 0$*, the Riccati equation*

$$\begin{aligned}&\dot{\Pi}(t) + \Pi(t)A + A^*\Pi(t) - \Pi(t)\left(BB^* - \frac{1}{\gamma^2}DD^*\right)\Pi(t) + H^*H = 0,\\ &\Pi(t_f) = 0,\end{aligned} \tag{A2.5}$$

has a unique solution $\Pi(t; t_f) \in \Sigma^+(X)$ *on* $[0, t_f]$ *with the property that* $\Pi(t; t_f) \leq \Pi$ *and for all* $x \in X$ *the limit*

$$\Pi x = \lim_{t_f \to \infty} \Pi(t; t_f)x$$

exists.

Proof. Consider the following optimization problem:

$$\varphi(s, x_0) = \sup_w \inf_u J_\gamma(x_0; u, w) = \sup_w \inf_u \int_s^{t_f} (\|Hx\|^2 + \|Bu\|^2 - \gamma^2\|w\|)dt,$$

where x is subject to

$$\dot{x}(s) = Ax(s) + Bu(s) + Dw(s), \quad x(t) = x_0. \tag{A2.6}$$

Note that the conditions of this proposition guarantee that $\varphi(t, x_0)$ is finite. According to Theorem 2.1 of [29], we have that

$$\varphi(t, x_0) = J_\gamma(x_0; u^*, w^*), \tag{A2.7}$$

where $u^* = B^*p,\ w^* = -\frac{1}{\gamma^2}D^*p$, and p is generated by

$$\begin{aligned}\dot{x} &= Ax + BB^*p - \frac{1}{\gamma^2}DD^*p, \quad x(t) = x_0,\\ \dot{p} &= -A^*p + H^*Hx, \quad p(t_f) = 0.\end{aligned} \tag{A2.8}$$

Thus the operator $x_0 \to -p = \partial\varphi(t, x_0)$ is linear and therefore self-adjoint on X (see example 2 in Chapter 2 of [5]). Moreover,

$$\varphi(t, x) = \langle P(t)x, x\rangle \quad \text{for} \quad x \in X, P(t) = \partial\varphi(t). \tag{A2.9}$$

Making use of (A2.6) and (A2.9), we have

$$\begin{aligned}\frac{d}{ds}\langle P(s)x(s), x(s)\rangle =& -\langle H^*Hx(s), x(s)\rangle - \langle P(s)BB^*P(s)x(s), x(s)\rangle\\ &+ \frac{1}{\gamma^2}\langle DD^*P(s)x(s), x(s)\rangle.\end{aligned}$$

Alternatively, from (A2.8),

$$\begin{aligned}&\frac{d}{ds}\langle P(s)x(s), x(s)\rangle\\ &= \langle \dot{P}(s)x(s), x(s)\rangle + \langle Ax(s), x(s)\rangle + \langle A^*x(s), x(s)\rangle\\ &\qquad + 2\langle P(s)BB^*P(s)x(s), x(s)\rangle - \frac{2}{\gamma^2}\langle DD^*P(s)x(s), x(s)\rangle.\end{aligned}$$

Therefore, from the last two equalities we obtain

$$\begin{aligned}&\dot{P}(t) + P(t)A + A^*P(t) - P(t)\left(BB^* - \frac{1}{\gamma^2}DD^*\right)P(t) + H^*H = 0,\\ &P(t_f) = 0.\end{aligned} \tag{A2.10}$$

Thus, $P(t)$ solves (A2.5). Let $\Pi(t;t_f) = P(t)$. From the above discussion we have

$$\langle \Pi(t;t_f)x_0, x_0\rangle = \sup_w \inf_u J_\gamma^{t_f}(x_0;u,w).$$

Note that $\langle \Pi(t;t_f)x_0, x_0\rangle$ is monotonically nondecreasing with increasing t_f, since the lower value of the game $J^{t_f}(x_0;\cdot,\cdot)$ defined on $[t,t_f]$ cannot be larger than that of the one defined on a longer interval, $[t,t_{f'}]$, $t_{f'} > t_f$, as the maximizing player can always play zero control on the subinterval $[t_f, t_{f'}]$. Let $\hat{w} \in \mathcal{W}$ be picked to be zero beyond $t = t_f$. Denoting all admissible feedback controls for (A2.6) by $\mathcal{M}$, we have the following estimate:

$$\begin{aligned}\langle \Pi(t;t_f)x_0, x_0\rangle &= \sup_w \inf_u J_\gamma^{t_f}(x_0;u,w)\\ &\le \inf_{\mu\in\mathcal{M}} \sup_w J_\gamma^\infty(x_0;u,\hat{w})\\ &\le \inf_{\mu\in\mathcal{M}} \sup_w J_\gamma^\infty(x_0;u,w) = \langle \Pi x_0, x_0\rangle.\end{aligned}$$

By the parallelogram law in a Hilbert space, we have that $\lim_{t\to\infty}\langle \Pi(t;t_f)x, y\rangle$ exists for any $x, y \in X$. Thus we can define $R \in \Sigma^+(X)$ such that

$$\lim_{t\to\infty}\langle \Pi(t;t_f)x, y\rangle = \langle Rx, y\rangle.$$

It can be verified by employing the standard argument (cf. [6]) that R in fact solves (A2.4).

Next we claim that $R = \Pi$, i.e., $\Pi(t;t_f)x \to \Pi x$ $\forall x \in X$: Since $A - (BB^* - \frac{1}{\gamma^2}DD^*)\Pi := K$ is exponentially stable, there exist $M > 0, \beta > 0$ such that

$$\|e^{Kt}\| \le Me^{-\beta t}.$$

Set $\Xi(t) = \Pi - \Pi(t;t_f)$; then Ξ is the mild solution to the Riccati equation

$$\begin{aligned}&\dot{\Xi} + K^*\Xi + \Xi K + \Xi\left(BB^* - \frac{1}{\gamma^2}DD^*\right)\Xi = 0,\\ &\Xi(t_f) = \Pi,\end{aligned} \tag{A2.11}$$

which is equivalent to the equation

$$\begin{aligned}(\Pi - \Pi(t;t_f))x =& e^{(t_f-t)K^*}\Pi e^{(t_f-t)K}\\ &+ \int_t^{t_f} e^{(s-t)K^*}\Xi(s)\left(BB^* - \frac{1}{\gamma^2}DD^*\right)\Xi(s)e^{(s-t)K}x ds\end{aligned} \tag{A2.12}$$

for any $x \in X$. Since for any $x \in X$, $\{\Pi(t;t_f)x\}$ is bounded, there exists a $C > 0$ such that

$$\|\Pi(t;t_f)\| \le C \qquad \text{for} \quad 0 \le t \le t_f \le \infty \tag{A2.13}$$

by the principle of uniform boundedness. Therefore (A2.12) and (A2.13) give

$$\begin{aligned}\|(\Pi - \Pi(t;t_f))x\| &\leq M^2 e^{-2\beta(t_f-t)}\|\Pi\|\|x\| \\ &\quad + M^2 \int_t^{t_f} e^{-\beta(s-t)}\|(\Pi - \Pi(t;t_f))x\| \left\|BB^* - \frac{1}{\gamma^2}DD^*\right\| (\|\Pi\| + C)e^{-\beta(s-t)}ds.\end{aligned}$$

By employing Gronwall's inequality, we obtain

$$\begin{aligned}&\|(\Pi - \Pi(t;t_f))x\| \\ &\quad\leq M^2 e^{-2\beta(t_f-t)}\|\Pi\|\|x\| \exp\left(M^2 \int_t^{t_f} \left\|BB^* - \frac{1}{\gamma^2}DD^*\right\| (\|\Pi\| + C)e^{-\beta(s-t)}ds\right) \\ &\quad\leq M^2 e^{-2\beta(t_f-t)}\|\Pi\|\|x\| \exp\left(M^2 \left\|BB^* - \frac{1}{\gamma^2}DD^*\right\| (\|\Pi\| + C)/\beta\right),\end{aligned}$$

which completes the proof. □

Similarly, we have the following.

PROPOSITION A.4. *Suppose that the Riccati equation*

$$\Pi^N A^N + A^{N*}\Pi^N - \Pi^N \left(B^N B^{N*} - \frac{1}{\gamma^2}D^N D^{N*}\right)\Pi^N + H^{N*}H^N = 0 \tag{A2.14}$$

has a positive definite solution Π^N *such that* $A^N - B^N B^{N*} + \frac{1}{\gamma^2}D^N D^{N*}$ *is exponentially stable. Then, for any* $t_f > 0$*, the Riccati equation*

(A2.15)
$$\begin{aligned}&\dot{\Pi}^N + \Pi^N A^N + A^{N*}\Pi^N - \Pi^N \left(B^N B^{N*} - \frac{1}{\gamma^2}D^N D^{N*}\right)\Pi^N + H^{N*}H^N = 0, \\ &\Pi^N(t_f;t_f) = 0,\end{aligned}$$

has a unique positive definite solution $\Pi^N(t;t_f)$ *with the property that* $\Pi^N(t;t_f) \leq \Pi$ *for* $t \in [0,\infty)$*, and for all* $x \in X$ *the limit*

$$\Pi^N x = \lim_{t_f\to\infty} \Pi^N(t;t_f)x$$

exists.

PROPOSITION A.5. *Suppose that both Riccati equations*

(A2.16)
$$\Pi^N A^N + A^{N*}\Pi^N - \Pi^N \left(B^N B^{N*} - \frac{1}{\gamma^2}D^N D^{N*}\right)\Pi^N + H^{N*}H^N = 0, \quad N = 1,2,\ldots,$$

and

$$\Pi A + A^*\Pi - \Pi\left(BB^* - \frac{1}{\gamma^2}DD^*\right)\Pi + H^*H = 0 \tag{A2.17}$$

have solutions $\Pi^N \in \Sigma^+(X^N)$, $\Pi \in \Sigma^+(X)$*, such that* $A^N - (B^N B^{N*} - \frac{1}{\gamma^2}D^N D^{N*})\Pi^N$ *and* $A - (BB^* - \frac{1}{\gamma^2}DD^*)\Pi$ *are both exponentially stable. For any* $t_f > 0$*, let* $\Pi^N(t;t_f)$*,* $\Pi(t;t_f)$ *be the solutions of the Riccati equations*

$$\begin{aligned}&\dot{\Pi}^N(t) + \Pi^N(t)A^N + A^{N*}\Pi^N(t) - \Pi^N(t)\left(B^N B^{N*} - \frac{1}{\gamma^2}D^N D^{N*}\right)\Pi^N(t) \\ &\qquad + H^{N*}H^N = 0, \\ &\Pi^N(t_f) = 0, \quad N = 1,2,\ldots,\end{aligned}$$

and

$$\begin{aligned}&\dot{\Pi}(t) + \Pi(t)A + A^*\Pi(t) - \Pi(t)\left(BB^* - \frac{1}{\gamma^2}DD^*\right)\Pi(t) + H^*H = 0,\\ &\Pi(t_f) = 0,\end{aligned}$$

respectively, where for convenience we suppress the dependence on the terminal time t_f. Then, for each $t_f > 0$,

$$\Pi^N(t;t_f) \to\!\!\to \Pi(t;t_f) \quad \textit{uniformly in } t \textit{ on} \quad [0,t_f] \quad \textit{as} \quad N \to \infty.$$

Proof. Consider the optimization problem

$$\sup_w \inf_u J_\gamma^N(x_0^N;u,w) = \sup_w \inf_u \int_\tau^{t_f} (\|H^N x^N\|^2 + \|u\|^2 - \gamma^2\|w\|^2)dt,$$

where x^N is subject to

$$\begin{aligned}\dot{x}^N &= A^N x + B^N u + D^N w,\\ x^N(\tau) &= x_0^N,\end{aligned} \tag{A2.18}$$

which is equivalent to

$$x^N(t) = e^{A^N(t-\tau)}x_0^N + \int_\tau^t e^{A^N(t-s)}\big(B^N u(s) + D^N w(s)\big)ds.$$

Let us denote $L_2(0,t_f;X^N)$ by $\mathcal{X}^N$, $L_2(0,t_f;U)$ by $\mathcal{U}$, and $L_2(0,t_f;W)$ by $\mathcal{W}$. Define $\mathcal{F}^N \in \mathcal{L}(\mathcal{X}^N,\mathcal{X}^N)$ by

$$(\mathcal{F}^N\phi)(s) = \int_\tau^t e^{A^N(t-s)}\phi(s)ds, \qquad \phi \in \mathcal{X}^N.$$

By identifying X^N, U with respective duals, it is easy to verify that

$$(\mathcal{F}^{N*}\phi)(s) = \int_t^{t_f} e^{A^{N*}(s-t)}\phi(s)ds, \qquad \phi \in \mathcal{X}^N,$$

where $\mathcal{F}^{N*} \in \mathcal{L}(\mathcal{X}^N,\mathcal{X}^N)$. Next we write the cost function J_γ^N as

(A2.19)
$$J_\gamma^N(x_0^N;u,w) = \|H^N(e^{A^N\cdot}x_0^N + \mathcal{F}^N(B^N u + D^N w))\|_{\mathcal{X}^N}^2 + \|u\|_{\mathcal{U}}^2 - \gamma^2\|w\|_{\mathcal{W}}^2.$$

For arbitrary but fixed $w \in \mathcal{W}$, there exists a unique u which minimizes J_γ^N, which further is the unique solution of

$$\delta_u J_\gamma^N(x_0^N;u,w)(v) = 0 \quad \forall v \in \mathcal{U}, \tag{A2.20}$$

where $\delta_u J_\gamma^N(x_0^N;u,w)(v)$ stands for the Gâteaux derivative of J_γ^N at u, applied to v. From (A2.19), we have

$$\begin{aligned}&\delta J_\gamma^N(x_0^N;u,w)(v)\\ &\quad = 2\langle H^{N*}H^N\big(e^{A^N\cdot}x_0^N + \mathcal{F}(B^N u + D^N w)\big), \mathcal{F}B^N v\rangle_{\mathcal{X}^N} + 2\langle u,v\rangle_{\mathcal{U}}\\ &\quad = 2\langle T_1^N u + T_2^N w + T_3^{N*}x_0^N, v\rangle,\end{aligned}$$

where

$$T_1^N = I + B^{N*}\mathcal{F}^{N*}H^{N*}H^N\mathcal{F}^N B^N \in \mathcal{L}(\mathcal{U},\mathcal{U}),$$

$$T_2^N = B^{N*}\mathcal{F}^{N*}H^{N*}H^N\mathcal{F}^N D^N \in \mathcal{L}(\mathcal{W},\mathcal{U}),$$

and

$$T_3^{N*} = B^{N*}\mathcal{F}^{N*}H^{N*}H^N e^{A^N\cdot} \in \mathcal{L}(X^N,\mathcal{U}).$$

According to (A2.20), a necessary and sufficient condition for u to be optimal for a fixed w is that

$$u^w(t) = -\big(T_1^{-N}(T_3^{N*}x_0^N + T_2^N w)\big)(t) \quad \text{a.e. in } [0,t_f],$$

where we define $T_1^{-N} = (T_1^N)^{-1}$. Hence we have

(A2.21)
$$\begin{aligned} &J_\gamma^N(x_0^N;u^w,w)\\ &= \|H^N\big(e^{A^N\cdot}x_0^N - \mathcal{F}^N B^N T_1^{-N}T_3^{N*}x_0^N + \mathcal{F}^N((-B^N T_1^{-N}T_2^N + D^N)w)\big)\|^2_{\mathcal{X}^\mathcal{N}}\\ &\quad + \|T_1^{-N}(T_3^{N*}x_0^N + T_2^N w)\|^2_{\mathcal{U}} - \gamma^2\|w\|^2_{\mathcal{W}}. \end{aligned}$$

We now study the problem of maximizing the expression $J_\gamma^N(x_0^N;u^w,w)$ for $w \in \mathcal{W}$. Note that although we have a linear quadratic problem, the concavity is not verified a priori. However, the Riccati equation (A2.16) provides the γ-attenuation level for system (A2.18) with cost function J_γ^N, and one can show that $w \to J_\gamma^N(x_0^N;u^w,w)$ is strictly concave. Thus the maximizer w will satisfy the equation

$$\delta_w J_\gamma^N(x_0^N;u^w,w)(m) = 0 \quad \forall m \in \mathcal{W}. \tag{A2.22}$$

Direct calculation results in

$$\delta_w J_\gamma^N(x_0^N;u^w,w)(m) = 2\langle \mathcal{T}_1^N w + \mathcal{T}_2^N x_0^N, m\rangle \qquad \forall m \in \mathcal{W}, \tag{A2.23}$$

where $\mathcal{T}_1^N \in \mathcal{L}(\mathcal{W},\mathcal{W})$ is given by

$$\begin{aligned} \mathcal{T}_1^N &= [\mathcal{F}^N(-B^N T_1^{-N}T_2^N + D^N)]^* H^{N*}H^N[\mathcal{F}^\mathcal{N}(-B^N T_1^{-N}T_2^N + D^N)]\\ &\quad + T_2^{N*}T_1^{-N*}T_1^{-N}T_2^N - \gamma^2 I \end{aligned}$$

and $\mathcal{T}_2^N \in \mathcal{L}(\mathcal{X}^N,\mathcal{W})$ is given by

$$\begin{aligned} \mathcal{T}_2^N &= [\mathcal{F}^N(-B^N T_1^{-N}T_2^N + D^N)]^* H^{N*}H^N(e^{A^N\cdot} - \mathcal{F}^N B^N T_1^{-N}T_3^{N*})\\ &\quad + T_3^N T_1^{-N*}T_1^{-N}T_3^{N*}. \end{aligned}$$

Recalling the proof of the first part of Proposition A.3, we know that

$$\langle \Pi^N(\tau)x_0^N, x_0^N\rangle = \sup_w \inf_u J_\gamma^N(x_0^N;u,w).$$

Thus, letting $x_0^N = 0$, we have

$$\sup_w \inf_u J_\gamma^N(0;u,w) = 0.$$

From (A2.21) we have

$$J_\gamma^N(0; u^w, w) = \langle \mathcal{T}_1^N w, w\rangle.$$

Since $w \to J_\gamma^N(0; u^w, w)$ is strictly concave, $w = 0$ is the only maximizer of the differential game $J_\gamma^N(0; u^w, w)$, which implies that $-\mathcal{T}_1^N$ is a positive definite operator of $\mathcal{L}(\mathcal{W}, \mathcal{W})$. Thus from (A2.19) we have

$$w = (-\mathcal{T}_1^N)^{-1}\mathcal{T}_2 x_0^N. \tag{A2.24}$$

Hence by (A2.20), (A2.21), and (A2.24), we have

$$\sup_w J_\gamma^N(x_0^N; u^w, w) = \langle (\mathcal{K}^{N*}\mathcal{K}^N + T_3^N T_1^{-N*} T_1^{-N} T_3^{N*} + \mathcal{T}_2^{N*}(-\mathcal{T}_1)^{-N}\mathcal{T}_2^N)x_0^N . x_0^N\rangle_{\mathcal{X}^N},$$

where

$$\mathcal{K}^N = H^N(e^{A^N\cdot} - \mathcal{F}^N B^N T_1^{-N} T_3^{N*}).$$

In view of this, we obtain

$$\Pi^N(\tau)x_0^N = (\mathcal{K}^{N*}\mathcal{K}^N + T_3^N T_1^{-N*} T_1^{-N} T_3^{N*} + \mathcal{T}_2^{N*}(-\mathcal{T}_1)^{-N}\mathcal{T}_2^N)x_0^N. \tag{A2.25}$$

Define $\mathcal{F}$, T_1, T_2, T_3 by

$$\begin{aligned}
&(\mathcal{F}\phi)(s) = \int_0^t e^{A(t-s)}\phi(s)ds, \qquad \phi \in \mathcal{X},\\
&T_1 = I + B^*\mathcal{F}^*H^*H\mathcal{F}B \in \mathcal{L}(\mathcal{U},\mathcal{U}),\\
&T_2 = B^*\mathcal{F}^*H^*H\mathcal{F}D \in \mathcal{L}(\mathcal{W},\mathcal{U}),\\
&T_3^* = B^*\mathcal{F}^*H^*He^{A\cdot} \in \mathcal{L}(X,\mathcal{U})
\end{aligned}$$

and $\mathcal{T}_1 \in \mathcal{L}(\mathcal{W}, \mathcal{W})$ and $\mathcal{T}_2 \in \mathcal{L}(\mathcal{X}, \mathcal{W})$, respectively, by

$$\begin{aligned}
\mathcal{T}_1 = {} & [\mathcal{F}(-BT_1^{-1}T_2 + D)]^*H^*H[\mathcal{F}(-BT_1^{-1}T_2 + D)]\\
& + T_2^*T_1^{-1*}T_1^{-1}T_2 - \gamma^2 I,
\end{aligned}$$

$$\begin{aligned}
\mathcal{T}_2 = {} & [\mathcal{F}(-BT_1^{-1}T_2 + D)]^*H^*H(e^{A\cdot} - \mathcal{F}BT_1^{-1}T_3^*)\\
& + T_3T_1^{-*}T_1^{-1}T_3^*.
\end{aligned}$$

Consider the optimization problem

$$J_\gamma(x_0; u, w) = \sup_w \inf_u \int_\tau^{t_f} (\|x\|^2 + \|u\|^2 - \gamma^2\|w\|^2)dt,$$

where x is generated by

$$\dot{x} = Ax + Bu + Dw, \qquad x(\tau) = x_0.$$

Then, following a discussion similar to the one that led to (A2.25),

$$\Pi(0)x_0 = (\mathcal{K}^*\mathcal{K} + T_3(T_1^{-1})^*T_1^{-1}T_3^* + \mathcal{T}_2^*(-\mathcal{T}_1)^{-1}\mathcal{T}_2)x_0,$$

where

$$\mathcal{K} = H(e^{A\cdot} - \mathcal{F}BT_1^{-1}T_3^*).$$

By assumptions (A1)–(A4), as $N \to \infty$ we have

$$\begin{aligned}
&e^{A^N t} \rightarrow\rightarrow e^{At} \qquad \text{uniformly on} \qquad [0, t_f],\\
&e^{A^{N*} t} \rightarrow\rightarrow e^{A^* t} \qquad \text{uniformly on} \quad [0, t_f],\\
&B^N \to B \qquad \text{strongly},\\
&D^N \to D \qquad \text{strongly},\\
&H^N \to H \qquad \text{strongly}.
\end{aligned}$$

Letting $x_0^N := P^N x_0$, the above discussion indicates that

$$\Pi^N(\tau) P^N x_0 \to \Pi(\tau) x_0 \quad \forall x_0 \in X \qquad \text{as} \quad N \to \infty,$$

and this completes the proof of Proposition A.5. □

Proof of Lemma 2.1. For each $\varepsilon > 0$ and $x \in X$, according to Propositions A.3 and A.4 there exists $t_f > 0$ such that

$$\begin{aligned}
&\|\Pi(0; t_f)x - \Pi x\| < \varepsilon,\\
&\|\Pi^N(0, t_f)P^N - \Pi^N P^N\| < \varepsilon \qquad \text{for each} \quad N = 1, 2, \dots .
\end{aligned}$$

Note that

$$\begin{aligned}
\|\Pi^N P^N x - \Pi x\| \le &\|\Pi^N P^N x - \Pi(0; t_f)^N P^N x\| + \|\Pi(0; t_f)^N P^N x - \Pi^N(0; t_f)x\|,\\
&\|\Pi^N(0; t_f)x - \Pi(0; t_f)x\| + \|\Pi(0; t_f)x - \Pi x\|\\
&\le \varepsilon\|x\| + \|\Pi^N(0; t_f)\|\|P^N x - x\| + \|\Pi^N(0; t_f)x - \Pi(0; t_f)x\| + \varepsilon.
\end{aligned}$$

Since $\|\Pi^N(0; t_f)\|$ is uniformly bounded in N by uniform boundedness theorem, and $\Pi^N(0; t_f)x \to \Pi(0; t_f)x$ by Proposition A.5, we thus have

$$\Pi^N \rightarrow\rightarrow \Pi \qquad \text{as} \quad N \to \infty,$$

and this completes the proof of Lemma 2.1. □

By using the duality theorem, Lemma 2.2 can be proven similarly.

REFERENCES

[1] T. Başar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Design Problems*, Birkhäuser, Boston, 1995.

[2] H.T. Banks, M.A. Demetriou, and R.C. Smith, *An H^∞/minmax periodic control in a two-dimensional structural acoustic model with piezoceramic actuators*, IEEE Trans. Automat. Control, 41 (1996), pp. 943–959.

[3] A. Bensoussan and P. Bernhard, *On the standard problem of H^∞-optimal control for infinite dimensional systems*, in Identification and Control in Systems Governed by Partial Differential Equations, South Hadley, MA, July 1992, Proc. Appl. Math. 68, H.T. Banks, R.H. Fabiano, and K Ito, eds., SIAM, Philadelphia, 1993, pp. 117–140.

[4] H.T. Banks and K. Kunisch, *The linear regulator problem for parabolic systems*, SIAM J. Control Optim., 22 (1984), pp. 684–698.

[5] V. Barbu and Th. Precupanu, *Convexity and Optimization in Banach Spaces*, D. Reidel, Dordrecht, 1986.

[6] A. Bensoussan, G.D. Prato, M.C. Delfour, and S.K. Mitter, *Representation and Control of Infinite Dimensional Systems*, Vol. I, Birkhäuser, Boston, MA, 1992.

[7] R.F. Curtain, *H_∞-control for distributed parameter systems: A survey*, in Proceedings of the 29th IEEE Conference on Decision and Control, Vol. 1, 1990, pp. 22–26.

[8] R.F. Curtain and D. Salamon, *Finite dimensional compensators for parabolic distributed systems with unbounded control and observation*, SIAM J. Control Optim., 24 (1984), pp. 255–276.

[9] R. Curtain and A.J. Pritchard, *The infinite-dimensional Riccati equation for systems defined by evolution operators*, SIAM J. Control Optim., 14 (1976), pp. 951–983.

[10] R.F. Curtain and Y. Zhou, *A weighted mixed-sensitivity H_∞-control design for irrational transfer matrices*, IEEE Trans. Automat. Control, 41 (1996), pp. 1312–1321.

[11] R. Datko, *Extending a theorem of A. M. Liapunov to Hilbert space*, J. Math. Anal. Appl., 32 (1970), pp. 610–616.

[12] J.S. Gibson, *Linear-quadratic optimal control of hereditary differential systems: Infinite dimensional Riccati equations and numerical approximations*, SIAM J. Control Optim., 21 (1983), pp. 95–139.

[13] J.S. Gibson, *The Riccati integral equations for optimal control problems on Hilbert spaces*, SIAM J. Control Optim., 17 (1979), pp. 537–565.

[14] J.S. Gibson and A. Adamian, *Approximation theory for linear-quadratic-Gaussian optimal control of flexible structures*, SIAM J. Control Optim., 29 (1991), pp. 1–37.

[15] K. Ito, *Finite-dimensional compensators for infinite-dimensional systems via Galerkin-type approximation*, SIAM J. Control Optim., 28 (1990), pp. 1251–1269.

[16] K. Ito and K.A. Morris, *An approximation theory of solutions to operator Riccati equations for H^∞ control*, SIAM J. Control Optim., 36 (1998), pp. 82–99.

[17] K. Ito, F. Kappel, and D. Salamon, *A variational approach to approximation of delay systems*, Differential Integral Equations, 4 (1991), pp. 51–72.

[18] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, New York, 1980.

[19] O.A. Ladyzhenskaya, *The Boundary Value Problems of Mathematical Physics*, Springer-Verlag, New York, 1985.

[20] H. Özbay, *Controller reduction in the two-block H^∞-optimal design for distributed plants*, Internat. J. Control, 54 (1991), pp. 1291–1308.

[21] H. Özbay and A. Tannenbaum, *On the structure of suboptimal H^∞ controllers in the sensitivity minimization problem for distributed stable plants*, Automatica J. IFAC, 27 (1991), pp. 293–305.

[22] H. Özbay and A. Tannenbaum, *A skew Toeplitz approach to H^∞ optimal control of multivariable distributed systems*, SIAM J. Control Optim., 28 (1990), pp. 653–670.

[23] C.V. Pao, *Semigroups and asymptotic stability of nonlinear differential equations*, SIAM J. Math. Anal., 3 (1972), pp. 371–379.

[24] A. Pazy, *Semigroup of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[25] Y. Sakawa, *Feedback control of second order evolution equations with damping*, SIAM J. Control Optim., 22 (1984), pp. 343–361.

[26] J.M. Schumacher, *A direct approach to compensator design for distributed parameter systems*, SIAM J. Control Optim., 21 (1983), pp. 823–836.

[27] A. Taylor and D. Lay, *Introduction to Functional Analysis*, John Wiley & Sons, New York, 1980.

[28] R.B. Vinter, *Filter stability for stochastic evolution equations*, SIAM J. Control Optim., 15 (1977), pp. 465–485.

[29] M. Xiao and T. Başar, *H^∞ control of a class of infinite-dimensional linear systems with nonlinear outputs*, in Annals of Dynamic Games, Vol. 5, V. Gaitsgory and J. Filar, eds., Birkhäuser, Boston, 1999, pp. 38–63.

[30] M. Xiao and T. Başar, *Solutions to Generalized Riccati Evolution Equations and H^∞ Optimal Control Problems on Hilbert Spaces*, Internal report, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, 1997.

[31] J. Zabczyk, *Remarks on the algebraic Riccati equation in Hilbert space*, Appl. Math. Optim., 2 (1976), pp. 251–258.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 6, pp. 1649–1675

DYNAMIC DOMAIN DECOMPOSITION OF OPTIMAL CONTROL PROBLEMS FOR NETWORKS OF STRINGS AND TIMOSHENKO BEAMS*

G. LEUGERING†

Abstract. We consider general networks of strings and/or Timoshenko beams. We apply controls at boundary nodes of the network and want to minimize some cost function along (part of) the structure. Optimality systems for the entire structure are far too complex to compute in reasonable time. In particular, in real-time applications one wants to reduce the size of the problem. Thus, dynamic decomposition into its physical elements appears to be a natural approach. We show how to iteratively decompose the global optimality system into a system related to a substructure. Then we interpret the local system as an optimality system corresponding to an optimal control problem for the substructure and finally we show convergence of the "outer" iteration.

Key words. dynamic domain decomposition, networks of strings and beams, saddle-point iteration, relaxation

AMS subject classifications. 49J20, 49K20, 49M27, 65N55

PII. S0363012997331986

1. Introduction.

We consider a connected network of smooth curves in $\mathbb{R}^3$ indexed by $i = 1 : n_e$. The curves are joined together at vertices $\{v_J, v_M, \dots\} = V$ and are representative of the physical location of strings or of centerlines of beams in their reference configuration. Each of those curves will undergo deformations which we describe by $r_i(x,t) : [0,l_i] \times [0,T] \longrightarrow \mathbb{R}^{p_i}$. Let v_J be a vertex and d_J the number of arcs adjacent to v_J, and let $\mathcal{E}_J$ be the set of indices corresponding to adjacent arcs at v_J, i.e., $d_J = |\mathcal{E}_J|$. In particular, we consider directed graphs, such that an edge with label i joining vertices v_J and v_M in the direction $v_J \ \to v_M$ is parametrized in such a way that $x = 0$ corresponds to the vertex v_J, while $x = \ell_i$ corresponds to v_M. For the sake of convenience, however, we will write $r_i(v_J,t)$ instead of $r_i(0,t)$ or $r_i(\sigma_i(0),t)$, where σ_i is the mapping from $[0,\ell_i]$ onto the edge e_i. With this convention in mind we define the edge-node incidence relation ϵ as

$$\epsilon_{iJ} = \begin{cases} -1 & \text{if } e_i \text{ starts at } v_J, \\ 1 & \text{if } e_i \text{ ends at } v_J, \\ 0 & \text{if } e_i \text{ is not incident at } v_J. \end{cases}$$

At a given vertex v_J the deformation $r_i(v_J,t)$ of arc i has $p_i \in \mathbb{N}$ degrees of freedom. At a given vertex v_J, such that arc i is adjacent to v_J, we suppose that $q_J (q_J \leq p_i)$ of these variables are geometrically constrained. At simple nodes, where $d_J = 1$, we always have $q_J = p_i$ (for $\epsilon_{iJ} \neq 0$). At multiple nodes, where $d_J > 1$, we assume that we have surjective linear mappings $C_{iJ} : \mathbb{R}^{p_i} \to \mathbb{R}^{q_J}$ $(p_i \geq q_J)$ such that

$$\lambda_J(t) := C_{iJ} r_i(v_J,t) = C_{jJ} r_j(v_J,t) \qquad \forall\, i,j \in \mathcal{E}_J,\ t \in (0,T).$$

[*] Received by the editors December 31, 1997; accepted for publication (in revised form) October 29, 1998; published electronically September 1, 1999. This research was supported by the Deutsche Forschungsgemeinschaft DFG under grant Le 595-9-2.

http://www.siam.org/journals/sicon/37-6/33198.html

[†] Institute of Mathematics, University of Bayreuth, D-95440, Bayreuth, Germany (leugering@uni-bayreuth.de).

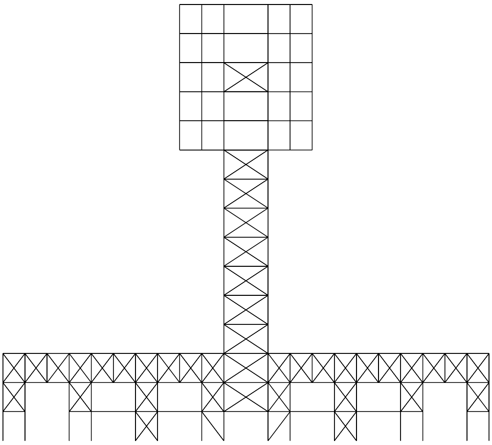


FIG. 1.1. *A typical multibay frame.*

Note that each arc carries along a local base so that C_{iJ} can be expressed as a matrix with respect to these bases. This requirement says that some (combinations of) variables are continuous across the joint v_J. For example, if $C_{iJ} = I \in \mathbb{R}^{p_i \times p_i}$, then all displacements are continuous across v_J, which for beams signifies a rigid joint. In particular, if one considers for simplicity a planar beam network with u_i, w_i, ψ_i as longitudinal, vertical, and shear displacements, then projecting with C_{iJ} onto the first two variables at v_J describes a pin joint, whereas $C_{iJ} = I_{3\times 3}$ describes a rigid joint. We refer the reader to [14], [15] (see also [27], [28] for "scalar" networks). As C_{iJ} has rank q_J, there is a right inverse C_{iJ}^+ such that $C_{iJ}C_{iJ}^+ = I$. Let Π_{iJ} denote the orthogonal projection onto the kernel of C_{iJ} and $\Pi_{iJ}^\perp$ the orthogonal projection onto $(\mathrm{ker} C_{iJ})^\perp = \mathrm{Im}\, C_{iJ}^T$, i.e., $C_{iJ}^+ C_{iJ} = \Pi_{iJ}^\perp$.

A typical situation is shown on a qualitative level in Figure 1.1. The graph shown there might be regarded as a part of a more complex space structure. Rather than decompose such a (sub-)structure into each individual element, one would use homogenization in order to replace the periodic substructures by dynamically equivalent homogenous elements and apply the decomposition principle to the structure composed of those homogenized structural elements. It is important to note that by this procedure one can deal with systems having local joint dynamics, where substructures interfere at common nodes. We might therefore include rigid bodies at those joints, as well as dry friction. Also it is possible and natural to decompose trusses, that is, structures with pin joints. Those situations cannot be handled with models obtained from three-dimensional elasticity by asymptotic analysis. See Saint Jean Paulin and Vanninathan [34] as an exemplary paper in this direction.

We introduce a mass matrix M_i, a stiffness matrix K_i, structural matrices R_i, S_i such that M_i, K_i, S_i are symmetric and positive definite (S_i is either positive definite or equal to zero). All matrices are sufficiently smooth with respect to x along arc i. With this notation one can define a kinetic and potential energy as well as a work functional as follows:

$$\mathcal{K}(t) := \frac{1}{2}\sum_{i=1}^{n_e} \int_0^{l_i} M_i(x)\dot r_i(x,t)\cdot \dot r_i(x,t)\,dx, \tag{1.1}$$

$$\mathcal{U}(t) := \frac{1}{2}\sum_{i=1}^{n_e}\int_0^{l_i}[K_i(x)(r_i'(x,t) + R_i(x)r_i(x,t)) \\ \cdot (r_i'(x,t) + R_i(x)r_i(x,t)) + S_i(x)r_i(x,t)\cdot r_i(x,t)]\,dx, \tag{1.2}$$

$$\mathcal{W}(t) := \sum_{i=1}^{n_e}\int_0^{l_i} F_i(x,t)r_i(x,t)\,dx + \sum_{v_J\in V} f_J(t)\lambda_J(t) \tag{1.3}$$

(dot and prime represent time and space derivatives, respectively). F_i and f_J are distributed body and nodal forces, respectively. The Lagrangian then is defined as

$$\mathcal{L}(r) := \int_0^T [\mathcal{K}(t) + \mathcal{W}(t) - \mathcal{U}(t)]\,dt. \tag{1.4}$$

Before we proceed to write down the equations of motion as in [14] we would like to indicate the range of applications of the formulation (1.1)–(1.3). For more details see [14].

Examples.

1. *Three-dimensional string networks.* $p_i = q_i = 3$, $C_{iJ} = I_{3\times 3}$,

$$R_i = S_i = 0,\ M_i = \rho_i I,\ S_i = 0,$$
$$K_i = Eh_i\left[\left(1 - \frac{1}{s_i}\right)I_{3\times 3} + \frac{1}{s_i}e_ie_i^T\right],$$

where e_i is the unit vector along the straight arc i, $s_i > 1$ is a stretching factor, E is Young's modulus, h_i is the thickness, and ρ_i is the density. The matrix K_i may also depend on x. This is the case when the original nonlinear problem is linearized around an equilibrium corresponding to a preloading of the network.

2. *Two-dimensional Timoshenko beam networks.* We have $e_i, e_i^\perp$ as above and n (orthogonal to the common $e_i - e_i^\perp$-plane). $M_i = \mathrm{diag}(\rho_i, \rho_i, \rho_i I_i)$,

$$K_i = \mathrm{diag}(Eh_i, Gh_i, EI_i),\ R_{i23} = 1,\ R_{ij} = 0 \text{ else},\ S_i = 0.$$

In this case the potential energy (1.2) reduces to

$$\mathcal{U}(t) = \frac{1}{2}\sum_{i=1}^{n_e}[Eh_i {u'}_i^2 + Gh_i({w'}_i + \psi_i)^2 + EI_i{\psi'}_i^2]\,dx,$$

where we have identified $r_i = (u_i, w_i, \psi_i)^T$ such that u_i, w_i denote longitudinal and vertical displacement, while ψ_i represents the shear deformation. $\mathcal{K}(t)$ and $\mathcal{W}(t)$ are obvious. In addition to the physical quantities introduced above, EI_i denotes the flexural rigidity, Eh_i the longitudinal stiffness, G_i the shear modulus, and I_i the moment of the cross section. The setup is seen to reproduce exactly the Timoshenko beam theory. For rigid joints $C_{iJ} = (e_i, e_i^\perp, n)$, for pin joints

$$C_{iJ} = (e_i, e_i^\perp)^T(e_i, e_i^\perp, n) \in \mathbb{R}^{2\times 3}.$$

3. *Three-dimensional Timoshenko beam networks.* This is analogous to Example 2; see [14].
4. *Networks of strings and beams.* See [14].

5. *Two-dimensional precurved Timoshenko beam (Bresse beam) networks.* This case is as in (1.2) but with

$$R_i = \begin{pmatrix} 0 & -\kappa_i & 0 \\ \kappa_i & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},$$

where $\kappa_i(x)$ is the curvature of arc i in the $e_i - e_i^\perp$ plane.

Once the formulation (1.1)–(1.4) is seen to be sufficiently general, we proceed to write down the equations of motion, which are derived from (1.1)–(1.4) by standard variational principles.

$$\text{(1.5)} \qquad M_i\ddot{r}_i = [K_i(r_i' + R_i r_i)]' - R_i^T K_i(r_i' + R_i r_i) - S_i r_i + F_i, \qquad (x,t) \in (0,l_i) \times (0,T),$$

$$\text{(1.6)} \qquad r_i(v_D) = 0,\ \epsilon_{iD} \neq 0,\ v_D \in V_D,\ i \in \mathcal{E}_D,\ t \in (0,T),$$

$$\text{(1.7)} \qquad C_{iJ} r_i(v_J) = C_{kJ} r_k(v_J) =: \lambda_J,\ \forall\, i,k \in \mathcal{E}_J,\ v_J \in V_M,\ t \in (0,T),$$

$$\text{(1.8)} \qquad \sum_{i\in\mathcal{E}_J} \epsilon_{iJ}(C_{iJ}^+)^T [K_i(r_i' + R_i r_i)](v_J) = f_J - M_J\ddot{\lambda}_J, \qquad v_J \in V_M \cup V_N \cup V_C,\ t \in (0,T),$$

$$\text{(1.9)} \qquad \epsilon_{iJ}\Pi_{iJ}[K_i(r_i' + R_i r_i)](v_J) = 0,\ v_J \in V_M \cup V_N \cup V_C,\ t \in (0,T),$$

$$\text{(1.10)} \qquad r_i(\cdot,0) = r_{i0},\ \dot{r}_i(\cdot,0) = r_{i1}, \quad x \in (0,l_i).$$

Here we have denoted by V_M, V_S the sets of multiple and simple nodes such that $V = V_M \dot\cup V_S$ and $V_S = V_N \dot\cup V_D \dot\cup V_C$ (disjoint union) separates into free Neumann nodes, fully clamped (Dirichlet) nodes, and controlled Neumann nodes. One can of course also consider partially clamped nodes. The control forces f_J are set to zero for $v_J \in V_N \cup V_M$. The matrix M_J represents mass and rotatory inertia of a rigid body at node v_J. In this paper, however, we will consider $M_J = 0$ and return to the more general case in some remarks below.

It should be noted that one might include nonlinear coupling dynamics in (1.8) and (1.9) such as dry friction, plasticity, obstacles, Winkler supports, etc..

The standing assumption throughout the paper is that

$$\text{(A)} \begin{cases} & \text{either} \\ \text{(i)} & \text{every } S_i \text{ is positive definite} \\ & \text{or} \\ \text{(ii)} & V_D \neq \emptyset \text{ and for each node there is a path to a node} \\ & \text{in } V_D \text{ along which all joints are rigid.} \end{cases}$$

The standard case is (Aii), which simply states that, in engineering terms, the structure is not a mechanism; in other words no part can perform rigid rotations, or the structure is statically determined. Its mathematical significance is revealed after defining appropriate spaces, as in [14]:

$$\text{(1.11)} \qquad \mathbf{H} = \prod_{i=1}^{n_e} L^2(0,l_i,\mathbb{R}^{p_i}),$$

$$\text{(1.12)} \qquad \mathbf{V} = \left\{ r \in \prod_{i=1}^{n_e} H^1(0,l_i,\mathbb{R}^{p_i}) | r_i \text{ satisfies (1.6), (1.7)} \right\}.$$

We define the inner products

$$(r,\hat r)_{\mathbf{H}} = \frac{1}{2}\sum_{i=1}^{n_e}\int_0^{l_i} M_i r_i \hat r_i \, dx, \tag{1.13}$$

$$(r,\hat r)_{\mathbf{V}} = \frac{1}{2}\sum_{i=1}^{n_e}\int_0^{l_i} [K_i(r_i' + R_i r_i)(\hat r_i' + R_i \hat r_i) + (S_i + I) r_i \hat r_i] \, dx. \tag{1.14}$$

By Lemma 3.1, Chapter IV of [14], we find that under assumption (A) the norm $\|r\|_{\mathbf{V}} = (r,r)_{\mathbf{V}}^{1/2}$, even without the term I, is equivalent to

$$\|r\| := \left(\sum_{i=1}^{n_e}\int_0^{l_i} [|r_i|^2 + |r_i'|^2] \, dx\right)^{1/2}. \tag{1.15}$$

We also cite Theorem 3.1, Chapter IV from [14]. We denote by $\mathbf{V}^*$ the dual of $\mathbf{V}$.

THEOREM 1.1. *Let* $(r_0, r_1) \in \mathbf{V}\times\mathbf{H}$*,* $f \in L^2(0,T,\mathbf{H})$*,* $M_J = 0\,\forall j$*,* $f_J \in L^2(0,T,\mathbb{R}^{q_J})$*,* $J : v_J \in V_C$*. Let* (A) *be satisfied. Then there exists a unique mild solution* r *of* (1.5)–(1.10) *with regularity* $r \in C(0,T,\mathbf{V}) \cap C^1(0,T,\mathbf{H}) \cap C^2(0,T,\mathbf{V}^*)$*.*

2. Control problems.

In [14] and other related work we asked the question as to whether networks described by (1.5)–(1.10) are controllable. In particular, given initial data $(r_0, r_1) \in \mathbf{V}\times\mathbf{H}$, $F \equiv 0$, is it possible to find controls $f_J \in L^2(0,T,\mathbb{R}^{q_J})$, $v_J \in V_C$, such that after time $T > 0$, reasonably large, the final values $r(\cdot,T)$, $\dot r(\cdot,T)$ are either exactly zero (or prescribed) or in an ϵ-ball around zero?

The first case is known as exact controllability (reachability), while the second case for $\epsilon > 0$ (arbitrary) is referred to as approximate controllability.

We have been able to give affirmative answers only for tree networks with $|V_D| = 1$ and $V_S = V_C$. See Theorem 5.1, Chapter IV of [14], and [36]. For string trees with $|V_D| = 1$, $V_S \supset V_C$ (and possibly joint masses) we have some positive results [18], but the picture is more complicated. If, however, the network contains circuits and if it is completely homogenous, then even approximate controllability fails to hold. On the other side positive results can be expected to hold whenever the individual optical lengths of the strings are rationally independent, even though this has not yet been rigorously investigated. All relevant results can be translated as well into the context of stabilizability.

Given this complex picture, one is tempted to rather look into *optimal* control problems related to such networks. Indeed, in many circumstances one does not want to exactly (approximately) control the entire structure. Rather, one wants to protect some substructure against vibrational energy. This can be achieved by minimizing the energy in that subregion or, alternatively, by steering the flux of energy away from the subregion. In order not to overload the presentation we choose the first case, since the second case would necessitate a different analysis.

We will consider the cost functional

$$J(f) := \frac{1}{2}\sum_{J\in V_C}\int_0^T |f_J|^2 \, dt + \frac{k}{2}\left\{\|r(T) - z_0\|_{\mathbf{V}}^2 + \|\dot r(T) - z_1\|_{\mathbf{H}}^2\right\}. \tag{2.1}$$

One might also take the $\mathbf{V}\times\mathbf{H}$ topology induced on the subgraph associated with a subset $\mathcal{C}$ of the edge set. If k grows large in (2.1), one comes close to controllability problems described above.

There are other reasonable choices for the cost function. In particular, one can work in the shifted energy norm and consider

$$J(f) := \frac{1}{2}\sum_{J\in V_C}\int_0^T |f_J|^2\,dt + \frac{k}{2}\left\{\|r(T)-z_0\|_{\mathbf{H}}^2 + \|\dot r(T)-z_1\|_{\mathbf{V}^*}^2\right\}. \tag{2.2}$$

One can also include $\int_0^T \|r\|_{\mathbf{H}}^2\,dt$, $\int_0^T \|r\|_{\mathbf{V}}^2\,dt$, or the total energy integrated over the time interval $[0,T]$, and, if the network is known to be exactly controllable, a general $L-Q-R$ infinite horizon problem would be of interest. The problem we discuss here is the following:

$$\min_{(f_J)_{J\in V_C}} J((f_J)) \text{ subject to (1.5)–(1.10)}, \tag{2.3}$$

with $J(\cdot)$ given by (2.1).

3. Global optimality conditions. It is a matter of applying standard arguments to derive the optimality system corresponding to the optimization problem (2.3). We introduce an adjoint state p localized to arc i as p_i, which then satisfies the adjoint "backward running" system

$$M_i\ddot p_i = [K_i(p_i'+R_ip_i)]' - R_i^TK_i(p_i'+R_ip_i) - S_ip_i, \qquad (x,t)\in(0,\ell_i)\times(0,T), \tag{3.1}$$

$$p_i(v_D)=0,\ \epsilon_{iD}\neq 0,\ v_D\in V_D,\ i\in\mathcal{E}_D, t\in(0,T), \tag{3.2}$$

$$C_{iJ}p_i(v_J) = C_{kJ}p_k(v_J) =: \Theta_J\ \forall\, i,k\in\mathcal{E}_J,\ v_J\in V_M, t\in(0,T), \tag{3.3}$$

$$\sum_{i\in\mathcal{E}_J}\epsilon_{iJ}(C_{iJ}^+)^T[K_i(p_i'+R_ip_i)](v_J)=0,\ v_J\in V_M\cup V_N\cup V_C,\ t\in(0,T), \tag{3.4}$$

$$\epsilon_{iJ}\Pi_{iJ}[K_i(p_i'+R_ip_i)](v_J)=0,\ v_J\in V_M\cup V_N\cup V_C,\ t\in(0,T), \tag{3.5}$$

$$p(T)=k(\dot r(T)-z_1),\ \dot p(T) = -kA(r(T)-z_0), \tag{3.6}$$

$$f_J = -p_{i_J},\ J: v_J\in V_C,\ i\in\mathcal{E}_J\ t\in(0,T). \tag{3.7}$$

The operator A in (3.6) is the Riesz isomorphism between $\mathbf{V}$ and $\mathbf{V}^*$. It is obvious from (3.6) that, because of the regularity of $r(T)\in\mathbf{V}$ and $\dot r(T)\in\mathbf{H}$, the final values of p have the regularity $p(T)\in\mathbf{H}$ and $\dot p(T)\in\mathbf{V}^*$. However, by the regularity result of Theorem 1.1, we know that the operator taking L^2-in-time Neumann data at external (simple) nodes into the final values of the solution is bounded with respect to the total energy. The adjoint of that operator selects the Dirichlet traces of the solutions of the homogenous problem at those external nodes. Hence, by transposition, the traces of p at nodes in V_C are L^2-in-time. The solution p satisfies (3.1)–(3.7) in the sense of transposition, as in Nicaise [29, Theorem 5.1]. It is also possible to derive the required regularity of traces directly using "direct" inequalities for the nonharmonic Fourier series associated with networks. We have the following theorem.

THEOREM 3.1. *Let the assumptions of Theorem* 1.1 *be satisfied. Then the optimality system* (1.1)–(1.10), (3.1)–(3.7) *has a unique solution* r, p *such that*

$$r\in C(0,T,\mathbf{V})\cap C^1(0,T,\mathbf{H})\cap C^2(0,T,\mathbf{V}^*),$$
$$p\in C(0,T,\mathbf{H})\cap C^1(0,T,\mathbf{V}^*)\cap C^2(0,T,D(A)^*).$$

In addition, we have $p_{iJ}(v_J,\dot{)}\in L^2(0,T)$ *for* $v_J\in V_C$, $i\in\mathcal{E}_J$.

Remark 3.1. If we consider data (r_0, r_1), $(z_0, z_1) \in \mathbf{D}(\mathbf{A}) \times \mathbf{V}$, then the adjoint state satisfies $p \in C(0, T, \mathbf{V}) \cap C^1(0, T, \mathbf{H})$, too.

This is seen as follows: we start with final data $(p_T, \dot{p}_T) \in \mathbf{V} \times \mathbf{H}$. We solve the adjoint system backward in time and take Dirichlet traces at controlled nodes which have $H^1(0, T; \mathbf{R}^{q_J})$ regularity. These traces are used as Neumann inputs to the forward running system, according to the optimality conditions. These inputs, together with the assumed regularity of the initial data, produce a solution r that has $C(0, T; \mathbf{D}(\mathbf{A})) \cap C^1(0, T, \mathbf{V})$ regularity. That regularity is shared by the final data $(r(T), \dot{r}(T))$. Since we assume the same regularity for the target, we have $(-\dot{r}(T),\ Ar(T)) := \Lambda(p_T, \dot{p}_T) \in \mathbf{V} \times \mathbf{H}$. Indeed, this process is precisely the one described in [12], [13]. It amounts to solving $\Lambda(p_T, \dot{p}_T) + \frac{1}{k} M(p_T, \dot{p}_T) = (-z_1, Az_0)$. It is plain that the parameter k serves as a Tychonov regularization of the Hilbert uniqueness method (HUM) operator Λ. The last equation can be uniquely solved in the finite energy space.

The main point of the paper is to provide an iterative procedure in order to solve (1.5)–(1.10), (3.1)–(3.7) based on decoupled subproblems. The reason for this is that the system (1.5)–(1.10), (3.1)–(3.7) is extremely large if the network is complex and the states are discretized by finite difference (FD), finite element (FEM) methods or wavelet approximations. Another important reason is that one arc might represent a string while another one represents a beam, so that different numerical approximations apply. It would, therefore, be extremely useful to be able to restrict calculations to a single element (arc), perform those *in parallel*, and communicate mutual results after the iteration step is completed.

For homogenous planar two-dimensional problems such a procedure has been derived by Benamou [1], [2], [3], and [4]. The procedure was inspired by Lions [25], who discussed an iterative nonoverlapping Schwarz decomposition for a static (elliptic) problem. It is straightforward to apply that technique to dynamic, parabolic, or hyperbolic problems and it was then realized by Benamou that the resulting dynamic domain decomposition (d^3m) method could be applied to optimality systems. However, it was by no means clear how to obtain such decoupling in the case of networks as discussed above.

In this respect, an observation made by Glowinski and Le Tallec [11] turned out to be very useful. Namely, they showed that Lions's method was equivalent to an augmented Lagrangian approach combined with a standard saddle-point iteration. We used the latter approach to derive dynamic domain decomposition methods for both simulation purposes and, more important for this paper, for optimal control problems for planar networks of strings and Euler–Bernoulli beams (with constant coefficients) [19], [23]. While in the case of string networks we worked with the cost analogous to (2.1), and we used the entire state throughout $(0, T)$ for Euler–Bernoulli beams. Note that the full state cost is easier to handle. Full state cost is the one used throughout in [1], [2], [3]. We have also looked into another possibility based on so-called Dirichlet–Neumann iterations as in [8] and [35]; see [18].

4. Dynamic domain decomposition.

Instead of (1.7), (1.8) and (3.3), (3.4) we consider, at multiple nodes v_J, the following mixed Robin-type boundary conditions for r_i^{n+1}, p_i^{n+1}, where n is the iteration index:

$$\begin{aligned}
&\epsilon_{iJ}(C_{iJ}^{+})^T[K_i((r_i^{n+1})' + R_i r_i^{n+1})](v_J) + \beta C_{iJ} p_i^{n+1}(v_J) \\
&= \beta\left(\frac{2}{d_J}\sum_{j\in\mathcal{E}_J} C_{jJ} p_j^n(v_J) - C_{iJ} p_i^n(v_J)\right)
\end{aligned} \tag{4.1}$$

$$-\left(\frac{2}{d_J}\sum_{j\in\mathcal{E}_J}\epsilon_{jJ}(C_{jJ}^{+})^T[K_j({r_j^n}'+R_jr_j^n)](v_J)\right.$$
$$\left.-\epsilon_{iJ}(C_{iJ}^{+})^T[K_i({r_i^n}'+R_ir_i^n)](v_J)\right)=:\lambda_{iJ}^n,$$

$$\begin{aligned}&\epsilon_{iJ}(C_{iJ}^{+})^T[K_i((p_i^{n+1})'+R_ip_i^{n+1})](v_J)-\beta C_{iJ}r_i^{n+1}(v_J)\\ &=-\beta\left(\frac{2}{d_J}\sum_{j\in\mathcal{E}_J}C_{jJ}r_j^n(v_J)-C_{iJ}r_i^n(v_J)\right)\\ &\quad-\left(\frac{2}{d_J}\sum_{j\in\mathcal{E}_J}\epsilon_{jJ}(C_{jJ}^{+})^T[K_j({p_j^n}'+R_jp_j^n)](v_J)\right.\\ &\quad\left.-\epsilon_{iJ}(C_{iJ}^{+})^T[K_i({p_i^n}'+R_ip_i^n)](v_J)\right)=:\mu_{iJ}^n\end{aligned}\tag{4.2}$$

for all $v_J\in V_M,\ t\in(0,T)$. The parameter $\beta>0$, which can be interpreted as an additional stiffness at the vertex v_J, turns out to be useful in the numerical simulations. Tuning β allows us to vary essentially between Dirichlet conditions ($\beta\to\infty$) and Neumann conditions ($\beta\to 0$).

The other conditions (1.5), (1.6), (1.8) at nodes in $V_N\cup V_C$, (1.9), (1.10) and (3.1), (3.2), (3.4) at nodes in $V_N\cup V_C$ (3.5)–(3.7) are kept unchanged.

Let us assume for a moment that the sequence (r_i^n,p_i^n) starting at some given (r_i^0,p_i^0) converges. We would like to check whether (4.1), (4.2) lead to the transmission conditions (1.6), (1.7), (1.8) and (3.2), (3.3), (3.4). Indeed, (4.1), (4.2) with the index n dropped amount to

$$\begin{aligned}0&=2\beta\left(\frac{1}{d_J}\sum_{j\in\mathcal{E}_J}C_{jJ}p_j(v_J)-C_{iJ}p_i(v_J)\right)\\ &\quad-\frac{2}{d_J}\sum_{j\in\mathcal{E}_J}\epsilon_{jJ}(C_{jJ}^{+})^T[K_j({r_j}'+R_jr_j)](v_J),\end{aligned}\tag{4.3}$$

$$\begin{aligned}0&=-2\beta\left(\frac{1}{d_J}\sum_{j\in\mathcal{E}_J}C_{jJ}r_j(v_J)-C_{iJ}r_i(v_J)\right)\\ &\quad-\frac{2}{d_J}\sum_{j\in\mathcal{E}_J}\epsilon_{jJ}(C_{jJ}^{+})^T[K_j({p_j}'+R_jp_j)](v_J).\end{aligned}\tag{4.4}$$

Upon summing (4.3), (4.4) over all arcs incident at v_J, we obtain

$$\begin{aligned}\sum_{j\in\mathcal{E}_J}\epsilon_{iJ}(C_{jJ}^{+})^T[K_i({r_i}'+R_ir_i)](v_J)&=0,\\ \sum_{j\in\mathcal{E}_J}\epsilon_{iJ}(C_{iJ}^{+})^T[K_i({p_i}'+R_ip_i)](v_J)&=0,\end{aligned}\tag{4.5}$$

which is (1.8), (3.4) at $v_J \in V_M$. If we now use (4.5) in (4.3), (4.4) we arrive at

$$\begin{aligned} C_{iJ}p_i(v_J) &= \frac{1}{d_J}\sum_{j\in\mathcal{E}_J} C_{jJ}p_j(v_J), \\ C_{iJ}r_i(v_J) &= \frac{1}{d_J}\sum_{j\in\mathcal{E}_J} C_{jJ}r_j(v_J) \end{aligned} \tag{4.6}$$

at v_J and for all $i \in \mathcal{E}_J$. But that implies

$$C_{iJ}p_i(v_J),\ C_{iJ}r_i(v_J) \text{ are independent of } i \in \mathcal{E}_J. \tag{4.7}$$

We thus see that, once the sequence (r_i^n, p_i^n) converges, the limiting (r, p) satisfies all boundary and transmission conditions in addition to the local balance equations (1.5) and (3.1). That is, in the limit, the global optimality conditions are satisfied. Note that this argument is independent of the choice of β.

We also consider some variants of the domain decomposition below:

- Instead of (4.1), (4.2) we can take into account the knowledge of those new iterates that have already been computed. While (4.1), (4.2) can be regarded as a Jacobi-type iteration, the latter strategy would be more in the spirit of a Gauss–Seidel-type iteration. One expects, and actually numerically confirms, faster convergence. However, with the latter strategy one sacrifices the inherent parallelism to some extent.
- A second extension utilizes another standard technique, namely, relaxation. This means that we reuse the old information on the actual edge with an underrelaxation parameter $\lambda \in [0, 1)$ as follows:

$$\begin{aligned} &\epsilon_{iJ}(C_{iJ}^{+})^T[K_i((r_i^{n+1})' + R_i r_i^{n+1})](v_J) + \beta C_{iJ}p_i^{n+1}(v_J) \\ &= (1-\lambda)\Bigg(\beta\Bigg(\frac{2}{d_J}\sum_{j\in\mathcal{E}_J} C_{jJ}p_j^n(v_J) - C_{iJ}p_i^n(v_J)\Bigg) \\ &\quad -\Bigg(\frac{2}{d_J}\sum_{j\in\mathcal{E}_J}\epsilon_{jJ}(C_{jJ}^{+})^T[K_j(r_j^{n\prime} + R_j r_j^n)](v_J) \\ &\qquad - \epsilon_{iJ}(C_{iJ}^{+})^T[K_i(r_i^{n\prime} + R_i r_i^n)](v_J)\Bigg)\Bigg) \\ &\quad + \lambda\Bigg(\epsilon_{iJ}(C_{iJ}^{+})^T[K_i((r_i^n)' + R_i r_i^n)](v_J) + C_{iJ}p_i^n(v_J)\Bigg), \end{aligned} \tag{4.8}$$

$$\begin{aligned} &\epsilon_{iJ}(C_{iJ}^{+})^T[K_i((p_i^{n+1})' + R_i p_i^{n+1})](v_J) - \beta C_{iJ}r_i^{n+1}(v_J) \\ &= (1-\lambda)\Bigg(-\beta\Bigg(\frac{2}{d_J}\sum_{j\in\mathcal{E}_J} C_{jJ}r_j^n(v_J) - C_{iJ}r_i^n(v_J)\Bigg) \\ &\quad -\Bigg(\frac{2}{d_J}\sum_{j\in\mathcal{E}_J}\epsilon_{jJ}(C_{jJ}^{+})^T[K_j(p_j^{n\prime} + R_j p_j^n)](v_J) \\ &\qquad - \epsilon_{iJ}(C_{iJ}^{+})^T[K_i(p_i^{n\prime} + R_i p_i^n)](v_J)\Bigg)\Bigg) \\ &\quad + \lambda\Bigg(\epsilon_{iJ}(C_{iJ}^{+})^T[K_i((p_i^n)' + R_i p_i^n)](v_J) - C_{iJ}r_i^n(v_J)\Bigg) \end{aligned} \tag{4.9}$$

for all $v_J \in V_M,\ t \in (0,T)$.

- A third variant, which is the extension to networks of an algorithm described by Deng [6] for planar elliptic problems, can be written as follows:

$$\begin{aligned} &\epsilon_{iJ}(C_{iJ}^{+})^T[K_i((r_i^{n+1})' + R_i r_i^{n+1})](v_J) + \beta C_{iJ} p_i^{n+1}(v_J) =: g_{iJ}^{n+1} \\ &= 2\beta\left(\frac{2}{d_J}\sum_{j\in\mathcal{E}_J} C_{jJ} p_j^n(v_J) - C_{iJ} p_i^n(v_J)\right) \\ &\qquad -\left(\frac{2}{d_J}\sum_{j\in\mathcal{E}_J} g_{jJ}^n - g_{iJ}^n\right), \end{aligned} \tag{4.10}$$

$$\begin{aligned} &\epsilon_{iJ}(C_{iJ}^{+})^T[K_i((p_i^{n+1})' + R_i p_i^{n+1})](v_J) - \beta C_{iJ} r_i^{n+1}(v_J) =: h_{iJ}^{n+1} \\ &= -2\beta\left(\frac{2}{d_J}\sum_{j\in\mathcal{E}_J} C_{jJ} r_j^n(v_J) - C_{iJ} r_i^n(v_J)\right) \\ &\qquad -\left(\frac{2}{d_J}\sum_{j\in\mathcal{E}_J} h_{jJ}^n - h_{iJ}^n\right) \end{aligned} \tag{4.11}$$

for all $v_J \in V_M,\ t \in (0,T)$. Reducing the iteration index in the defining equations (4.10), (4.11) to n and summing over all incident edges, it is easy to see that this variant is equivalent to the original conditions (4.1), (4.2). One can then also introduce a relaxation parameter into this variant. We conclude this discussion of alternative constructions with the remark that the last approach, even though equivalent to the first, is easier to implement, as no derivatives have to be computed for the update.

Let us, for the sake of easier reference, collect the conditions for the decoupled system (r_i^{n+1}, p_i^{n+1})

$$\begin{cases} M_i \ddot{r}_i^{n+1} = [K_i((r_i^{n+1})' + R_i r_i^{n+1})]' & - R_i^T K_i(r_i^{n+1\prime} + R_i r_i^{n+1}) - S_i r_i^{n+1}, \\ M_i \ddot{p}_i^{n+1} = [K_i((p_i^{n+1})' + R_i p_i^{n+1})]' & - R_i^T K_i(p_i^{n+1\prime} + R_i p_i^{n+1}) - S_i p_i^{n+1}, \\ & \qquad (x,t) \in (0,\ell_i) \times (0,T), \end{cases} \tag{4.12}$$

$$r_i(v_D) = 0,\ p_i(v_D) = 0,\ \epsilon_{iD} \neq 0,\ v_D \in V_D,\ t \in (0,T), \tag{4.13}$$

$$\begin{cases} \epsilon_{iC}[K_i((r_i^{n+1})' + R_i r_i^{n+1})](v_C) &= -p_i^{n+1}(v_C), \\ \epsilon_{iC}[K_i((p_i^{n+1})' + R_i p_i^{n+1})](v_C) &= 0,\ v_C \in V_C,\ t \in (0,T), \end{cases} \tag{4.14}$$

$$\begin{cases} \epsilon_{iN}[K_i((r_i^{n+1})' + R_i r_i^{n+1})](v_N) &= 0, \\ \epsilon_{iN}[K_i((p_i^{n+1})' + R_i p_i^{n+1})](v_N) &= 0,\ v_N \in V_N,\ t \in (0,T), \end{cases} \tag{4.15}$$

$$\begin{cases} \epsilon_{iJ}\Pi_{iJ}[K_i((r_i^{n+1})' + R_i r_i^{n+1})](v_J) &= 0, \\ \epsilon_{iJ}\Pi_{iJ}[K_i((p_i^{n+1})' + R_i p_i^{n+1})](v_J) &= 0,\ v_J \in V_M,\ t \in (0,T), \end{cases} \tag{4.16}$$

$$r_i^{n+1}, p_i^{n+1} \text{ satisfies (4.1), (4.2) (or (4.8), (4.9)) at } v_J \in V_M,\ t \in (0,T), \tag{4.17}$$

$$\begin{cases} r^{n+1}(\cdot,0) &= r_0,\ \dot{r}^{n+1}(\cdot,0) = r_1, \\ p^{n+1}(\cdot,T) &= k(\dot{r}^{n+1}(\cdot,T) - z_1), \\ \dot{p}^{n+1}(\cdot,T) &= -kA(r^{n+1}(\cdot,T) - z_0),\ x \in (0,l_i). \end{cases} \tag{4.18}$$

In order to discuss problems (4.12)–(4.18) for each individual $i = 1 : n_e$, we have to distinguish four cases:

(i) $\epsilon_{iC}, \epsilon_{iJ} \neq 0$ for $v_C \in V_C$, $v_J \in V_M$,

(ii) $\epsilon_{iJ}, \epsilon_{iM} \neq 0$ for $v_J, v_M \in V_M$,

(iii) $\epsilon_{iD}, \epsilon_{iD} \neq 0$ for $v_D \in V_D$, $v_J \in V_M$,

(iv) $\epsilon_{iN}, \epsilon_{iJ} \neq 0$ for $v_N \in V_N$, $v_J \in V_M$.

The single edge network $\epsilon_{iD}, \epsilon_{iB} \neq 0$ $v_D, v_B \in V_S$ is trivial. The most complicated case is (i), which we will consider below. The other cases then follow similarly. Now, as (4.1), (4.2) are known boundary inputs λ_{iJ}^n and μ_{iJ}^n, respectively, one can look at the system (4.12)–(4.18) as an optimality system corresponding to an optimal control problem on the individual arc i as follows:

$$
\begin{aligned}
&\min_{f_J, f_C} \frac{k}{2} \left\{ \|r_i^{n+1}(T) - z_i^0\|_{H^1(0,\ell_i)^{p_i}}^2 + \|\dot r_i^{n+1}(T) - z_i^1\|_{L^2(0,\ell_i)^{p_i}}^2 \right\} \\
&\qquad + \frac{1}{2} \int_0^T \left\{ \frac{1}{\beta} |f_J|^2 + |f_C|^2 \right\} dt \\
&\qquad + \frac{\beta}{2} \int_0^T \left| C_{iJ} r_i^{n+1}(v_J) + \frac{1}{\beta} \mu_{iJ}^n \right|^2 dt,
\end{aligned} \tag{4.19}
$$

$$
\begin{aligned}
&M_i \ddot r_i^{n+1} = [K_i((r_i^{n+1})' + R_i r_i^{n+1})]' - R_i^T K_i((r_i^{n+1})' + R_i r_i^{n+1}) - S_i r_i^{n+1}, \\
&\qquad\qquad\qquad\qquad (x,t) \in (0,l_i) \times (0,T), \\
&\epsilon_{iC}[K_i((r_i^{n+1})' + R_i r_i^{n+1})](v_C) = f_C,\ t \in (0,T), \\
&\epsilon_{iJ}(C_{iJ}^+)^T [K_i((r_i^{n+1})' + R_i r_i^{n+1})](v_J) = f_J + \lambda_{iJ}^n,\ t \in (0,T), \\
&\epsilon_{iJ} \Pi_{iJ} [K_i((r_i^{n+1})' + R_i r_i^{n+1})](v_J) = 0,\ t \in (0,T), \\
&r_i^{n+1}(0) = r_{i0},\ \dot r_i^{n+1}(0) = r_{i1},\ x \in (0,\ell_i).
\end{aligned} \tag{4.20}
$$

The proof that (4.19), (4.20) is actually the corresponding optimal control problem is done by standard variational arguments. Now, regularity theory in [14, Chapter IV] yields Theorem 4.1.

THEOREM 4.1. *Let* $(r_{i0}, r_{i1}) \in H^1(0,l_i) \times L^2(0,l_i)$, $\lambda_{iJ}^n, \mu_{iJ}^n \in L^2(0,T,\mathbb{R}^{q_J})$. *Then problem* (4.19), (4.20) *has unique minimizers* $f_C \in L^2(0,T,\mathbb{R}^{p_i})$, $f_J \in L^2(0,T,\mathbb{R}^{q_J})$. *The necessary optimality conditions are given by* (4.12)–(4.18).

Note that a similar result holds for the relaxed algorithm, where the λ_{iJ}^n, μ_{iJ}^n now depend on the relaxation parameter λ as in (4.8), (4.9).

Remark 4.1. The same arguments as in Remark 3.1 apply to the local optimality system. Hence, working with the local regularity $H^2(0,\ell_i) \times H^1(0,\ell_i)$ for initial and target data, and assuming in addition $\lambda_{iJ}^n, \in H^1(0,T,\mathbb{R}^{q_J})$, then the adjoint system has local finite-energy regularity.

5. Convergence of the algorithm.

Let us denote the unique solution of (4.12)–(4.18) by $(\hat r_i, \hat p_i)$ and by (r_i, p_i) the local restrictions of the solution (r,p) of the global optimality system (1.5)–(1.10) (with $F_i, M_J = 0$), (3.1)–(3.7). We then introduce the errors $\tilde r_i^n := r_i - \hat r_i^n$, $\tilde p_i^n := p_i - \hat p_i^n$. Now, observe that p_i, r_i also satisfy (4.1), (4.2), and, of course also the other conditions (4.12)–(4.16), (4.18) for the decoupled system.

This shows that $\tilde r_i^{n+1}, \tilde p_i^{n+1}$ satisfy (4.12)–(4.18), where (4.18) is changed to

$$\begin{cases} \tilde r_i^{n+1}(0) &= 0,\ \dot{\tilde r}_i^{n+1}(0) = 0, \\ \tilde p_i^{n+1}(T) &= k\dot{\tilde r}_i^{n+1}(T), \\ \dot{\tilde p}_i^{n+1}(T) &= -kA_i \tilde r_i^{n+1}(T), \end{cases} \tag{5.1}$$

and A_i, defined by $(Ar)_i =: A_i r_i$, is the restriction of A to the edge with index i. The same is true for all other cases (ii)–(iv). We proceed to show that the errors $\tilde r_i^n$, $\tilde p_i^n$ converge to zero in a natural energy sense.

The main work involves the condition (4.1), (4.2). For the sake of simplicity, we take $\beta = 1$, the arguments being valid for general $\beta > 0$. We multiply (4.1), (4.2) by $C_{iJ}\tilde p_i^{n+1}(v_J)$, $C_{iJ}\tilde r_i^{n+1}(v_J)$, respectively, and then sum over all incident arcs.

$$\begin{aligned} &\sum_{i\in\mathcal{E}_J} \epsilon_{iJ} (C_{iJ}^+)^T [K_i((\tilde r_i^{n+1})' + R_i \tilde r_i^{n+1})](v_J) \cdot C_{iJ}\tilde p_i^{n+1}(v_J) + \sum_{i\in\mathcal{E}_J} |C_{iJ}\tilde p^{n+1}(v_J)|^2 \\ &= \frac{2}{d_J} \sum_{j\in\mathcal{E}_J} C_{jJ}\tilde p_j^n(v_J) \sum_{i\in\mathcal{E}_J} C_{iJ}\tilde p_i^{n+1}(v_J) - \sum_{i\in\mathcal{E}_J} C_{iJ}\tilde p_i^n(v_J) C_{iJ}\tilde p_i^{n+1}(v_J) \\ &\quad -\Bigg(\frac{2}{d_J} \sum_{j\in\mathcal{E}_J} \epsilon_{jJ} (C_{jJ}^+)^T [K_j(\tilde r_j^{n\prime} + R_j \tilde r_j^n)](v_J) \sum_{i\in\mathcal{E}_J} C_{iJ}\tilde p_i^{n+1}(v_J) \\ &\qquad - \sum_{i\in\mathcal{E}_J} \epsilon_{iJ} (C_{iJ}^+)^T [K_i(\tilde r_i^{n\prime} + R_i \tilde r_i^n)](v_J) \cdot C_{iJ}\tilde p_i^{n+1}(v_J) \Bigg), \end{aligned} \tag{5.2}$$

$$\begin{aligned} &\sum_{i\in\mathcal{E}_J} \epsilon_{iJ} (C_{iJ}^+)^T [K_i((\tilde p_i^{n+1})' + R_i \tilde p_i^{n+1})](v_J) C_{iJ}\tilde r_i^{n+1}(v_J) - \sum_{i\in\mathcal{E}_J} |C_{iJ}\tilde r_i^{n+1}(v_J)|^2 \\ &= -\Bigg(\frac{2}{d_J} \sum_{j\in\mathcal{E}_J} C_{jJ}\tilde r_j^n(v_J) \sum_{i\in\mathcal{E}_J} C_{iJ}\tilde r_i^{n+1}(v_J) - \sum_{i\in\mathcal{E}_J} C_{iJ}\tilde r_i^n(v_J) C_{iJ}\tilde r_i^{n+1}(v_J) \Bigg) \\ &\quad -\Bigg(\frac{2}{d_J} \sum_{j\in\mathcal{E}_J} \epsilon_{jJ} (C_{jJ}^+)^T [K_j(\tilde p_j^{n\prime} + R_j \tilde p_j^n)](v_J) \sum_{i\in\mathcal{E}_J} C_{iJ}\tilde r_i^{n+1}(v_J) \\ &\qquad - \sum_{i\in\mathcal{E}_J} \epsilon_{iJ} C_{iJ}^+ [K_i(\tilde p_i^{n\prime} + R_i \tilde p_i^n)](v_J) C_{iJ}\tilde r_i^{n+1}(v_J) \Bigg). \end{aligned} \tag{5.3}$$

On the other hand, taking squares in (4.1), (4.2) we obtain

$$\begin{aligned} &|(C_{iJ}^+)^T [K_i((\tilde r_i^{n+1})' + R_i \tilde r_i^{n+1})](v_J)|^2 = |C_{iJ}\tilde p_i^{n+1}(v_J)|^2 \\ &+ |(C_{iJ}^+)^T [K_i(\tilde r_i^{n\prime} + R_i \tilde r_i^n)](v_J)|^2 \\ &- 2C_{iJ}\tilde p_i^{n+1}(v_J) \Bigg(\frac{2}{d_J} \sum_{j\in\mathcal{E}_J} C_{jJ}\tilde p_j^n(v_J) - C_{iJ}\tilde p_i^n(v_J) \Bigg) \\ &-\Bigg(\frac{2}{d_J} \sum_{j\in\mathcal{E}_J} \epsilon_{jJ} (C_{jJ}^+)^T [K_j(\tilde r_j^{n\prime} + R_j \tilde r_j^n)](v_J) \\ &\qquad - \epsilon_{iJ} (C_{iJ}^+)^T [K_i(\tilde r_i^{n\prime} + R_i \tilde r_i^n)](v_J) \Bigg) \end{aligned} \tag{5.4}$$

$$
\begin{aligned}
&+\frac{4}{d_J^2}\left|\sum_{j\in\mathcal{E}_J} C_{jJ}\tilde{p}_j^n(v_J)\right|^2 + \left|C_{iJ}\tilde{p}_i^n(v_J)\right|^2\\
&-\frac{4}{d_J}\sum_{j\in\mathcal{E}_J} C_{jJ}\tilde{p}_j^n(v_J)C_{iJ}\tilde{p}_i^n(v_J)\\
&+\frac{4}{d_J^2}|\sum_{j\in\mathcal{E}_J}\epsilon_{jJ}(C_{jJ}^+)^T[K_j(\tilde{r}_j^{n\prime}+R_j\tilde{r}_j^n)](v_J)\big|^2\\
&-\frac{4}{d_J^2}\sum_{j\in\mathcal{E}_J}\epsilon_{jJ}(C_{jJ}^+)^T[K_j(\tilde{r}_j^{n\prime}+R_j\tilde{r}_j)](v_J)\\
&\qquad\qquad -\epsilon_{iJ}(C_{iJ}^+)^T[K_i(\tilde{r}_i^{n\prime}+R_i\tilde{r}_i^n)](v_J)\\
&-2\left(\frac{2}{d_J}\sum_{j\in\mathcal{E}_J} C_{jJ}\tilde{p}_j^n(v_J)-\tilde{p}_i^n(v_J)\right)\\
&\qquad\qquad -\left(\frac{2}{d_J}\sum_{j\in\mathcal{E}_J}\epsilon_{jJ}(C_{jJ}^+)^T[K_j(\tilde{r}_j^{n\prime}+R_j\tilde{r}_j^n)](v_J)\right.\\
&\qquad\qquad\qquad \left.-\epsilon_{iJ}(C_{iJ}^+)^T[K_i(\tilde{r}_i^{n\prime}+R_i\tilde{r}_i^n)](v_J)\right).
\end{aligned}
$$

A similar expression holds for the $\tilde{p}_i^n$-variable. We sum (5.4) over incident edges and obtain

$$
\begin{aligned}
&\sum_{i\in\mathcal{E}_J}|(C_{iJ}^+)^T[K_i((\tilde{r}_i^{n+1})'+R_i\tilde{r}_i^{n+1})](v_J)|^2\\
&\quad=\sum_{i\in\mathcal{E}_J}|C_{iJ}\tilde{p}_i^{n+1}(v_J)|^2+\sum_{i\in\mathcal{E}_J}|C_{iJ}\tilde{p}_i^n(v_J)|^2\\
&\qquad+\sum_{i\in\mathcal{E}_J}|(C_{iJ}^+)^T[K_i(\tilde{r}_i^{n\prime}+R_i\tilde{r}_i^n(v_J))]|^2\\
&\quad+2\Bigg[-\sum_{i\in\mathcal{E}_J}C_{iJ}\tilde{p}_i^{n+1}(v_J)\frac{2}{d_J}\sum_{j\in\mathcal{E}_J}C_{jJ}\tilde{p}_j^n(v_J)\\
&\qquad+\sum_{i\in\mathcal{E}_J}C_{iJ}\tilde{p}_i^{n+1}(v_J)C_{iJ}\tilde{p}_i^n(v_J)\\
&\qquad+\frac{2}{d_J}\sum_{i\in\mathcal{E}_J}C_{iJ}\tilde{p}_i^{n+1}(v_J)\sum_{j\in\mathcal{E}_J}\epsilon_{jJ}(C_{jJ}^+)^T[K_j(\tilde{r}_j^{n\prime}+R_j\tilde{r}_j^n)](v_J)\\
&\qquad+\sum_{i\in\mathcal{E}_J}\epsilon_{iJ}(C_{iJ}^+)^T[K_i(\tilde{r}_i^{n\prime}+R_i\tilde{r}_i^n)](v_J)C_{iJ}\tilde{p}_i^{n+1}(v_J)\Bigg]\\
&\quad-2\sum_{i\in\mathcal{E}_J}\left(\frac{2}{d_J}\sum_{j\in\mathcal{E}_J}C_{jJ}\tilde{p}_j^n(v_J)-C_{iJ}\tilde{p}_i^n(v_J)\right)\\
&\quad-\left(\frac{2}{d_J}\sum_{j\in\mathcal{E}_J}\epsilon_{jJ}(C_{jJ}^+)^T[K_j(\tilde{r}_j^{n\prime}+R_i\tilde{r}_j^n)](v_J)-\epsilon_{iJ}(C_{iJ}^+)^T[K_i(\tilde{r}_i^{n\prime}+R_i\tilde{r}_i^n)](v_J)\right).
\end{aligned}
\tag{5.5}
$$

Again, a similar expression can be derived for the adjoint variable. Factoring the last term in (5.5), we see that it reduces to

$$-2\sum_{i\in\mathcal{E}_J}\left\{\epsilon_{iJ}(C_{jJ}^{+})^T[K_i(\tilde{r}_i^{n\prime}+R_i\tilde{r}_i^n)](v_J)C_{iJ}\tilde{p}_i^n(v_J)\right\}\,.$$

Furthermore, the bracketed term in (5.5) is minus the right-hand side of (5.2). Therefore, we arrive at

$$\begin{aligned}
&\sum_{i\in\mathcal{E}_J}|(C_{iJ}^{+})^T[K_i((\tilde{r}_i^{n+1})'+R_i\tilde{r}_i^{n+1})](v_J)|^2+\sum_{i\in\mathcal{E}_J}|C_{iJ}\tilde{p}_i^{n+1}(v_J)|^2\\
&=-2\sum_{i\in\mathcal{E}_J}\epsilon_{iJ}(C_{iJ}^{+})^T[K_i((\tilde{r}_i^{n+1})'+R_i\tilde{r}_i^{n+1})](v_J)C_{iJ}\tilde{p}_i^{n+1}(v_J)\\
&-2\sum_{i\in\mathcal{E}_J}\epsilon_{iJ}(C_{iJ}^{+})^T[K_i(\tilde{r}_i^{n\prime}+R_i\tilde{r}_i^n)](v_J)C_{iJ}\tilde{p}_i^n(v_J)\\
&+\sum_{i\in\mathcal{E}_J}\left|(C_{iJ}^{+})^T[K_i(\tilde{r}_i^{n\prime}+R_i\tilde{r}_i^n)](v_J)\right|^2+\sum_{i\in\mathcal{E}_J}\left|C_{iJ}\tilde{p}_i^n(v_J)\right|^2.
\end{aligned}\tag{5.6}$$

An analogous relation holds for the adjoint variable, but with the opposite signs at the mixed terms. We sum up (5.6) and its counterpart and integrate with respect to time. We obtain for $v_J\in V_M$,

$$\begin{aligned}
\tilde{E}_J^{n+1}:=\int_0^T\sum_{i\in\mathcal{E}_J}\Big\{&|(C_{iJ}^{+})^T[K_i((\tilde{r}_i^{n+1})'+R_i\tilde{r}_i^{n+1})](v_J)|^2\\
&+|(C_{iJ}^{+})^T[K_i((\tilde{p}_i^{n+1})'+R_i\tilde{p}_i^{n+1})](v_J)|^2\\
&+|C_{iJ}\tilde{r}_i^{n+1}(v_J)|^2+|C_{iJ}\tilde{p}_i^{n+1}(v_J)|^2\Big\}\,dt=\tilde{E}_J^n\\
+2\int_0^T\sum_{i\in\mathcal{E}_J}\big[&\epsilon_{iJ}(C_{iJ}^{+})^T[K_i((\tilde{p}_i^{n+1})'+R_i\tilde{p}_i^{n+1})](v_J)C_{iJ}\tilde{r}_i^{n+1}(v_J)\\
&-\epsilon_{iJ}(C_{iJ}^{+})^T[K_i((\tilde{r}_i^{n+1})'+R_i\tilde{r}_i^{n+1})](v_J)C_{iJ}\tilde{p}_i^{n+1}(v_J)\big]\,dt\\
+\,2\int_0^T\sum_{i\in\mathcal{E}_J}\big[&\epsilon_{iJ}(C_{iJ}^{+})^T[K_i(\tilde{p}_i^{n\prime}+R_i\tilde{p}_i^{n+1})](v_J)C_{iJ}\tilde{r}_i^n(v_J)\\
&-\epsilon_{iJ}(C_{iJ}^{+})^T[K_i(\tilde{r}_i^{n\prime}+R_i\tilde{r}_i^n)](v_J)C_{iJ}\tilde{p}_i^n(v_J)\big]\,dt.
\end{aligned}\tag{5.7}$$

We first note that $C_{iJ}^{+}\,(C_{iJ}^{+})^T$ is symmetric and positive definite, and so is K_i. Hence, $\tilde{E}_J^n$ is equivalent to the "energy trace" at v_J

$$\begin{aligned}
E_J^n:=\int_0^T\sum_{i\in\mathcal{E}_J}\Big\{&K_i(\tilde{r}_i^{n\prime}+R_i\tilde{r}_i)(v_J)\cdot(\tilde{r}_i^{n\prime}+R_i\tilde{r}_i^n)(v_J)\\
&+K_i(\tilde{p}_i^{n\prime}+R_i\tilde{p}_i^n)(v_J)\cdot(\tilde{p}_i^{n\prime}+R_i\tilde{p}_i^n)(v_J)\\
&+|C_{iJ}\tilde{r}_i^n(v_J)|^2+|C_{iJ}\tilde{p}_i^n(v_J)|^2\Big\}\,dt.
\end{aligned}\tag{5.8}$$

Let us now recall that the total energy associated with (1.1), (1.2) is

$$\begin{aligned}\mathcal{E}(r,t) &:= \frac{1}{2}\sum_{i=1}^{n_e}\int_0^{l_i}\left[M_i\dot r_i\cdot\dot r_i + K_i(r_i'+R_ir_i)\cdot(r_i'+R_ir_i)+S_ir_i\cdot r_i\right]dx \\ &=: \sum_{i=1}^{n_e}\mathcal{E}_i(r_i,t)\ .\end{aligned}\tag{5.9}$$

Now, according to [14, Lemma 3.2, Chapter IV], there is a constant C such that

$$\begin{aligned}&\int_0^T\left[M_i\dot r_i\cdot\dot r_i + K_i(r_i'+R_ir_i)\cdot(r_i'+R_ir_i)+S_ir_i\cdot r_i\right](v_J)\,dt \\ &\leq C\left\{\int_0^t \mathcal{E}(r_i,s)ds + \mathcal{E}_i(r_i,0)+\mathcal{E}_i(r_i,t)\right\}\end{aligned}\tag{5.10}$$

for sufficiently regular solutions of (1.5). Moreover, for solutions r_i of (1.5) satisfying (1.9) and with the understanding that $C_{iJ}=I$ for $v_J\in V_S$ ($\epsilon_{iJ}\neq 0$),

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(r,t) &= \sum_J\sum_{i\in\mathcal{E}_J}\epsilon_{iJ}(C_{iJ}^+)^T[K_i(r_i'+R_ir_i)](v_J)C_{iJ}\dot r_i(v_J) \\ &\leq \sum_J\sum_{i\in\mathcal{E}_J}\left\{\frac{1}{4\epsilon}|(C_{iJ}^+)^T[K_i(r_i'+R_ir_i)](v_J)|^2+\epsilon|C_{iJ}\dot r_i(v_J)|^2\right\}.\end{aligned}\tag{5.11}$$

Upon integration, (5.11) leads to

$$\begin{aligned}\mathcal{E}(r,t)\leq \mathcal{E}(r,0)+\int_0^t\sum_J\sum_{i\in\mathcal{E}_J}\Bigg\{&\frac{1}{4\epsilon}|(C_{iJ}^+)^T[K_i(r_i'+R_ir_i)](v_J)|^2 \\ &+\epsilon|C_{iJ}\dot r_i(v_J)|^2\Bigg\}\,ds,\end{aligned}$$

which by (5.10) yields

$$\begin{aligned}\mathcal{E}(r,t)\leq \mathcal{E}(r,0)+\frac{1}{4\epsilon}\int_0^t\sum_J\sum_{i\in\mathcal{E}_J}&|(C_{iJ}^+)^T[K_i(r_i'+R_ir_i)](v_J)|^2\,dt \\ &+\epsilon C\left\{\int_0^t\mathcal{E}(r,s)\,ds+\mathcal{E}(r,0)+\mathcal{E}(r,t)\right\}\ .\end{aligned}\tag{5.12}$$

Now, absorbing $\epsilon C\,\mathcal{E}(r,t)$ into the left side of (5.12) and using Gronwall's inequality we obtain

$$\mathcal{E}(r,t)\leq C\left\{\mathcal{E}(r,0)+\int_0^t\sum_J\sum_{i\in\mathcal{E}_J}|(C_{iJ}^+)^T[K_i(r_i'+R_ir_i)](v_J)|^2\,dt\right\}\tag{5.13}$$

with a generic constant $C>0$. We may apply the energy estimate to our local solutions $\tilde r_i^n,\tilde p_i^n$, as no coupling conditions have been used to derive (5.13). Even though, by Remark 4.1, we can achieve energy-space regularity for the local adjoint system, if we assume $H^1(0,T)$-regularity of λ_{iJ}, we want to derive energy estimates in the shifted energy space, as we want to deal with $\lambda's$ being in L^2. The application of estimate (5.13) to $\tilde{p}_i^n$ is done as follows. Assuming, as always, L^2-regularity of the data μ_{iJ}^n (see (5.21) below for the justification), and using the fact that $\tilde{r}_i^n(v_J)$ is also L^2-in-time, we have a Neumann-type problem for $\tilde{p}_i^n$with L^2-boundary inputs and final data in the shifted energy space. By transposition, we obtain the estimate (5.13) with $\mathcal{E}(r,t)$ and $\mathcal{E}(r,0)$ replaced by $\mathcal{E}^*(p,t)$ and $\mathcal{E}^*(p,T)$, respectively. However, because $\|v\|_{\mathbf{V}^*} = \|A^{-1}v\|_{\mathbf{V}}$ we have that $\mathcal{E}^*(\tilde{p}^n,T) := \|\tilde{p}^n(T)\|_{\mathbf{H}} + \|\dot{\tilde{p}}^n(T)\|_{\mathbf{V}^*}$ and $k^2\mathcal{E}(\tilde{r}^n,T)$ are equivalent. Since $\mathcal{E}(\tilde{r}^n,0)=0$, we therefore obtain

$$\begin{aligned}\mathcal{E}^*(\tilde{p}^n,t) \le C\int_0^T \sum_J \sum_{i\in\mathcal{E}_J} \Bigg\{ &|(C_{iJ}^+)^T[K_i(\tilde{p}_i^{n\prime}+R_i\tilde{p}_i^n)](v_J)|^2 \\ &+|(C_{iJ}^+)^T[K_i(\tilde{r}_i^{n\prime}+R_i\tilde{r}_i^n)](v_J)|^2 \Bigg\}\, dt\ ,\end{aligned} \tag{5.14}$$

whenever the right-hand side is defined. Note that the summation is taken over all joint indices J such that $v_J \in V_M \cup V_N \cup V_C$. Therefore, (5.14) can be rewritten as

$$\begin{aligned}\mathcal{E}^*(\tilde{p}^n,t) \le C\int_0^T \sum_{v_J\in V_M} \sum_{i\in\mathcal{E}_J} \Bigg\{ &|(C_{iJ}^+)^T[K_i(\tilde{p}_i^{n\prime}+R_i\tilde{p}_i^n)](v_J)|^2 \\ &+ |(C_{iJ}^+)^T[K_i(\tilde{r}_i^{n\prime}+R_i\tilde{r}_i^n)](v_J)|^2 \Bigg\}\, dt \\ + C\int_0^T \sum_{v_C\in V_C} &|(\tilde{p}_{i_C}^n(v_C))|^2\, dt \qquad (\mathcal{E}_C=\{i_C\}).\end{aligned} \tag{5.15}$$

By applying the same argument to $(\tilde{r}_i^n)$ and denoting $\tilde{E}^n = \sum_J \tilde{E}_J^n$, $J : v_J \in V_M$, we obtain

$$\mathcal{E}(\tilde{r}^n,t) + \mathcal{E}^*(\tilde{p}^n,t) \le C\left(\tilde{E}^n + \int_0^T \sum_{v_C\in V_C} |\tilde{p}_{i_C}^n(v_C)|^2\, dt\right). \tag{5.16}$$

Again, this estimate holds as long as the right-hand side is finite. This will turn out to be true in the following.

We go back to (5.7) and evaluate the mixed terms on the right-hand side:

$$\begin{aligned}0 &= \int_0^T\int_0^{l_i} [M_i\ddot{\tilde{p}}_i^n - (K_i((\tilde{p}_i^n)' + R_i\tilde{p}^n))' + R_i^T(K_i(\tilde{p}_i^{n\prime}+R_i\tilde{p}_i^n)) \\ &\qquad\qquad + S_i\tilde{p}_i^n]\tilde{r}_i^n\, dx\, dt \\ &= \int_0^{l_i} (M_i\dot{\tilde{p}}_i^n\tilde{r}_i^n - M_i\tilde{p}_i^n\dot{\tilde{r}}_i^n)|_0^T\, dx \\ &\quad + \int_0^T\int_0^{l_i} \tilde{p}_i^n M_i\ddot{\tilde{r}}_i^n\, dx\, dt - \sum_J \int_0^T \epsilon_{iJ}K_i(\tilde{p}_i^{n\prime}+R_i\tilde{p}_i^n)(v_J)\tilde{r}_i^n(v_J)\, dt \\ &\quad + \int_0^T\int_0^{l_i} (\tilde{p}_i^{n\prime}+R_i\tilde{p}_i^n)K_i\tilde{r}_i^{n\prime}\, dx\, dt\end{aligned}$$

$$\begin{aligned}
&+\sum_J \epsilon_{iJ}\int_0^T \tilde p_i^n(v_J)(K_iR_i\tilde r_i^n)(v_J)\,dt - \int_0^T\int_0^{l_i}\tilde p_i^n(K_iR_i\tilde r^n)'\,dx\,dt\\
&+\int_0^T\int_0^{l_i}\tilde p_i^n(R_i^TK_iR_i\tilde r_i^n + S_i\tilde r_i^n)\,dx\,dt\\
=&\int_0^T\sum_J\epsilon_{iJ}[K_i(\tilde r_i^{n\prime}+R_i\tilde r_i^n)](v_J)\tilde p_i^n(v_J)\,dt\\
&-\int_0^T\sum_J\epsilon_{iJ}[K_i(\tilde p_i^{n\prime}+R_i\tilde p_i^n)](v_J)\tilde r_i^n(v_J)\,dt\\
&-k\|\tilde r_i^n(T)\|^2_{\mathbf{V}_i} - k\|\dot{\tilde r}_i^n(T)\|^2_{\mathbf{H_i}}\ ,
\end{aligned}\tag{5.17}$$

where $\mathbf{V}_i := H^1(0,\ell_i)^{p_i}$ and $\mathbf{H}_i := L^2(0,\ell_i)^{p_i}$. As $(Ar)_i =: A_ir_i$, A_i is the Riesz isomorphism between V_i and V_i^* with respect to the norm

$$\|r_i\|^2_{\mathbf{V_i}} = \int_0^{l_i}[K_i(r_i'+R_ir_i)(r_i'+R_ir_i)+(S_i+I)r_ir_i]\,dx.$$

Now

$$\begin{aligned}
\tilde p_i^n(v_J) &= \Pi_{iJ}^{\perp}\tilde p_i^n(v_J) + \Pi_{iJ}\tilde p_i^n(v_J)\\
&= C_{iJ}^{+}C_{iJ}\tilde p_i^n(v_J) + \Pi_{iJ}\tilde p_i^n(v_J),
\end{aligned}$$

and hence, by (1.9)

$$\begin{aligned}
&\epsilon_{iJ}[K_i(\tilde r_i^{n\prime}+R_i\tilde r_i^n)](v_J)\tilde p_i^n(v_J)\\
&=\epsilon_{iJ}(C_{iJ}^{+})^T[K_i(\tilde r_i^{n\prime}+R_i\tilde r_i^n)](v_J)C_{iJ}\tilde p_i^n(v_J)\\
&\qquad+\epsilon_{iJ}\Pi_{iJ}[K_i(\tilde r_i^{n\prime}+R_i\tilde r_i^n)](v_J)\tilde p_i^n(v_J)\\
&=\epsilon_{iJ}(C_{iJ}^{+})^T[K_i(\tilde r_i^{n\prime}+R_i\tilde r_i^n)](v_J)C_{iJ}\tilde p_i^n(v_J).
\end{aligned}\tag{5.18}$$

The same holds for the other terms in (5.17). We thus have by (5.17), (5.18)

$$\begin{aligned}
&\int_0^T\sum_{v_J\in V_M}\epsilon_{iJ}(C_{iJ}^{+})^T[K_i(\tilde r_i^{n\prime}+R_i\tilde r_i^n)](v_J)C_{iJ}\tilde p_i^n(v_J)\,dt\\
&-\int_0^T\sum_{v_J\in V_M}\epsilon_{iJ}(C_{iJ}^{+})^T[K_i(\tilde p_i^{n\prime}+R_i\tilde p_i^n)](v_J)C_{iJ}\tilde r_i^n(v_J)\,dt\\
&= k\left\{\|\tilde r_i^n(T)\|^2_{\mathbf{V}_i}+\|\tilde r_i^n(T)\|^2_{\mathbf{H_i}}\right\}+\sum_{v_C\in V_C}\int_0^T|\tilde p_{i_C}^n(v_C)|^2\,dt.
\end{aligned}\tag{5.19}$$

Upon using (5.19) in (5.7) we obtain

$$\begin{aligned}
\tilde E^{n+1} = \tilde E^n &- 2k\big\{\|\tilde r^n(T)\|^2_{\mathbf V}+\|\dot{\tilde r}^n(T)\|^2_{\mathbf H}\\
&\qquad +\|\tilde r^{n+1}(T)\|^2_{\mathbf V}+\|\dot{\tilde r}^{n+1}(T)\|^2_{\mathbf H}\big\}\\
&-2\Bigg\{\int_0^T\sum_{V_C\in V_C}|\tilde p_{i_C}^n(v_C)|^2\,dt\\
&\qquad+\int_0^T\sum_{V_C\in V_C}|\tilde p_{i_C}^{n+1}(v_C)|^2\,dt\Bigg\}.
\end{aligned}\tag{5.20}$$

Taking the initial data to be zero, we iterate the recursion (5.20) down to the index zero:

$$
\begin{aligned}
\tilde{E}^{n+1} = \tilde{E}^0 &- 4k \sum_{k=0}^{n+1}{}' \big\{ \|\tilde{r}^k(T)\|^2_{\mathbf{V}} + \|\dot{\tilde{r}}^k(T)\|^2_{\mathbf{H}} \big\} \\
&- 4k \sum_{k=0}^{n+1}{}' \sum_{v_C \in V_C} \int_0^T |\tilde{p}^k_{i_C}(v_C)|^2 \, dt,
\end{aligned}
\tag{5.21}
$$

where $\sum_{i=0}^{n}{}' a_i := \frac{1}{2} a_0 + \frac{1}{2} a_n + \sum_{j=1}^{n-1} a_j$. Letting n tend to infinity we conclude

$$
\begin{cases}
\displaystyle\sum_{k=0}^{\infty} & \big\{ \|\tilde{r}^k(T)\|^2_{\mathbf{V}} + \|\dot{\tilde{r}}^k(T)\|^2_{\mathbf{H}} \big\} < \infty, \\
\displaystyle\sum_{k=0}^{\infty} \sum_{v_C \in V_C} & \displaystyle\int_0^T |\tilde{p}^k_{i_C}(v_C)|^2 \, dt < \infty, \\
 & \tilde{E}^n \text{ bounded,}
\end{cases}
\tag{5.22}
$$

whenever E^0 is finite. This can be achieved by properly choosing the initial guess for the iterates $\hat{r}^0_i$, $\hat{p}^0_i$ and taking initial and target data of the global system to be in $\mathbf{D}(\mathbf{A}) \times \mathbf{V}$.

Going back to (4.10), (4.11) (with $\beta = 1$ for simplicity) it is easily seen that the recursion (5.20) takes the equivalent form:

$$
\begin{aligned}
\int_0^T \sum_{J \in V_M, i \in \mathcal{E}_J} \|(\tilde{g}^{n+1}_{ij}, \tilde{h}^{n+1}_{iJ})\|^2 dt &= \int_0^T \sum_{J \in V_M, i \in \mathcal{E}_J} \|(\tilde{g}^{n}_{ij}, \tilde{h}^{n}_{iJ})\|^2 dt \\
-2k\big\{ \|\tilde{r}^n(T)\|^2_{\mathbf{V}} + \|\dot{\tilde{r}}^n(T)\|^2_{\mathbf{H}} &+ \|\tilde{r}^{n+1}(T)\|^2_{\mathbf{V}} + \|\dot{\tilde{r}}^{n+1}(T)\|^2_{\mathbf{H}} \big\} \\
- 2 \left\{ \int_0^T \sum_{V_C \in V_C} |\tilde{p}^n_{i_C}(v_C)|^2 \, dt \right. &+ \left. \int_0^T \sum_{V_C \in V_C} |\tilde{p}^{n+1}_{i_C}(v_C)|^2 \, dt \right\}.
\end{aligned}
\tag{5.23}
$$

Of course, the analogue of (5.21) and the conclusion that the $L^2(0,T)$-norm of $(\tilde{g}, \tilde{h})_{J \in V_M, i \in \mathcal{E}_J}$ remains bounded also hold.

We conclude from (5.22) that

$$
\begin{aligned}
&(\tilde{r}^n(T), \dot{\tilde{r}}^n(T)) \longrightarrow 0 \text{ in } \mathbf{V} \times \mathbf{H}, \\
&\tilde{p}^n_{i_C}(v_C) \longrightarrow 0 \text{ in } L^2(0,T).
\end{aligned}
\tag{5.24}
$$

But, as $\tilde{E}^n$ is bounded in n by (5.22), we obtain from (5.24) and (5.16) that

$$
\mathcal{E}_i(\tilde{r}^n_i, t), \; \mathcal{E}^*_i(\tilde{p}^n_i, t) \le C \qquad \forall\, i, n, t.
\tag{5.25}
$$

Thus, on subsequences

$$
\begin{cases}
(\tilde{r}^n_i, \dot{\tilde{r}}^n_i) \longrightarrow (\eta_i, \dot{\eta}_i), & w - \star - \text{in } L^\infty(0, T, \mathbf{V}_i \times \mathbf{H}_i), \\
(\tilde{p}^n_i, \dot{\tilde{p}}^n_i) \longrightarrow (\mu_i, \dot{\mu}_i), & w - \star - \text{in } L^\infty(0, T, \mathbf{H}_i \times \mathbf{V}^*_i),
\end{cases}
\tag{5.26}
$$

$$
\begin{cases}
C_{iJ} \tilde{r}^n_i(v_J) \longrightarrow \alpha_{iJ}, \; C_{iJ} \tilde{p}^n_{ij}(v_J) \longrightarrow \gamma_{iJ}, \\
(C^+_{iJ})^T [K_i(\tilde{r}^{n\prime}_i + R_i \tilde{r}^n_i)](v_J) \longrightarrow \beta_{iJ}, \\
(C^+_{iJ})^T [K_i(\tilde{p}^{n\prime}_i + R_i \tilde{p}^n_i)](v_J) \longrightarrow \delta_{iJ}, \qquad w - \text{in } L^2(0,T).
\end{cases}
\tag{5.27}
$$

However, as only a subsequence is involved, we cannot use (4.1), (4.2) directly in order to conclude convergence to the global solutions. We therefore resort to a result given by Opial [33] as follows.

We consider the map taking $(\tilde{g}_{iJ},\ \tilde{h}_{iJ})$ of (4.10) and (4.11) into the right-hand sides of those equations (for the errors), where we delete the iteration index n. That is, given $(\tilde{g}_{iJ},\ \tilde{h}_{iJ})$ we solve the local error system with mixed Robin conditions (4.10), (4.11) and inhomogeneities $(\tilde{g}_{iJ},\ \tilde{h}_{iJ})$ for $\tilde{r}_j,\ \tilde{p}_j$, and then compute all the traces appearing in (4.10), (4.11) on the right-hand side (again with the iteration index n deleted). To be precise, we introduce a map S_J by

$$(S_J(z))_i := \frac{2}{d_J} \sum_{j \in \mathcal{E}_J} z_{jJ} - z_{iJ}.$$

Note that this map satisfies $S_J^2 = I$. We identify network quantities, such as $C_{iJ}\tilde{r}_i(V_J)$, with $\tilde{r}_{iJ}$. With this notation, (4.10), (4.11) can be written briefly as

$$(\tilde{g}_{iJ}^{n+1}, \tilde{h}_{iJ}^{n+1}) = ((S_J(2\tilde{p}^n - \tilde{g}^n))_i, S_J(-2\tilde{r}^n - \tilde{h}^n)_i).$$

In an obvious way we can collect all the components into single vectors $\tilde{g}, \tilde{h}, \tilde{p}, \tilde{r}$. We extend the map S_J to all multiple nodes and denote it by S. Now, we define the map $\mathcal{T}$ by

$$\mathcal{T}(\tilde{g}, \tilde{h}) := S(2(\tilde{p}, -\tilde{r}) - (\tilde{g}, \tilde{h})). \tag{5.28}$$

Speaking in terms of our iteration (involving either (4.1), (4.2) or (4.10), (4.11) (with $\beta = 1$)), the new boundary terms $\tilde{g}_{iJ}^{n+1}, \tilde{h}_{iJ}^{n+1}$ are obtained from the previous ones by applying $\mathcal{T}$; i.e., with $x := (\tilde{g}, \tilde{h})$ we have a fixed point iteration

$$x^{n+1} = \mathcal{T}(x^n). \tag{5.29}$$

If $\beta > 0$ we have instead $\mathcal{T}(\tilde{g}, \tilde{h}) := S(2\beta(\tilde{p}, -\tilde{r}) - (\tilde{g}, \tilde{h}))$.

Now, the existence of a fixed-point is guaranteed by the existence (and in fact uniqueness) of solutions to the global optimality system. It is apparent from (5.23) (or equivalently (5.20)) that the mapping $\mathcal{T}$ is nonexpansive, i.e., $\|\mathcal{T}(x)\| \le \|x\|$. We are going to show that the map $\mathcal{T}$ is asymptotically regular in the sense of Opial [33], i.e., $\mathcal{T}^{n+1}x - \mathcal{T}^n x = \mathcal{T}^n(\mathcal{T} - I)x \to 0$. Then, according to Opial [33, Theorem 2], the sequence defined by (5.29) converges weakly to a (in fact the unique) fixed point. Now, with $(\bar{\tilde{g}}, \bar{\tilde{h}}) := \mathcal{T}(\tilde{g}, \tilde{h})$ we obtain by (5.23)

$$\begin{aligned} \|(\bar{\tilde{g}}, \bar{\tilde{h}})\|^2_{L^2(0,T;\mathbf{R}^{ne \times \# V_M})} &= \|(T(\tilde{g}, \tilde{h}))\|^2_{L^2} \\ &= \|(\tilde{g}, \tilde{h})\|^2_{L^2} - 2k\big\{\|\tilde{r}(T)\|^2_{\mathbf{V}} + \|\dot{\tilde{r}}(T)\|^2_{\mathbf{H}} \\ &\qquad + \|\bar{\tilde{r}}(T)\|^2_{\mathbf{V}} + \|\dot{\bar{\tilde{r}}}(T)\|^2_{\mathbf{H}}\big\} \\ &\quad - 2\left\{ \int_0^T \sum_{V_C \in V_C} |\tilde{p}_{i_C}(v_C)|^2 \, dt + \int_0^T \sum_{V_C \in V_C} |\bar{\tilde{p}}_{i_C}(v_C)|^2 \, dt \right\}. \end{aligned} \tag{5.30}$$

Assume that λ is an eigenvalue of $\mathcal{T}$ corresponding to the eigenpair (g, h). Then

$$
\begin{aligned}
(5.31)\qquad \|T(\tilde g, \tilde h)\|^2 &= |\lambda|^2 \|(\tilde g, \tilde h)\|^2\\
&= \|(\tilde g, \tilde h)\|^2 - 2k\big\{\|\tilde r(T)\|^2_{\mathbf{V}} + \|\dot{\tilde r}(T)\|^2_{\mathbf{H}}\\
&\qquad\qquad + \|\bar{\tilde r}(T)\|^2_{\mathbf{V}} + \|\dot{\bar{\tilde r}}(T)\|^2_{\mathbf{H}}\big\}\\
&\quad - 2\left\{\int_0^T \sum_{V_C \in V_C} |\tilde p_{i_C}(v_C)|^2 \, dt\right.\\
&\qquad\qquad \left. + \int_0^T \sum_{V_C \in V_C} |\bar{\tilde p}_{i_C}(v_C)|^2 \, dt\right\}.
\end{aligned}
$$

Dividing by $\|(\tilde g, \tilde h)\|^2 \neq 0$, we see that $|\lambda| \leq 1$. Assume now that $\lambda = 1$. Then equality (5.31) immediately gives

$$\tilde r(T) = \dot{\tilde r}(T) = \bar{\tilde r}(T) = \dot{\bar{\tilde r}}(T) = 0,\ \tilde p_{i_C}(v_C) = \bar{\tilde p}_{i_C}(v_C) = 0.$$

As shown above this implies that $\tilde r = \tilde p = \bar{\tilde r} = \bar{\tilde p} = 0$. Hence, all eigenvalues of $\mathcal{T}$ have modulus less than 1. We recall the structure of T given by (5.28). Up to the isomorphism S this map is the identity shifted by a map that takes appropriate traces of $\tilde r, \tilde p$ at all multiple nodes. Now, by our "energy inequalities" (5.20), (5.21), we know that even the Neumann traces at those nodes are L^2 in time. However, $\tilde r, \tilde p$ solve hyperbolic equations on each edge. Interchanging the space and time variables (as usual in one-dimensional problems) one obtains the same regularity for the velocities at multiple nodes. Therefore, picking Dirichlet data is a compact operation. Hence, the map $\mathcal{T}$ is—up to an isomorphism—the identity plus a compact operator K. Therefore, K and hence the entire map $\mathcal{T}$ is completely determined by its point spectrum and the essential spectrum at $\lambda = 1$. In terms of asymptotic regularity, the worst that can happen is that $x = \lim_{k\to\infty} x_k$, where x_k is the eigenelement $(= (\tilde g_k, \tilde h_k))$ corresponding to $\lambda_k \to 1$. On this sequence we have $\|\mathcal{T}^n(\mathcal{T} - I)x_k\| = |\lambda_k^n(\lambda_k - 1)| \to 0$. Thus, $\mathcal{T}$ is asymptotically regular and Opial's result applies.

We conclude by the same argument as in (4.3)–(4.6) that, in a weak-L^2 sense, all boundary and transmission conditions are satisfied in the limit. In fact, as seen above, the convergence of traces at multiple nodes takes place in the strong sense, and going back to the original iteration, we infer strong convergence of Neumann traces there. Therefore, we have strong convergence of the solutions $\tilde r_i^n$ in $C(0, T; H^1(0, \ell_i)) \cap C^1(0, T; L^2(0, \ell_i))$, and $\tilde p_i^n$ in $C(0, T; L^2(0, \ell_i)) \cap C^1(0, T; (H^1(0, \ell_i))^*)$. On the other hand by (5.24) all global and, hence, local final conditions for the adjoint system strongly tend to zero, and so do the traces at controlled nodes. The strong limits $\tilde r_i$ and $\tilde p_i$ of the local solutions thus correspond to zero initial and final data, respectively. Also, in the limit the global transmission conditions are satisfied in the strong sense. By uniqueness, the limits have to be equal to zero. We obtain Theorem 5.1.

THEOREM 5.1. *Let* (r_0, r_1), $(z_0, z_1) \in \mathbf{D}(\mathbf{A}) \times \mathbf{V}$ *and assumption* (A) *be satisfied. Then the sequence* $(r_i^n, p_i^n)_i$, $i = 1 : n_e$, $n \in \mathbb{N}_0$, *defined by the solution to* (4.12)–(4.18) *(or* (4.19), (4.20)*) converges to* (r_i, p_i), $i = 1 : n_e$ *in* $C(0, T; V \times H) \cap C^1(0, T; H \times V^*)$, *where* $(r, p) = (r_i, p_i)_{i=1}^{n_e}$ *solves the global optimality system* (1.5)–(1.10), (3.1)–(3.7).

This result can be extended to the relaxed version (4.8), (4.9) of (4.1), (4.2) as well as to the relaxed version of (4.10), (4.11). The point is that by underrelaxation the iteration mapping $\mathcal{T}_\lambda := \lambda I + (1 - \lambda)\mathcal{T}$ has the same set of fixed points as the original map $\mathcal{T}$, and, according to Opial [33, Theorem 3], the map $\mathcal{T}_\lambda$ is asymptotically regular. Hence for each $\lambda \in (0,1)$ and for any initial data, the fixed point sequence converges weakly in L^2 to the unique fixed point of $\mathcal{T}$. Therefore, independently of λ, the transmission conditions are satisfied in the weak limit and, by the argument given above, also in the strong limit. Now, in order to conclude convergence of the relaxed scheme, one needs to extend the basic recursion relation (5.20) to the relaxed scheme, as one has to use the fact that the final values and the traces at controlled nodes tend to zero also for the sequence generated by the relaxed scheme. This extension is tedious but nevertheless possible.

6. Remarks on other cost functionals. One might consider functions different from (2.1). In particular, we might want to optimize the flux of energy, rather than to dissipate energy. This is of interest in real-time applications where one wants to protect certain subregions from perturbations. Without going into details, we mention that in the string case the flux at a given vertex v_J in the direction e_i of edge i is given by

$$\begin{aligned} &-\epsilon_{iJ}K_i r_i'(v_J)\dot{r}_i(v_J) \\ &\quad = \frac{1}{4}\left\{|\epsilon_{iJ}K_i r_i'(v_J) - \dot{r}_i(v_J)|^2 - |\epsilon_{iJ}K_i r_i'(v_J) + \dot{r}_i(v_J)|^2\right\}, \end{aligned}$$

so that maximizing the flux of energy at v_J in that direction comes down to minimizing the quadratic cost

$$\frac{k}{2}\int_0^T |\epsilon_{iJ}K_i r_i'(v_J) + \dot{r}_i(v_J)|^2\, dt.$$

It is apparent that upon letting k grow large the optimal control would have the tendency to install an absorbing condition at v_J. More on this issue can be found in [17]. The complete analysis of the corresponding domain decomposition is yet to be done. Similar cost criteria apply also to the Timoshenko beam case.

Another cost function, which is much simpler and more classical, is provided by the full-state cost

$$J(f) = \frac{1}{2}\sum_{v_J \in V_C}\int_0^T |f_J|^2\, dt + \frac{1}{2}\int_0^T \sum_{i=1}^{n_e} |r_i|^2\, dx\, dt. \tag{6.1}$$

This is the kind of cost function that has been frequently used by Benamou [1], [2], etc.. The point is that recursions similar to (5.21) involve the error with respect to the entire state, so that convergence follows without the necessity of applying energy estimates. In our present situation, we are able to show that in the case of (6.1), (5.21) changes to

$$\begin{aligned} \tilde{E}^{n+1} = \tilde{E}^0 - 4\sum_{k=0}^{n+1'}\sum_{i=1}^{n_e}\int_0^T\int_0^{l_i} |\tilde{r}_i^k|^2\, dx\, dt \\ -4\sum_{k=0}^{n+1'}\sum_{v_C \in V_C}\int_0^T |\tilde{p}_{iC}^k(v_C)|^2\, dt. \end{aligned} \tag{6.2}$$

In this case it is obvious that we obtain convergence in $L^2(0,T,H)$ of the corresponding errors directly. We note that, by standard arguments, the coupling in the local and global optimality systems is much simpler then, as the full state r_i appears as distributed input in the adjoint equation.

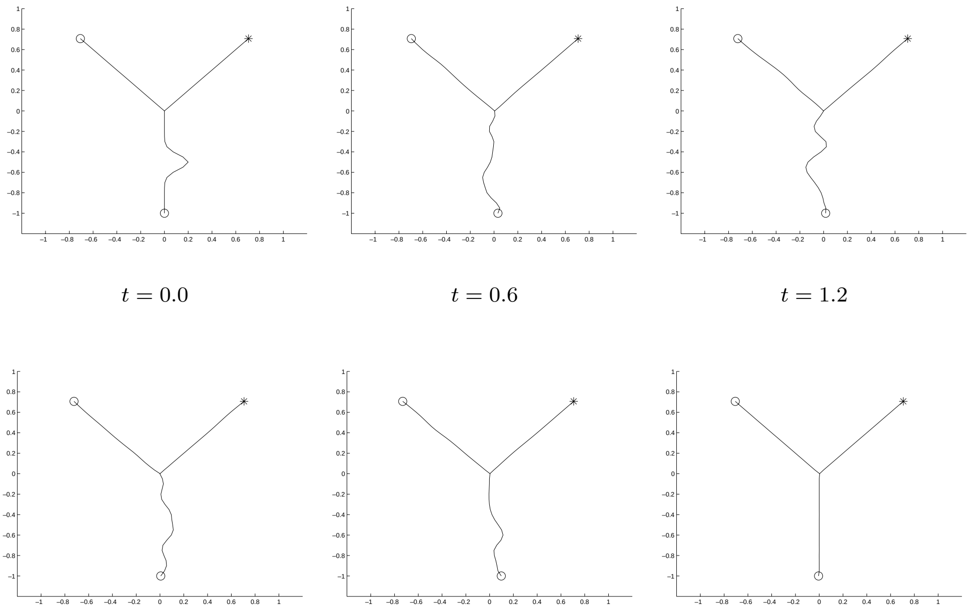
As for cost function (2.2), we remark that results similar to those corresponding to cost function (2.1) can be shown. We refer the reader to Lagnese and Leugering [16], where transmission problems in higher dimensions are discussed. See also Leugering [19] for a discussion of this problem in the context of networks of strings.

We also remark that it is possible to show similar results—in the case of full state cost—for networks of Rayleigh beams. This can be done by letting the shear stiffness of each member tend to infinity; see Example 2. The analysis, however, goes beyond the scope of this paper. We shall consider this problem in a forthcoming publication.

7. Numerical simulations. We consider only networks of strings. For Timoshenko beams the simulations are very similar. As an example we take a tripod (divining rod) consisting of one-dimensional elastic material. We put Neumann boundary controls at the south node and the northwest node, whereas the northeast node is clamped. This system is known to be exactly controllable in $T \geq 4$ time units *in general*, that is, for *general* initial conditions [14]. We have picked $T = 3$ as final time because this particular example shows that controls can be applied successfully in a shorter time interval for a particular choice of the initial data. However, as is obvious, a control time of one time unit is not sufficient to steer the system to rest. This is verified numerically in Figure 7.1. Figures 7.2 and 7.3 show a rectangular tripod with the south vertex controlled in the Neumann condition. In Figure 7.2 the west node is clamped, while the east node is subjected to a free Neumann condition. That system is also known to be controllable [18]. Here it really takes 4 time units (a shorter time duration is not sufficient in this configuration; we have only one control) to control the configuration to rest. In Figure 7.3, however, the east and west nodes are clamped. This situation is known to be uncontrollable [18] in any time interval by Neumann controls at the south node. The underlying phenomenon is exactly as in the midspan control case for a string clamped at its boundaries. Figure 7.4 clearly shows that the upright element does not absorb any energy from the horizontal members, simply because a zero node deflection occurs at the multiple node. The initial data are chosen as eigenelements in an obvious way. There is no longitudinal component in the initial data; therefore, the upright element performs no vertical deflection, which is in complete agreement with the theory.

The final experiment concerns the problem of controlling networks with circuits. It has been shown in [14] that homogenous networks of strings, i.e., the network contains a circuit and all physical constants are equal, are not even approximately controllable. It was further demonstrated in [14] that upon using completely absorbing boundary controls at simple nodes, a residual motion settles in such a circuit that has zero displacements (and velocities) at the nodes constituting the closed path. Furthermore, it has also been remarked in [14] that exact controllability is obtained for string networks, once the prestretching factors (which in turn determine the "optical length" of an individual string) are rational independent. However, rational independence is a matter that is hard to realize (with finitely many machine numbers). We take a square with two strings attached to it. The simple nodes of the string network are controlled in the Neumann conditions. Instead of taking an eigenmode on the square with zero nodal values that would give a picture similar to Figure 7.4, we decided to produce a canonical short range input (bump) as before; see Figure 7.5. The pictures clearly show that only very minor (literally "invisible") oscillations are left after the canonical timespan.

The numerical experiments were done using a Newmark-(1/2,1/4) scheme for the forward and backward wave equations appearing in the optimality conditions.



$t = 0.0$

$t = 0.6$

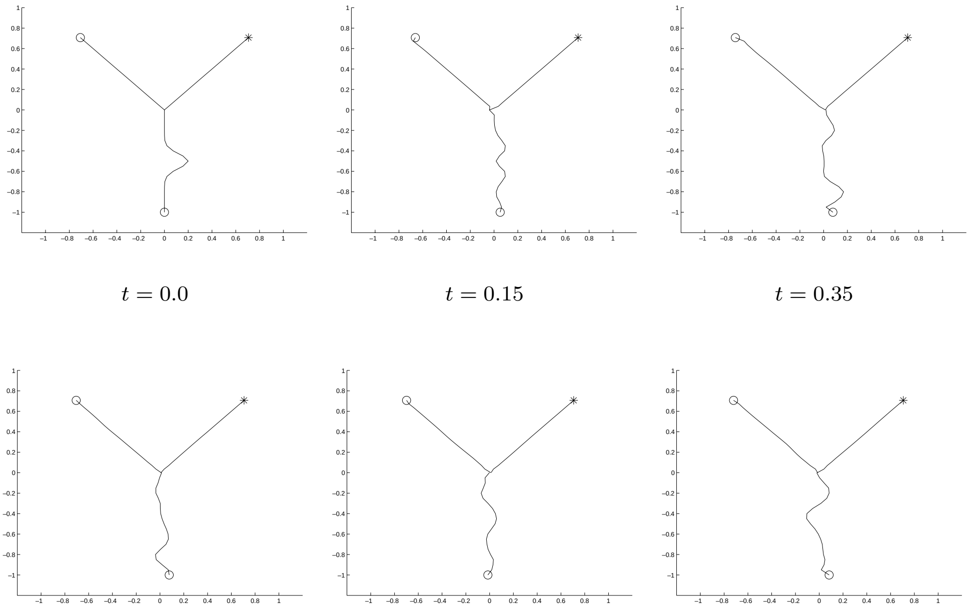
$t = 1.2$

$t = 1.8$

$t = 2.4$

$t = 3.0$

FIG. 7.1. *Control time sufficiently large.*



$t = 0.0$

$t = 0.15$

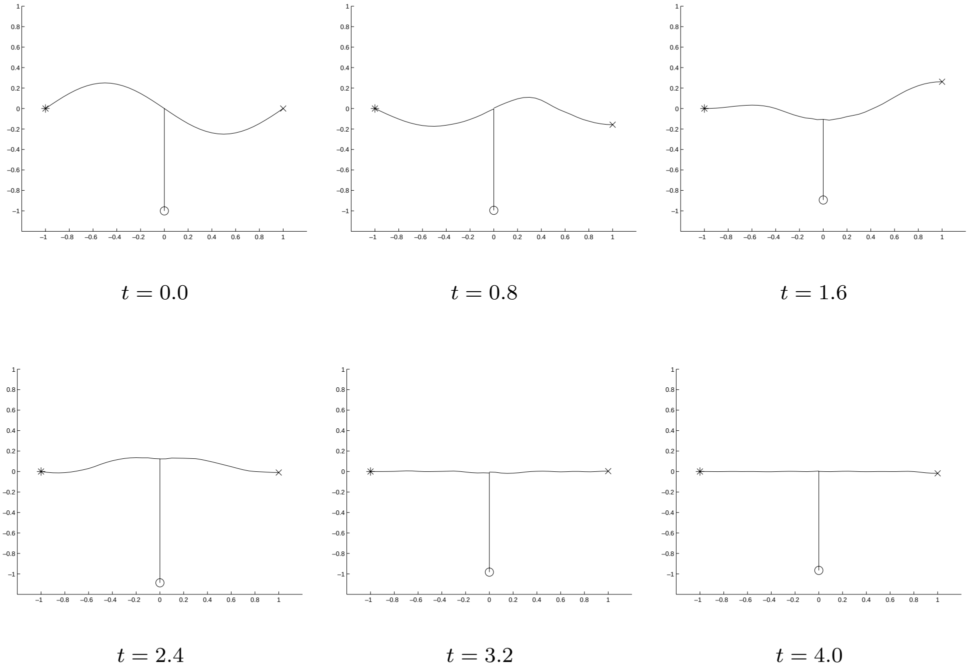
$t = 0.35$

$t = 0.55$

$t = 0.75$

$t = 1.0$

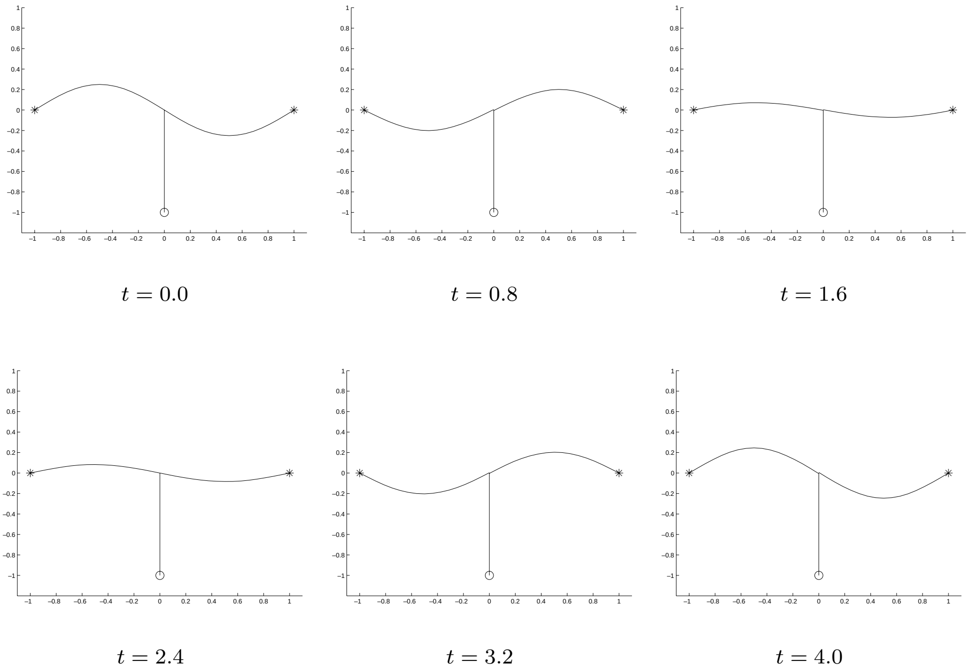
FIG. 7.2. *Control time too small.*



$t = 0.0$ $t = 0.8$ $t = 1.6$

$t = 2.4$ $t = 3.2$ $t = 4.0$

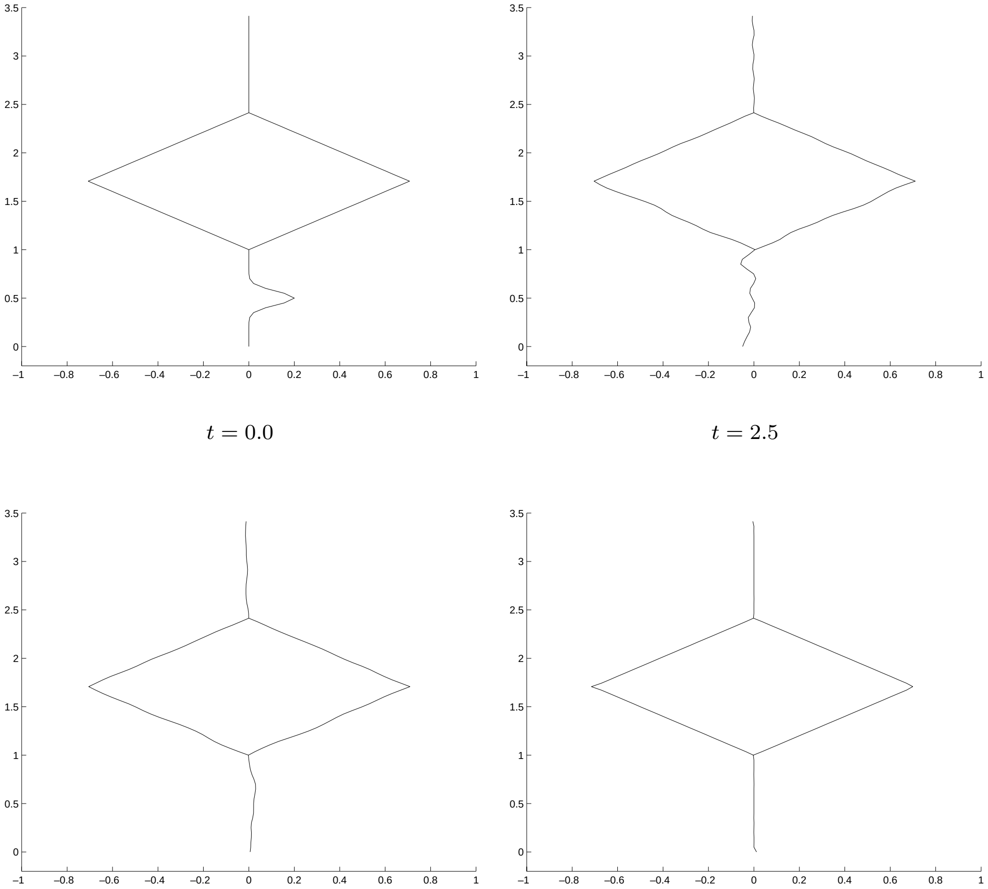
Fig. 7.3. *Control at one simple node only.*



$t = 0.0$ $t = 0.8$ $t = 1.6$

$t = 2.4$ $t = 3.2$ $t = 4.0$

Fig. 7.4. *Lack of controllability due to rational dependence.*



$t = 0.0$

$t = 2.5$

$t = 7.0$

$t = 8.0$

FIG. 7.5. *Optimal control of a simple network containing a circuit.*

We worked out also the discrete optimality conditions for the discretized problem and implemented the corresponding scheme, which is slightly different from the one obtained by first deriving the optimality condition on the partial differential equation level and then discretizing. In fact, this is an important and well-known observation: control and discretization do not commute in general.

The physical constants are always put to unity for easier reference. We used 20 discretization points in space for each member. The outer iterations were terminated usually after 10 iterations to yield 10^{-4} accuracy. In order to be close to controllability phenomena, we have chosen the penalty parameter k to be in the range of 10^4.

We applied several variations of the algorithms and also different discretizations of the equations and costs; however, we do not have the space to dwell on this and refer to a forthcoming publication where the numerical part will be made more explicit.

Acknowledgment. We gratefully acknowledge Ralf Hundhammer's work on the implementation.

REFERENCES

[1] J.-D. Benamou, *Décomposition de domaine pour le contrôle de systèmes gouvernés par des équations d'évolution*, C.R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 1065-1070.

[2] J.-D. Benamou, *A domain decomposition method for the optimal control of systems governed by the Helmholtz equation*, in Proc. Mathematical and Numerical Aspects of Wave Propagation, G. Cohen, ed., SIAM, Philadelphia, 1995, pp. 653–662.

[3] J.-D. Benamou, *A domain decomposition method for control problems*, in Proceedings 9th International Conference on Domain Decomposition Methods, Bergen, Norway, P. Bjørstad et al., eds., 1998, pp. 266–273. http://www.ddm.org/DD9/index.html

[4] J.-D. Benamou, *A domain decomposition method with coupled transmission conditions for the optimal control of systems governed by elliptic partial differential equations*, SIAM J. Numer. Anal., 33 (1996), pp. 2401–2416.

[5] J.-D. Benamou and B. Desprès, *A domain decomposition method for the Helmholtz equation and related optimal control problems*, J. Comput. Phys., 136 (1997), pp. 68–82.

[6] Q. Deng, *An analysis for a nonoverlapping domain decomposition iterative procedure*, SIAM J. Sci. Comput., 18 (1997), pp. 1517–1527.

[7] C. Castro and E. Zuazua, *Boundary controllability of a hybrid system consisting in two flexible beams connected by a point mass*, SIAM J. Control Optim., 36 (1998), pp. 1576–1595.

[8] J. F. Bourgat, R. Glowinski, P. Le Tallec, and M. Vidrascu, *Variational formulation and algorithm for trace operator in domain decomposition calculations*, in Proc. Domain Decomposition Methods, T. Chan, R. Glowinski, J. Periaux, and O. B. Widlund, eds., SIAM, Philadelphia, 1989, pp. 3–16.

[9] G. Chen, M.C. Delfour, A.M. Krall, and G. Payre, *Modeling, stabilization and control of serially connected beams*, SIAM J. Control Optim., 25 (1987), pp. 526–546.

[10] B. Després, *Méthodes de décomposition domaine pour les problèmes de propagation d'ondes en régimes harmoniques*, Ph.D. thesis, Université de Paris IX, INRIA, Roquencourt, 1991.

[11] P. Glowinski and P. Le Tallec, *Augmented Lagrangian interpretation of the nonoverlapping Schwarz alternating method*, in Proc. Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, T. Chan and R. Glowinski, eds., SIAM, Philadelphia, 1990, pp. 224–231.

[12] P. Glowinski and J.L. Lions, *Exact and approximate controllability for distributed parameter systems*, in Acta Numerica 1994, Cambridge University Press, UK, 1994, pp. 269–378.

[13] R. Glowinski and J.L. Lions, *Exact and approximate controllability for distributed parameter systems*, in Acta Numerica 1996, Cambridge University Press, UK, 1996, pp. 159–333.

[14] J.E. Lagnese, G. Leugering, and E.J.P.G. Schmidt, *Modeling, Analysis and Control of Dynamic Elastic Multi-Link Structures*, Birkhäuser, Boston, Basel, Berlin, 1994.

[15] J.E. Lagnese, G. Leugering, and E.J.P.G. Schmidt, *On the analysis and control of hyperbolic systems associated with vibrating networks*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 77–104.

[16] J.E. Lagnese and G. Leugering, *Dynamic domain decomposition in approximate and exact boundary control in problems of transmission for wave equations*, SIAM J. Control Optim., to appear.

[17] G. Leugering, *Reverbation analysis and control of networks of elastic strings*, in Control of Partial Differential Equations and Applications, Lecture Notes in Pure and Appl. Math. 174, Springer-Verlag, Berlin, Heidelberg, New York, 1995, pp. 193–206.

[18] G. Leugering, *On dynamic domain decomposition of controlled networks of elastic strings and joint masses*, in Proc. Control of Distributed Parameter Systems, F. Kappel, ed., Birkhäuser, Boston, Basel, Berlin, 1998, pp. 191–205.

[19] G. Leugering, *A domain decomposition of optimal control problems for dynamic networks of elastic strings*, Comput. Optim. Appl., to appear.

[20] G. Leugering, *A Domain Decomposition Technique for Dynamic Networks of Timoshenko Beams*, manuscript.

[21] J. von Below, *Parabolic network equations*, Habilitation, Tübingen, 1993.

[22] V. Brauer and G. Leugering, *Semi-discretization of control and observation problems for a network of strings*, Control Cybernet., submitted.

[23] G. Leugering, *Dynamic domain decomposition of optimal control problems for networks of Euler-Bernoulli beams*, in ESAIM Proc. Control and Partial Differential Equations, Vol. 4, J.P. Puel and M. Tuscnak, eds., ESAIM, 1998, pp. 223-233.

[24] G. Leugering and E.J.P.G. Schmidt, *On the control of networks of vibrating strings and beams*, in Proc. 28th IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 1989, pp. 2287–2290.

[25] P.L. Lions, *On the Schwarz alternating method* III. *A variant for nonoverlapping subdomains*, in Proc. Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, T. Chan and R. Glowinski, eds., SIAM, Philadelphia, 1990, pp. 202–223.

[26] J.L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications* I, Die Grundlehren der mathematischen Wissenschaften, Springer-Verlag, Berlin, Heidelberg, New York, 1972.

[27] F. Ali Mehmeti, *Nonlinear Waves in Networks*, Akademie-Verlag, Berlin, 1994.

[28] S. Nicaise, *Polygonal Interface Problems*, Methods and Procedures in Mathematical Physics 39, Verlag Peter D. Lang, Frankfurt am Main, 1993.

[29] S. Nicaise, *Control of networks of Euler-Bernoulli beams*, in ESAIM Control Optim. Calc. Var., 4 (1999), pp. 57–81 (electronic).

[30] J. Infante and E. Zuazua, *Boundary observability for the space-discretization of the* 1*-d wave equation*, C. R. Acad. Sci. Paris Ser. I Math., 326 (1998), pp. 713–718.

[31] J.K. Bennighof and R.L. Boucher, *Exact minimum-time control of a distributed system using a traveling wave formulation*, J. Optim. Theory Appl., 73 (1992), pp. 149–167.

[32] F.P. Vasilyev, M.A. Kurzhanskii, and M.M. Potapov, *Method of straight lines in boundary control and observation problems for the equation of string oscillation*, Moscow Univ. Comput. Math. Phys., 3 (1993), pp. 5–11.

[33] S. Opial, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.

[34] J. Saint Jean Paulin and M. Vanninathan, *Vibrations of thin elastic structures and exact controllability*, RAIRO Modél. Math. Anal. Numér., 31 (1997), pp. 765–803.

[35] A. Quarteroni and A. Valli, *Theory and application of Steklov-Poincaré operators for boundary-value problems. The heterogenous operator case*, in Proc. Fourth International Symposium on Domain Decomposition Methods for Partial Differential Equations, T. Chan and R. Glowinski, eds., SIAM, Philadelphia, 1991, pp. 58–81.

[36] E.J.P.G. Schmidt, *On the modelling and exact controllability of networks of strings*, SIAM J. Control Optim., 30 (1992), pp. 229–245.

[37] S.W. Taylor, *Exact boundary controllability of a beam and mass system*, in Proc. Computation and Control IV, K.L. Bowers and J. Lund, eds., Birkhäuser, Basel, 1992, pp. 305–321.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 6, pp. 1676–1700

NECESSARY CONDITIONS OF OPTIMIZATION FOR PARTIALLY OBSERVED CONTROLLED DIFFUSIONS[*]

CHARALAMBOS D. CHARALAMBOUS[†] AND JOSEPH L. HIBEY[‡]

Abstract. Necessary conditions are derived for stochastic partially observed control problems when the control enters the drift coefficient and correlation between signal and observation noise is allowed. The problem is formulated as one of complete information, but instead of considering directly the equation satisfied by the unnormalized conditional density of nonlinear filtering, measure-valued decompositions are used to decompose it into two processes. The minimum principle and the stochastic partial differential equation satisfied by the adjoint process are then derived, and the optimality conditions are shown to be the exact necessary conditions derived by Bensoussan [*Maximum principle and dynamic programming approaches of the optimal control of partially observed diffusions*, Stochastics, 9 (1983), pp. 169–222; *Stochastic Control of Partially Observable Systems*, Cambridge University Press, Cambridge, UK, 1992] when the correlation is zero.

Key words. stochastic control, minimum principle, partially observable diffusions, nonlinear filtering, measure-valued decompositions

AMS subject classification. 93E20

PII. S0363012994259379

1. Introduction.

The stochastic control problem under consideration is

$$\min\{J(u); u \in U_{ad}\}, \tag{1.1}$$

$$J(u) = E^u \left\{ \int_0^T \pi(t, x_t, u_t)dt + \kappa(x_T) \right\}, \tag{1.2}$$

$$dx_t = f(t, x_t, u_t)dt + \sigma(t, x_t)dw_t, \; x_0 = x, \tag{1.3}$$

$$dy_t = h(t, x_t)dt + g(t)dw_t + \hat{g}(t)db_t, \; y_0 = 0, \tag{1.4}$$

where, assuming $\{w_s; 0 \leq s \leq t\}, \{b_s; 0 \leq s \leq t\}$ are uncorrelated Wiener processes, the correlation between the signal process and the observation process is

$$\langle x, y \rangle_t = \int_0^t \sigma(s, x_s)g(s)^T ds. \tag{1.5}$$

The set of all admissible controls, denoted by U_{ad}, may depend on the observations $\{y_s; 0 \leq s \leq t\}$, thereby describing a strict-sense control problem (output feedback) rather than a wide-sense one which requires additional dependence of U_{ad} on the previous controls (see, for example, Fleming and Nisio [3]). Recent approaches addressing this problem with output feedback when the correlation between the signal

[*] Received by the editors December 9, 1994; accepted for publication (in revised form) November 18, 1998; published electronically September 1, 1999.
http://www.siam.org/journals/sicon/37-6/25937.html

[†] Department of Electrical Engineering, McGill University, Montreal, P.Q., Canada H3A 2A7 (chadcha@cim.mcgill.ca).

[‡] Department of Electrical Engineering, University of Colorado at Denver, Denver, CO 80217 (jhibey@carbon.cudenver.edu).

process and the observation process is zero (i.e., $g = 0$) are based on strong and weak variations, and can be found in Bensoussan [1, 2], Haussmann [4], Baras, Elliott, and Kohlmann [5], and Elliott and Yang [6]. A stochastic partial differential equation satisfied by the adjoint process is given by Bensoussan [1, 2] when $g = 0$ and $\sigma(t, x)$ is degenerate, whereas the other authors above provide an explicit representation for the adjoint process in terms of a conditional expectation. The derivations found in Bensoussan [1] and Haussmann [4] are based on the robust version of the unnormalized conditional density equation which does not exist when the correlation considered here is present (due to h_i, h_j being noncommutative). Bensoussan [2] introduced a new approach in solving the problem by applying the variational methods of partial differential equations, weak control variations, and using Galerkin's approximations for Sobolev spaces to approximate the adjoint process through a finite-dimensional basis.

Here, for the stochastic control problem (1.1)–(1.4), we shall derive optimality conditions that are generalizations of the optimality conditions derived in Bensoussan [1, 2]. The main idea of our methodology is the use of measure-valued decompositions as discussed by Kunita [7, 8], whereby the unnormalized conditional density ρ_t is written as a composition of two measure-valued processes such that $\rho_t = \nu_t\mu_t$. In Kunita [7, 8], this decomposition is used to obtain necessary and sufficient conditions for the smoothness of ρ_t. Here we recognize that ν_t satisfies a parabolic partial differential equation containing the Kolmogorov operator, and hence the explicit dependence on the control, while μ_t satisfies a stochastic partial differential equation containing the observations and, therefore, has no explicit dependence on the control. Therefore, in the control situation, this decomposition serves as a separation argument similar to the separation principle given by Wonham [9] because any control variation affects explicitly only the measure-valued process ν_t. A similar approach in deriving necessary conditions of optimization for state-valued processes satisfying the separation principle of Wonham [9] is given by Bensoussan [10].

In section 2, we state our main assumptions and give a summary of the results from filtering theory and Kunita's [7, 8] decomposition which will be used throughout. In section 3, we formulate the stochastic control problem by expressing the performance index in terms of the measure-valued processes ν_t, μ_t. In addition, we present our separation argument and relate the perturbed process $\rho_t^{\mathcal{B}}$ of Bensoussan [1, 2] to our perturbed processes z_t which are, respectively, the byproducts of applying weak variations to the equations satisfied by ρ_t, ν_t. In section 4, we derive the minimum principle using the interplay between Euclidean variational methods and variational methods for measure-valued processes. We also derive the stochastic partial differential equation satisfied by the adjoint process using a generalization of the martingale representation theorem. In section 5, we show that the minimum principle established by Bensoussan [1, 2] is a special case of Theorem 4.3, and that the stochastic partial differential equation satisfied by the adjoint process found in Bensoussan [1, equation (2.66)] is related to the adjoint process equation given by (4.10) (in fact, Bensoussan's adjoint process equation is again a special case of the adjoint process given by (4.10)).

2. Problem formulation.

The control parameter u will take values in a compact, convex subset U of some Euclidean space R^k. We shall make the following assumptions.

2.1. Assumptions.

(A1) $f : [0, T] \times R^n \times U \to R^n$, bounded continuous in t, continuous in u, C^∞ in

x, with bounded derivatives of any order in x, bounded first derivatives in u, with a constant k_1 such that

$$|f(t,x,u) - f(t,z,u)| \leq k_1|x-z|;$$

(A2) $\sigma : [0,T] \times R^n \to R^n \otimes R^m$, bounded continuous in t, C^∞ in x, with bounded derivatives of any order in x, and a constant k_2 such that

$$||\sigma(t,x) - \sigma(t,z)|| \leq k_2|x-z|;$$

(A3) $h : [0,T] \times R^n \to R^d$, bounded continuous in t, C^∞ in x, with bounded derivatives of any order in x;

(A4) $g : [0,T] \to R^d \otimes R^m$, $\hat{g} : [0,T] \to R^d \otimes R^d$, continuous, with constants $\beta_1, \beta_2 > 0$, such that

$$g(t)g(t)^T + \hat{g}(t)\hat{g}(t)^T \geq \beta_1 I_d, \; \hat{g}(t)\hat{g}(t)^T \geq \beta_2 I_d,$$

where I_d denotes the $d \times d$ identity matrix;

(A5) $\pi : [0,T] \times R^n \times U \to R$, bounded, with bounded L^2-norm independent of u, continuous in u, and satisfying

$$\pi(.,.,u) \in L^2([0,T] \times R^n), \; \left|\frac{\partial}{\partial u}\pi(t,x,u)\right| \leq \gamma(x), \; \gamma(\cdot) \in L^2(R^n);$$

(A6) $\kappa : R^n \to R$, Borel bounded with $\kappa(\cdot) \in L^2(R^n)$;

(A7) $x_0 = x$ is assumed to be known. If x_0 is random, then $p_0 \in L^2(R^n)$ is the density of x_0 and is assumed to be independent of $\{w_s, b_s; 0 \leq s \leq t\}$.

2.2. Evolution of unnormalized conditional density equation. We start with a reference probability space $(\Omega, \mathcal{F}, \mathcal{P})$ with a complete filtration $\{\mathcal{F}_t; t \in [0,T]\}$, two adapted Wiener processes, $\{w(t); t \in [0,T]\}$, $\{b(t); t \in [0,T]\}$, and an $\mathcal{F}_0$ measurable random variable x_0 such that

$w : [0,T] \times \Omega \to R^n$ is a standard Wiener process independent of $b(\cdot)$,
$b : [0,T] \times \Omega \to R^d$ is a standard Wiener process independent of $w(\cdot)$,
$x(0) : \Omega \to R^n$ is a random variable independent of $w(\cdot), b(\cdot)$.

Furthermore, suppose an observation process $y(\cdot)$ is given by

$$dy_t = g(t)dw_t + \hat{g}(t)db_t, \; y_0 = 0. \tag{2.1}$$

We write $\{\mathcal{F}_t^y; t \in [0,T]\}$ for the complete filtration generated by the observation σ-algebra $\sigma\{y_s; 0 \leq s \leq t \leq T\}$, and we denote by E^u (resp., E) the expectation with respect to measure $\mathcal{P}^u$ (resp., $\mathcal{P}$).

Definition 2.1. *Denote by $L_y^2([0,T]; R^k)$ the set of square integrable stochastic processes adapted to $\{\mathcal{F}_t^y; t \in [0,T]\}$ with values in R^k. The set of admissible controls, denoted by U_{ad}, is defined as*

$$U_{ad} \stackrel{\triangle}{=} \{u(\cdot) \in L_y^2([0,T]; R^k) u(t,y) \in U, (a.e. \text{ on } [0,T]), \mathcal{P} \; a.s.\}.$$

Given the system $(\Omega, \mathcal{F}, \mathcal{P}; \mathcal{F}_t)$ and a $u \in U_{ad}$, consider the diffusion process $x(\cdot)$ satisfying the Ito equation

$$dx_t = f(t,x_t,u_t)dt - \sigma(t,x_t)g(t)^T k(t,x_t)dt + \sigma(t,x_t)dw_t, \; x_0 \in R^n, \tag{2.2}$$

where the $d \times 1$ vector $k(t, x_t)$ is defined by

$$k(t, x_t) \triangleq (g(t)g(t)^T + \hat{g}(t)\hat{g}(t)^T)^{-1} h(t, x_t). \tag{2.3}$$

By assumptions (A3) and (A4), $k(t, x)$ is a well-defined bounded vector. Furthermore, by assumptions (A1)–(A3) and (A7), there exists a unique solution (see Bensoussan [2]) such that

$$x(.) \in L^2(\Omega, \mathcal{F}_T, \mathcal{P}; C([0, T]; R^n)).$$

Define the martingale $\{m_t; t \in [0, T]\}$ with respect to $(\{\mathcal{F}_t; t \in [0, T]\}, \mathcal{P})$ by

$$m_T \triangleq \int_0^T k(s, x_s)^T (g(s)dw_s + \hat{g}(s)db_s),$$

and introduce a new probability measure through the likelihood-ratio Λ_T^u defined as

$$\Lambda_T^u \triangleq E\left[\frac{d\mathcal{P}^u}{d\mathcal{P}} | \mathcal{F}_T\right] = \exp\left\{m_T - \frac{1}{2}\langle m, m\rangle_T\right\}. \tag{2.4}$$

On this new probability measure $\mathcal{P}^u$ (and after incorporating the martingale translation theorem) the processes $\{x_s, y_s; 0 \le s \le t\}$ become for the system $(\Omega, \mathcal{F}_t, \mathcal{P}^u), t \in [0, T]$, the weak solutions of (1.3), (1.4), which are unique in probability law.

Given any $\tilde{f} \in C_b^\infty(R^n)$ (space of continuous, real-valued, infinitely continuously differentiable functions with bounded derivatives), we obtain from Bayes formula (see Bensoussan [2] and Kunita [11], assuming existence of a conditional distribution which is also absolutely continuous with respect to Lebesque measure)

(2.5)

$$E^u[\tilde{f}(x_t)|\mathcal{F}_t^y] = \frac{E[\tilde{f}(x_t)\Lambda_t^u|\mathcal{F}_t^y]}{E[\Lambda_t^u|\mathcal{F}_t^y]} = \frac{\int_{R^n} \tilde{f}(z)\rho_t(dz)}{\int_{R^n} \rho_t(dz)} = \frac{\rho_t(\tilde{f})}{\rho_t(1)} = \frac{\langle \tilde{f}(z), \rho_t(z)\rangle}{\langle 1, \rho_t(z)\rangle},$$

where ρ_t is the unnormalized conditional density of x_t given the information $\mathcal{F}_t^y$.

THEOREM 2.2. *For any $\tilde{f} \in C_b^\infty(R^n)$, $\rho_t(\tilde{f})$ is the solution of*

$$\rho_t(\tilde{f}) = p_0(\tilde{f}) + \int_0^t \rho_s(A^u(s)\tilde{f})ds + \int_0^t \rho_s(M(s)\tilde{f})d\tilde{y}_s, \tag{2.6}$$

where p_0 is a delta measure δ_x and

$$A^u(t) = \sum_{i=1}^n f^i(t, x, u)\frac{\partial}{\partial x^i} + \frac{1}{2}\sum_{i,j=1}^n a^{ij}(t, x)\frac{\partial^2}{\partial x^i \partial x^j}, \ a(t, x) = \sigma(t, x)\sigma(t, x)^T, \tag{2.7}$$

$$M_k(t) = \sum_{i=1}^d (\tilde{k}(t)^{-1/2})^{ik} h^i(t, x) + \sum_{i=1}^n (\sigma(t, x)g(t)^T \tilde{k}(t)^{-1/2})^{ik}\frac{\partial}{\partial x^i}, \ 1 \le k \le d, \tag{2.8}$$

$$\tilde{k}(t) = g(t)g(t)^T + \hat{g}(t)\hat{g}(t)^T, \tag{2.9}$$

with $(\cdot)^{i,j}$ the (i, j)th component of a matrix, $(\cdot)^i$ the ith component of a vector, and $\{\tilde{y}_s; 0 \le s \le t\}$ a standard Wiener process.

Proof. See Pardoux [12]. □

Remark 2.1. Notice that (2.6) is a degenerate version of the equation presented by Pardoux [12, equation (4.3), pp. 206–216] in the case when no control is present, $g(t) \in R^d \otimes R^n, \sigma(t,x) \in R^n \otimes R^n$, and $\tilde{k}(t) = I_d$. A similar equation is derived by Kunita [8, pp. 154–167] using the interplay between Kushner's equation for the normalized conditional density and the unnormalized conditional density when no control is present, $\hat{g} = I_d, g = 0$, and a martingale $\tilde{g}(t)db_t$ is added to the right side of (1.3). Different proofs of (2.6), when $\sigma(t,x) \in R^n \otimes R^n$ and $g(t) \in R^d \otimes R^n$, are also found in Bensoussan [2]. It is important to note that the case when g depends on x still remains open. For reasons of simplicity, and without loss of generality, we shall assume that $\tilde{k}(t) = I_d$ so that $\{y_s; 0 \leq s \leq t\}$ is the standard Wiener process described by (2.1).

2.3. Decomposition and representations of unnormalized conditional density. Suppose the operators $A^u(t), M_k(t)$ are defined by

$$A^u(t) \triangleq \frac{1}{2}\sum_{j=1}^{m} X_j(t)^2 + X_0^u(t);\ M_k(t) \triangleq h_k(t) + Y_k(t),\ 1 \leq k \leq d, \tag{2.10}$$

where

$$X_0^u(t) \triangleq X_0(t,x,u) = \sum_{i=1}^{n}\left(f^i(t,x,u) - \frac{1}{2}\sum_{k=1}^{m}\sum_{j=1}^{n}\sigma^{jk}(t,x)\frac{\partial}{\partial x^j}\sigma^{ik}(t,x)\right)\frac{\partial}{\partial x^i},$$

$$X_j(t) \triangleq X_j(t,x) = \sum_{i=1}^{n}\sigma^{ij}(t,x)\frac{\partial}{\partial x^i},\ 1 \leq j \leq m,$$

$$Y_k(t) \triangleq Y_k(t,x) = \sum_{j=1}^{m}\gamma(t)^{kj}X_j(t),\ h_k(t) = h^k(t,x),\ 1 \leq k \leq d,$$

and γ^{ij} are components of $g(t)$. Writing the stochastic integral of (2.6) in terms of the Stratonovich integral and defining

$$L^u(t) \triangleq A^u(t) - \frac{1}{2}\sum_{k=1}^{d} M_k(t)^2 = \frac{1}{2}\sum_{j=1}^{m}\tilde{X}_j(t)^2 + \tilde{X}_0^u(t) + \tilde{h}_0(t), \tag{2.11}$$

then (2.6) is equivalent to

$$\rho_t(\tilde{f}) = p_0(\tilde{f}) + \int_0^t \rho_r(L^u(r)\tilde{f})dr + \sum_{k=1}^{d}\int_0^t \rho_r(M_k(r)\tilde{f}) \circ dy_r^k. \tag{2.12}$$

Here, "∘" denotes the Stratonovich integral. Furthermore,

$$\tilde{X}_0^u(t) \triangleq \tilde{X}^u(t,x) = X_0^u(t) - \sum_{k=1}^{d} h_k(t)Y_k(t),\ \tilde{X}_j(t) \triangleq \tilde{X}_j(t,x) = \sum_{k=1}^{m}\theta^{ik}(t)X_k(t),\ 1 \leq j \leq m, \tag{2.13}$$

$$M_k(t) = h_k(t) + Y_k(t),\ \tilde{h}_0(t) = -\frac{1}{2}\sum_{k=1}^{d}(h_k(t)^2 + Y_k(t)h_k(t)), \tag{2.14}$$

and $\theta^{ij}(t)$ are the components of the positive definite matrix $\theta(t)$, satisfying $\theta(t)^T\theta(t) = I_m(t) - g(t)^T g(t)$.

At this point, we invoke the representation results of stochastic partial differential equations given by Kunita [8, 7]. Write the space $(\Omega, \mathcal{F}_T, \mathcal{P})$ as a product of two probability spaces as follows:

$$\Omega = \Omega^y \otimes \Omega^w = C([0,T]; R^d) \otimes C([0,T]; R^m),$$

$$\mathcal{F}_T = \mathcal{F}_T^y \otimes \mathcal{F}_T^w,\ \mathcal{P}(dy, dw) = \mathcal{P}^y(dy) \otimes \mathcal{P}^w(dw),$$

where $\mathcal{F}_T^y, \mathcal{F}_T^w$ are the Borel σ-algebras on $C([0,T]; R^d), C([0,T]; R^m)$ (spaces of continuous functions), respectively. Here $\mathcal{P}^y, \mathcal{P}^w$ denote the Wiener measures $\mu_w^d(dy), \mu_w^m(dw)$ on $C([0,T]; R^d), C([0,T]; R^m)$, respectively. For $u \in U_{ad}$ consider the process $\xi_{0,t} = \xi_{0,t}(x, y, w)$ starting at $\xi_{0,0} = x$ and described by the stochastic differential equation

$$(\Omega^y \otimes \Omega^w, \mathcal{F}_t^y \otimes \mathcal{F}_t^w, \mathcal{P}^y \otimes \mathcal{P}^w):$$

$$d\xi_{0,t} = \tilde{X}^u(t, \xi_{0,t})dt + \sum_{j=1}^{m}\tilde{X}_j(t, \xi_{0,t}) \circ dw_t^j + \sum_{k=1}^{d} Y_k(t, \xi_{0,t}) \circ dy_t^k. \tag{2.15}$$

Since the coefficients of (2.15) are of C^∞-class and their derivatives are bounded, by Kunita [13, Theorem 2.3], the solution map $\xi_{0,t}(\cdot, y, w) : R^n \to R^n$ is a.s., for each $y \in C([0,T]; R^d)$ (which specifies the control function at a fixed time t), a C^∞-diffeomorphism in x for any t. Define $\phi_{0,t} \overset{\triangle}{=} \phi_{0,t}(x, y, w)$ by

$$\phi_{0,t} \overset{\triangle}{=} \exp\left\{\sum_{k=1}^{d}\int_0^t h_k(r, \xi_{0,r}(x)) \circ dy_r^k + \int_0^t \tilde{h}_0(r, \xi_{0,r})(x)dr\right\}. \tag{2.16}$$

Then (see Kunita [8, 7]) the solution of (2.12) can be represented by

$$\rho_t(\tilde{f})(x, y) = E_{\mathcal{P}^w}[\tilde{f}(\xi_{0,t}(x, y, \cdot))\phi_{0,t}(x, y, \cdot)]. \tag{2.17}$$

Also, since $u(t, .)$ is measurable with respect to t and $\{y_s; 0 \leq s \leq t\}$, the results of Kunita [14] extend easily and show that $\rho_t(\tilde{f})$ of (2.12) satisfies $E_{\mathcal{P}}|\rho_t(\tilde{f})(x)|^2 < \infty$ for all t and x.

Remark 2.2. Notice that if the initial state $x = x_0$ is random, then (2.19) is given by

$$\rho_t(\tilde{f})(\tilde{w}) = \int_{R^n} E_{\mathcal{P}^w}[\tilde{f}(\xi_{0,t}(x, \tilde{w}, .))\phi_{0,t}(x, \tilde{w}, .)]p_0(dx);$$

thus an additional integration with respect to p_0 is required, which is also consistent with the derivation of (2.6), where $x_0 = x$ is assumed to be deterministic; see Kunita [8, 11].

A different representation describing the solution of (2.12), which also constitutes our main tool in deriving the stochastic minimum principle, is given in Kunita [8] by treating the second term on the right side of (2.12) as the principal part and the last term as the perturbation part. Using this approach, which also holds in the controlled case as well, the solution to (2.12) can be expressed as $\rho_t = \nu_t \mu_t$, where

$$d\mu_t(\tilde{f})(x,y) = \sum_{k=1}^{d} \mu_t(M_k(t)\tilde{f})(x,y) \circ dy_t^k, \ \lim_{t \downarrow 0} \mu_t(\tilde{f}) = \tilde{f}(x), \tag{2.18}$$

$$\frac{\partial}{\partial t}\nu_t(\tilde{f})(x,y) = \nu_t(\mu_t L^u(t)\mu_t^{-1}\tilde{f})(x,y), \ \lim_{t \downarrow 0} \nu_t(\tilde{f}) = \tilde{f}(x), \tag{2.19}$$

with $\mu_t L^u(t)\mu_t^{-1}$ a second-order differential operator of parabolic type (see Kunita [8], [11, Chapter 6]). In fact, if we let $n_{0,t}(x,y)$ correspond to the solution of

$$dn_{0,t} = \sum_{k=1}^{d} Y_k(t, n_{0,t}) \circ dy_t^k, \ n_{0,0} = x, \tag{2.20}$$

which is also the solution to (2.15) when $\tilde{X}_j(t) = 0, \ 0 \leq j \leq m$, then the solution to (2.18) can be represented as

$$\mu_t(\tilde{f})(x,y) = \tilde{f}(n_{0,t}(x,y)) \exp\left\{ \sum_{k=1}^{d} \int_0^t h_k(r, n_{0,r}(x,y)) \circ dy_r^k \right\}. \tag{2.21}$$

By the assumptions on $Y_1, \ldots, Y_d$ the solution map $n_{0,t}(.,y) : R^n \to R^d$ is, a.s., one-to-one, onto, and C^∞ in x for each t. The representation (2.21) can be shown easily by an application of Ito's formula. Similarly, a representation for $\nu_t(\tilde{f})(x,y)$ is given in Kunita [7, 13].

Remark 2.3. For the uncontrolled case, and assuming $L(t)$ is hypoelliptic, it was shown by Kunita [7] that the solution measure $\rho_t(x,y,dz)$ of (2.6) has C^∞-density function for almost all $y \in C([0,T];R^d)$ by first showing that the solution measure $\nu_t(x,y,dz)$ of (2.19) has a C^∞-density function (because $\rho_t = \nu_t\mu_t$ and $\mu_t(.,y) \in C_b^\infty(R^n)$ is a one-to-one and onto map). Also, since for each t, $u(t,.)$ depends on $\{y_s; 0 \leq s \leq t\}$, the result of Kunita [7] can be applied to our problem by assuming that the Lie algebra generated by the vector fields of the diffusion coefficient of (1.3) is, a.s., of dimension n for all $t \in [0,T], x \in R^d$.

Remark 2.4. Notice that the measure-valued process μ_t satisfying (2.18) does not explicitly depend on the control u, as is easily seen in (2.14); only the measure-valued process ν_t satisfying (2.19) depends on u. This is the separation argument we shall concentrate on in our analysis.

2.4. Variational methods applied to the information state. Introduce the Hilbert space $H(R^n) = L^2(R^n)$ with scalar product denoted by $(\cdot,\cdot)$. Define the Sobolev space

$$H^1(R^n) \triangleq \left\{ \hat{u} \in L^2(R^n), \frac{\partial \hat{u}}{\partial x^i} \in L^2(R^n), 1 \leq i \leq n \right\}.$$

Let $H^{-1}(R^n)$ be the dual of $H^1(R^n)$. The norm in $H(R^n)$ is denoted by $|\cdot|$ while the norm in $H^1(R^n)$ is denoted by $||\cdot||$, that is,

$$|| \hat{u} ||^2 = | \hat{u} |^2 + \sum_{i=1}^{n} \left| \frac{\partial}{\partial x^i} \hat{u} \right|^2 .$$

The norm in $H^{-1}(R^n)$ is denoted by $||\cdot||_*$. The pairing between $H^1(R^n)$ and $H^{-1}(R^n)$ is denoted by $\langle \cdot, \cdot \rangle$. Define

$$\begin{aligned} &L_y^2((0,T);H^1) \overset{\triangle}{=} \\ &\{\hat{u} \in L^2(\Omega^y, \mathcal{F}_t^y, \mathcal{P}^y; L^2((0,T);H^1)); \text{a.e. on } [0,\mathrm{T}], \hat{u}(t) \in L^2(\Omega^y, \mathcal{F}_t^y, \mathcal{P}^y; H^1)\}. \end{aligned}$$

Integrating (2.19) by parts, we get the equation

$$\frac{\partial}{\partial t}\nu_t = L_{w^y}^u(t)^* \nu_t, \;\; \nu_0 = \delta_x \tag{2.22}$$

written in strong form, where $L_{w^y}^u(t)^*$ is the formal adjoint of $L_{w^y}^u(t) = \mu_t L^u(t) \mu_t^{-1}$, which can be represented by

$$L_{w^y}^u(t) = \frac{1}{2}\sum_{i,j=1}^n \alpha_{w^y}^{ij}(t,x)\frac{\partial^2}{\partial x^i \partial x^j} + \sum_i^n b_{w^y}^i(t,x,u)\frac{\partial}{\partial x^i} + d_{w^y}(t,x,u).$$

Lemma 2.3. *Suppose* $u \in U_{ad}$, (A1)–(A7) *hold, and* (A8) $w_t \in R^n, \exists \beta_3 > 0$ *such that*

$$\sigma(t,x)\sigma^T(t,x) \geq \beta_3 I_n > 0 \; \forall t \geq 0, \forall x \in R^n.$$

Then

$$-A^u(\cdot), A^u(\cdot)^* \in L^\infty((0,T); \mathcal{L}(H^1(R^n); H^{-1}(R^n))),$$

$$M(\cdot), M(\cdot)^* \in L^\infty((0,T); \mathcal{L}(H^1(R^n); H^d(R^n))),$$

the coercivity condition associated with the strong form of (2.6) *holds, that is, for some* $\lambda_1, \lambda_2 > 0$,

$$-2\langle A^u(t), \hat{u}, \hat{u}\rangle + \lambda_1 | \hat{u} |^2 \geq \lambda_2 || \hat{u} ||^2 + \sum_{k=1}^d |M_k(t)^* \hat{u} |^2$$

and for each $y(\cdot) \in C([0,T]; R^d)$, *there exists a unique solution of* (2.22) *in the space*

$$\nu(\cdot) \in L_y^2((0,T); H^1(R^n)), \frac{\partial}{\partial t}\nu(\cdot) \in L_y^2((0,T); H^{-1}(R^n)).$$

Proof. The coercivity condition is a direct consequence of (A1)–(A4) and (A8) (see Bensoussan [2]). For each $u \in U_{ad}$, define

$$A_{w^y}^u(t) f \overset{\triangle}{=} \mu_t(A^u(t)\mu_t^{-1}) f = \mu_t\left(\left[L^u(t) + \frac{1}{2}\sum_{k=1}^d M_k(t)^2\right]\mu_t^{-1}\right) f.$$

The operators $L_{w^y}^u(t)$ and $A_{w^y}^u$ are for almost all $w^y \in \Omega^y$ second-order linear parabolic partial differential operators. By assumptions (A1)–(A4) and (A8), the random functions $\alpha_{w^y}^{i,j}, b_{w^y}^i, d_{w^y}$ are a.s. smooth and bounded. Therefore, for any $w^y \in C([0,T]; R^d)$, (2.22) can be viewed as a deterministic partial differential equation. Using the standard theory of these equations, there is one and only one solution of (2.22) as specified. □

Next, consider the strong form of (2.6):

$$\rho_t = p_0 + \int_0^t A^u(s)\rho_s ds + \sum_{i=1}^d \int_0^t M_k(s)\rho_s dy_s^i. \tag{2.23}$$

The previous lemma enables us to obtain the following results.

Theorem 2.4. *Suppose* $u \in U_{ad}$ *and* (A1)–(A8) *hold. Then there exists one and only one solution to* (2.23) *in the space*

$$\rho(\cdot) \in L_y^2((0,T); H^1(R^n)) \cap L^2(\Omega^y, \mathcal{F}_t^y, \mathcal{P}^y; C((0,T); H)).$$

Proof. This follows from Lemma 2.3 and the representation $\rho_t = \nu_t \mu_t$. Alternatively, we can employ the energy equality

$$|\rho_t|^2 = |p_0|^2 + 2\int_0^t \langle A^u(r)\rho_r, \rho_r\rangle dr + \sum_{k=1}^d \int_0^t |M_k(r)\rho_r|^2 dr + 2\sum_{k=1}^d \int_0^t (M_k(r)\rho_r, \rho_r) dy_r^k$$

and the coercivity condition of Lemma 2.3, and then proceed as in Bensoussan [2]. □

3. Decomposed control problem. With the aid of (2.4), the expected cost (1.2) can be written as

$$J(u) = E\left\{\Lambda_T \left[\kappa(x_T) + \int_0^T \pi(r, x_r, u_r) dr\right]\right\}.$$

In addition, by (2.5), the above cost is equivalent to

$$J(u) = E_{\mathcal{P}}\left\{\langle \kappa, \rho_T\rangle + \int_0^T \langle \pi(r, ., u_r), \rho_r\rangle dr\right\}. \tag{3.1}$$

Since $\rho_t = \nu_t \mu_t$, we also have the representation

$$J(u) = E_{\mathcal{P}}\left\{\nu_T \mu_T(\kappa) + \int_0^T \nu_r \mu_r(\pi^u(r)) dr\right\}, \tag{3.2}$$

where $\pi^u(t) \overset{\triangle}{=} \pi(t, x, u)$.

Suppose $u^* \in U_{ad}$ is an optimal control. For any other control $u \in U_{ad}$ and for $\epsilon \in [0,1]$, we know that

$$u_t^\epsilon = u_t^* + \epsilon(u_t - u_t^*) \in U$$

and

$$J(u^\epsilon) \geq J(u^*).$$

In addition, if the Gâteaux derivative of $J(u)$ as a functional on the Hilbert space $L_y^2((0,T); R^k)$ is well defined, then by differentiating with respect to ϵ we obtain

$$\frac{d}{d\epsilon} J(u^\epsilon)|_{\epsilon=0} \geq 0 \ \forall u \in U_{ad}.$$

In the following, we assume for convenience that $u \in U_{ad}$ is such that

$$u^\delta = u^* + \delta u \in U_{ad}, \ \ \delta \in [0, \alpha].$$

As a result of the decomposition (2.18) and (2.19), any control variation $u^\delta \in U_{ad}$ would affect only the measure-valued process ν_t. Indeed, it is clear that the right side of (2.18) does not contain the control variable because $M_k(t), 1 \leq k \leq d$, as defined by (2.14), is independent of u. Thus the solution measure μ_t of (2.18) is not affected by control variations (see Kunita [13, 15] and Bensoussan [2]).

Let us now introduce the equation

$$\rho_t^{\mathcal{B}}(\tilde{f}) = \int_0^t \rho_r \left(\frac{\partial}{\partial u} L^{u^*}(r) u_r \tilde{f} \right) dr + \int_0^t \rho_r^{\mathcal{B}}(L^{u^*}(r)\tilde{f}) dr + \sum_{k=1}^d \int_0^t \rho_r^{\mathcal{B}}(M_k(r)\tilde{f}) \circ dy_r^k, \tag{3.3}$$

which is obtained by formally considering weak control variations of (2.12). Next, we shall show that the solution $\rho_t^{\mathcal{B}}$ of (3.3) can also be represented by the composition $\rho_t^{\mathcal{B}} = z_t \mu_t$, where z is given by (3.4).

LEMMA 3.1. *Let $\mu_t^{-1}(\cdot)(x, y)$ be the inverse operator of $\mu_t(\cdot)(x, y)$. Then $\rho_t^{\mathcal{B}}$ is the solution to* (3.3) *if and only if*

$$z_t(\tilde{f}) = \rho_t^{\mathcal{B}}(\mu_t^{-1}(\tilde{f}))$$

is the solution of

$$\frac{\partial}{\partial t} z_t(\tilde{f}) = z_t(\mu_t L^{u^*}(t) \mu_t^{-1} \tilde{f}) + \nu_t \left(\mu_t \frac{\partial}{\partial u} L^{u^*}(t) u_t \mu_t^{-1} \tilde{f} \right), \quad \lim_{t \downarrow 0} z_t = 0. \tag{3.4}$$

Proof. Suppose $\rho_t^{\mathcal{B}}$ is the solution to (3.3). We shall show that $z_t(\tilde{f}) = \rho_t^{\mathcal{B}}(\mu_t^{-1}(\tilde{f}))$ is a solution to (3.4). By (2.21), the inverse operator $\mu_{s,t}^{-1}$ is given by

$$\mu_{s,t}^{-1}(\tilde{f}) = \tilde{f}(n_{s,t}^{-1}(x)) \exp \left\{ - \sum_{k=1}^d \int_s^t h_k(r, n_{r,t}^{-1}(x)) \circ \hat{d}\, y_r^k \right\}, \tag{3.5}$$

where $n_{s,t}^{-1}$ (which is one-to-one and onto, and $C^\infty(R^n)$ a.s. for each t) is the inverse process of $n_{s,t}$ which is the solution of (2.20) starting at $n_{s,s} = x$. Letting $\tilde{n}_{s,t}, s \in [0, t]$, be the solution of the backward stochastic differential equation

$$d\tilde{n}_{s,t} = \sum_{k=1}^d Y_k(s, \tilde{n}_{s,t}) \circ \hat{d}\, y_s^k, \ \ \tilde{n}_{t,t} = x,$$

from Kunita [13, 15] we have $n_{s,t}^{-1} = \tilde{n}_{s,t}$. Applying the Ito differential rule to $\tilde{f}(\tilde{n}_{s,t}(x))\phi_{s,t}^{-1}(x)$, which is the right side of (3.5), it follows that

$$\begin{aligned} \tilde{f}(\tilde{n}_{s,t}(x))\phi_{s,t}^{-1}(x) = \tilde{f}(x) &- \sum_{k=1}^d \int_s^t Y_k(r) \tilde{f}(\tilde{n}_{r,t}(x)) \phi_{r,t}^{-1}(x) \circ \hat{d} y_r^k \\ &- \sum_{k=1}^d \int_s^t h_k(r) \tilde{f}(\tilde{n}_{r,t}(x)) \phi_{r,t}^{-1}(x) \circ \hat{d} y_r^k. \end{aligned}$$

Interchanging the forward and backward variables of integration (see Kunita [8, Part II]) and setting $s = 0, \tilde{n}_t = \tilde{n}_{0,t}$ we arrive at

$$\mu_t^{-1}(\tilde{f}) = \tilde{f}(x) - \sum_{k=1}^{d} \int_0^t M_k(r)\mu_r^{-1}(\tilde{f}) \circ dy_r^k. \tag{3.6}$$

Applying the result of Kunita [11, Lemma 6.2.7, Theorem 6.2.8, pp. 312–313] to $\tilde{z}_t(\tilde{f}) = \rho_t^{\mathcal{B}}(\mu_t^{-1}(\tilde{f}))$, we have

$$\begin{aligned} d\tilde{z}_t(\tilde{f}) &= d[\rho_t^{\mathcal{B}}(\mu_t^{-1}(\tilde{f}))] \\ &= \rho_t\left(\frac{\partial}{\partial u} L^{u^*}(t) u_t \mu_t^{-1}\tilde{f}\right) dt + \rho_t^{\mathcal{B}}(L^{u^*}(t)\mu_t^{-1}\tilde{f})dt \\ &\quad + \sum_{k=1}^{d} \rho_t^{\mathcal{B}}(M_k(t)\mu_t^{-1}\tilde{f}) \circ dy_t^k - \sum_{k=1}^{d} \rho_t^{\mathcal{B}}(M_k(t)\mu_t^{-1}\tilde{f}) \circ dy_t^k \\ &= \nu_t\mu_t\left(\frac{\partial}{\partial u} L^{u^*}(t) u_t \mu_t^{-1}\tilde{f}\right) dt + \tilde{z}_t(\mu_t L^{u^*}(t)\mu_t^{-1}\tilde{f})dt, \end{aligned}$$

where the second equality follows by substituting $\nu_t\mu_t$ for ρ_t and using the fact that μ_t is a one-to-one and onto map. The sufficiency can be shown similarly by letting $z_t(\tilde{f})$ be the solution of (3.4) and applying Ito's extended formula to the composition $\tilde{\rho}_t^{\mathcal{B}} = z_t\mu_t(\tilde{f})$, where $\mu_t(\tilde{f})$ is the solution to (2.18). □

Write (3.4) in the strong form (by integrating by parts):

$$\frac{\partial}{\partial t} z_t = L_{w^y}^{u^*}(t)^* z_t + \frac{\partial}{\partial u} L_{w^y}^{u^*}(t)^* \nu_t u_t, \ \lim_{t\downarrow 0} z_t = 0. \tag{3.7}$$

From (2.10), (2.11) we obtain

$$\frac{\partial}{\partial u} L^u(t)^*\tilde{f} = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(\frac{\partial}{\partial u} f^i(t,x,u)\tilde{f}\right).$$

Since $\frac{\partial}{\partial u} f^i(t,x,u)$ is bounded and continuous in u, then for $\tilde{f} \in H(R^n)$ we have

$$\frac{\partial}{\partial u} L^u(t)^*\tilde{f} \in \mathcal{L}(R^k; \mathcal{L}(H(R^n); H^{-1}(R^n))).$$

In addition, since U is compact, we also have

$$\frac{\partial}{\partial u} L^u(t)^*\rho_t u_t \in L_y((0,T); H^{-1}(R^n)).$$

Recall that $L_{w^y}^u(t)$ can be represented by

$$L_{w^y}^u(t) = \frac{1}{2}\sum_{i,j=1}^{n} \alpha_{w^y}^{i,j}(t,x)\frac{\partial^2}{\partial x^i \partial x^j} + \sum_{i=1}^{n} b_{w^y}^i(t,x,u)\frac{\partial}{\partial x^i} + d_{w^y}(t,x,u),$$

where the coefficients are a.s. smooth and bounded. Consequently,

$$\frac{\partial}{\partial u} L_{w^y}^u(t)\tilde{f} = \frac{\partial}{\partial u}\left(\sum_{i=1}^{n} b_{w^y}^i(t,x,u)\frac{\partial}{\partial x^i} + d_{w^y}(t,x,u)\right)\tilde{f},$$

and by Lemma 2.3 we deduce that for any w^y,

$$\frac{\partial}{\partial u} L_{w^y}^{u^*}(t)^* \nu_t u_t \in L^2((0,T); H^{-1}(R^n)).$$

Thus, by Lemma 2.3, we conclude that for each w^y there exists one and only one solution to (3.7) in the space

$$z(\cdot) \in L_y^2((0,T); H^1(R^n)),\ \frac{\partial}{\partial t} z(\cdot) \in L_y^2((0,T); H^{-1}(R^n)).$$

In view of Lemma 3.1 we also conclude existence and uniqueness of solutions $\rho_t^{\mathcal{B}}$ corresponding to (3.3).

Lemma 3.2. *Suppose u^*, u are admissible. Then there exists one and only one solution of* (3.3) *(in the strong sense) in the space*

$$\rho^{\mathcal{B}} \in L_y^2((0,T); H^1(R^n)) \cap L^2(\Omega^y, \mathcal{F}_t^y, \mathcal{P}^y; C((0,T); H(R^n))).$$

Proof. The proof follows from the above construction and Lemma 3.1, or by applying the variational methods to (3.3) using the compactness of U. □

Lemma 3.3. *Suppose ν_t^δ denotes the solution to* (2.19) *when control u^δ is used and ν_t denotes its solution when control u^* is used. Then*

$$\sup_{0 \le t \le T} E_{\mathcal{P}^y} \left| \frac{\nu_t^\delta - \nu_t}{\delta} - z_t \right|^2 \to 0 \ as\ \delta \to 0.$$

Proof. Set $\tilde{\nu}_t^\delta = \frac{\nu_t^\delta - \nu_t}{\delta} - z_t$, where ν_t^δ, ν_t, z_t are solutions of the corresponding equations expressed in strong form. Clearly, the proof is complete if we can show that

$$\sup_{0 \le t \le T} E_{\mathcal{P}^y} |\tilde{\nu}_t^\delta|^2 \to 0 \ as\ \delta \to 0,$$

and thus it is sufficient to show that

$$\sup_{0 \le t \le T} E_{\mathcal{P}^y} |\tilde{\nu}_t^\delta \mu_t|^2 \to 0 \ as\ \delta \to 0.$$

From the above definition of $\tilde{\nu}_t^\delta$ we also have

$$\tilde{\nu}_t^\delta(\tilde{f}) = \frac{\nu_t^\delta(\tilde{f}) - \nu_t(\tilde{f})}{\delta} - z_t(\tilde{f}),$$

and, as a direct consequence, we obtain

$$\tilde{\nu}_t^\delta(\tilde{f}) = \int_0^t \tilde{\nu}_r^\delta(\mu_r L^{u^*+\delta u}(r) \mu_r^{-1} \tilde{f}) dr + \int_0^t z_r(\mu_r [L^{u^*+\delta u}(r) - L^{u^*}(r)] \mu_r^{-1} \tilde{f}) dr$$
$$+ \delta \int_0^t \nu_r \left(\mu_r \left[L^{u^*+\delta u}(r) - L^{u^*}(r) - \delta \frac{\partial}{\partial u} L^{u^*}(r) u_r \right] \mu_r^{-1} \tilde{f} \right) dr.$$

Integration by parts yields

$$\tilde{\nu}_t^\delta(\tilde{f}) = \int_0^t \tilde{\nu}_r^\delta(\mu_r L^{u^*+\delta u}(r) \mu_r^{-1} \tilde{f}) dr$$
$$+ \int_0^t \int_0^1 d\lambda \nu_r \left(\mu_r \left[\frac{\partial}{\partial u} L^{u^*+\lambda\delta u}(r) - \frac{\partial}{\partial u} L^{u^*}(r) \right] \mu_t^{-1} \tilde{f} \right) u_r dr$$
$$+ \delta \int_0^t \int_0^1 d\lambda z_r \left(\mu_r \frac{\partial}{\partial u} L^{u^*+\lambda\delta u}(r) \mu_r^{-1} \tilde{f} \right) u_r dr.$$

Denote by $\rho_t^\delta(\tilde{f})$ the solution to (2.12) corresponding to control $u^\delta \in U_{ad}$. From section 2.3 and Lemma 3.1, we now have

$$\rho_t^\delta = \nu_t^\delta \mu_t,\ \tilde{\rho}_t^\delta = \tilde{\nu}_t^\delta \mu_t,\ \tilde{\nu}_t^\delta \mu_t = \frac{\nu_t^\delta \mu_t - \nu_t \mu_t}{\delta} - z_t \mu_t.$$

Thus

$$\tilde{\rho}_t^\delta(\tilde{f}) = \int_{R^n} \tilde{f}(z)\tilde{\rho}_t^\delta(x,z)dz = \int_{R^n} \tilde{\nu}_t^\delta(x,z)\tilde{f}(n_{0,t}(z))\phi_{0,t}(z)dz,$$

where $\phi_{0,t}$ is the exponential term in (2.21). It is now a matter of simple algebra to show that

$$\begin{aligned} d\tilde{\rho}_t^\delta = L^{u^*+\delta u}(t)^* \tilde{\rho}_t^\delta dt &+ \int_0^1 d\lambda \left[\frac{\partial}{\partial u} L^{u^*+\lambda\delta u}(t)^* - \frac{\partial}{\partial u} L^{u^*}(t)^* \right] \rho_t u_t dt \\ &+ \delta \int_0^1 d\lambda \frac{\partial}{\partial u} L^{u^*+\lambda\delta u}(t)^* \rho_t^{\mathcal{B}} u_t dt + \sum_{k=1}^d M_k(t)^* \tilde{\rho}_t^\delta \circ dy_t^k. \end{aligned}$$

Rewriting the Stratonovich stochastic integral in terms of the Ito stochastic integral and using the equality

$$\frac{\partial}{\partial u} L^{u^*+\lambda\delta u}(t) = \frac{\partial}{\partial u}\left(A^{u^*+\lambda\delta u}(t) - \frac{1}{2}M_k(t)^2 \right) = \frac{\partial}{\partial u} A^{u^*+\lambda\delta u}(t),$$

we now obtain

$$\begin{aligned} d\tilde{\rho}_t^\delta = A^{u^*+\delta u}(t)^* \tilde{\rho}_t^\delta dt &+ \int_0^1 d\lambda \left[\frac{\partial}{\partial u} A^{u^*+\lambda\delta u}(t)^* - \frac{\partial}{\partial u} A^{u^*}(t)^* \right] \rho_t u_t dt \\ &+ \delta \int_0^1 d\lambda \frac{\partial}{\partial u} A^{u^*+\lambda\delta u}(t)^* \rho_t^{\mathcal{B}} u_t dt + \sum_{k=1}^d M_k(t)^* \tilde{\rho}_t^\delta dy_t^k. \end{aligned}$$

Applying the Ito differential rule to $|\tilde{\rho}_t^\delta|^2$ we deduce the energy equation

$$\begin{aligned} |\tilde{\rho}_t^\delta|^2 = 2 &\int_0^t \langle A^{u^*+\delta u}(r)\tilde{\rho}_r^\delta, \tilde{\rho}_r^\delta \rangle dr \\ &+ 2\int_0^t \left\langle \tilde{\rho}_r^\delta, \int_0^1 d\lambda \left[\frac{\partial}{\partial u} A^{u^*+\lambda\delta u}(r)^* - \frac{\partial}{\partial u} A^{u^*}(r)^* \right] \rho_r \right\rangle u_r dr \\ &+ 2\delta \int_0^t \left\langle \tilde{\rho}_r^\delta, \int_0^1 d\lambda \frac{\partial}{\partial u} A^{u^*+\lambda\delta u}(r)^* \rho_r^{\mathcal{B}} \right\rangle u_r dr \\ &+ \sum_{k=1}^d \int_0^t |M_k(r)^* \tilde{\rho}_r^\delta|^2 dr + \sum_{k=1}^d \int_0^t (\tilde{\rho}_r^\delta, M_k(r)^* \tilde{\rho}_r^\delta) dy_r^k. \end{aligned}$$

Using the coercivity condition (assumption (A8)) in the above equation and proceeding as in Bensoussan [2, Lemma 8.2.1], we deduce the estimate

$$\sup_{0\le t\le T} E_{\mathcal{P}^y} |\tilde{\rho}_t^\delta|^2 \to 0 \ as\ \delta \to 0.$$

Finally, since $\tilde{\rho}_t^\delta = \tilde{\nu}_t^\delta \mu_t$, we have the desired results. □

The next step is to establish the differentiability of $J(u)$ as a function on the Hilbert space $L^2_y([0,T];R^k)$.

LEMMA 3.4. *The cost function $J(u)$ is Gâteaux differentiable and*

$$\begin{aligned}(3.8)\quad J^{u^\delta} &\triangleq \frac{d}{d\delta}J(u^*+\delta u)|_{\delta=0}\\ &= E_{\mathcal{P}^y}\left\{z_T\mu_T(\kappa(x))+\int_0^T\left[z_r\mu_r(\pi^{u^*}(r))+\nu_r\mu_r\left(\frac{\partial}{\partial u}\pi^{u^*}(r)u_r\right)\right]dr\right\}.\end{aligned}$$

Proof. Suppose we set $\hat{J}$ to be equal to the right side of (3.8), and $\nu^\delta_t=\delta\tilde{\nu}^\delta_t+\delta z_t+\nu_t$; then

$$\begin{aligned}\frac{1}{\delta}\{J(u^*+\delta u)-J(u^*)\} &= E_{\mathcal{P}^y}\int_0^T\left\{\tilde{\nu}^\delta_r\mu_r\left(\pi^{u^*+\delta u}(r)-\pi^{u^*}(r)-\delta\frac{\partial}{\partial u}\pi^{u^*}(r)u(r)\right)\right\}dr\\ &\quad+E_{\mathcal{P}^y}\int_0^T\{z_r\mu_r(\pi^{u^*+\delta u}(r)-\pi^{u^*}(r))\}dr\\ &\quad+E_{\mathcal{P}^y}\int_0^T\tilde{\nu}^\delta_r\mu_r(\pi^{u^*+\delta u}(r))dr.\end{aligned}$$

Letting δ tend to zero, we notice that the first and second terms of the right side of the previous expression tend to zero due to assumption (A5). The last term also tends to zero by Lemma 3.3. □

4. Necessary conditions.

The perturbed process z_t satisfying (3.4) and the variational cost (3.8) can be viewed as the analogue of the deterministic variational problem given by Fleming and Rishel [16, Theorem 10.2, p. 38 and Theorem 11.1, p. 41] as follows. Suppose we introduce the new equation

$$(4.1)\quad \frac{\partial}{\partial t}P_t(x)=-(\mu_tL^{u^*}(t)\mu_t^{-1})P_t(x)-\mu_t(\pi^{u^*}(t))(x),\ \lim_{t\uparrow T}P_t(x)=\mu_T(\kappa)(x),$$

where $\mu_t(\pi^{u^*}(t))$ corresponds to the integral cost and $\mu_T(\kappa)$ corresponds to the terminal cost.

LEMMA 4.1. *Let $\mu_t^{-1}(\cdot)(x,y)$ be the inverse operator of $\mu_t(\cdot)(x,y)$. Then $\tilde{q}_t=\mu_t^{-1}P_t$ is a solution of*

$$d\tilde{q}_t=-L^{u^*}(t)\tilde{q}_tdt-\pi^{u^*}(t)dt-\sum_{k=1}^d M_k(t)\tilde{q}_t\circ dy^k_t,\lim_{t\uparrow T}\tilde{q}_t=\kappa.$$

Moreover, there exists one and only one solution

$$\tilde{q}_t(\cdot)\in L^2_y((0,T);H^1(R^n))\cap L^2(\Omega,\mathcal{F}_{0,t},\mathcal{P};C((0,T);H))$$

such that the preceding equation holds.

Proof. The first part is obtained as in Lemma 3.1, using (3.6). The second part follows from the variational methods of section 2.4. □

If $\pi^{u^*}(t)$ is set to zero, the evaluation of $z_T(P_T)$, using (4.1) and the homogeneous part of (3.4) (e.g., with the second term on the right side of (3.4) set to zero), yields $z_T\mu_T(\kappa)$ (see Kunita [11, Lemma 6.2.7, p. 312]). Comparing (3.4) and (4.1), we have the following result by an application of Kunita [11, Lemma 6.2.7, Theorem 6.2.8, pp. 312–313].

LEMMA 4.2. *The variational cost of Lemma* 3.4 *is given by*

$$J^{u^\delta} = E_{\mathcal{P}^y}\left\{\sum_{j=1}^{k}\int_0^T u_t^j \int_{R^n}\left[\frac{\partial}{\partial u^j}L^{u^*}(t)\tilde{P}_t(x) + \frac{\partial}{\partial u^j}\pi^{u^*}(t)\right]\rho_t(x)dxdt\right\}, \tag{4.2}$$

where $\tilde{P}_t(x) \overset{\triangle}{=} \mu_t^{-1}P_t(x)$.

Proof. Applying the Ito formula given by Kunita [11, Lemma 6.2.7, Theorem 6.2.8, pp. 312–313] to $z_t(P_t)$ gives

$$\begin{aligned} z_T(P_T) &= z_0(P_0) + \int_0^T z_t(\mu_t L^{u^*}(t)\mu_t^{-1}P_t)dt + \int_0^T \nu_t\left(\mu_t\frac{\partial}{\partial u}L^{u^*}(t)u_t\mu_t^{-1}P_t\right)dt \\ &\quad - \int_0^T z_t(\mu_t L^{u^*}(t)\mu_t^{-1}P_t)dt - \int_0^T z_t(\mu_t\pi^{u^*}(t))dt \\ &= \int_0^T\left[\nu_t\mu_t\left(\frac{\partial}{\partial u}L^{u^*}(t)u_t\mu_t^{-1}P_t\right) - z_t(\mu_t\pi^{u^*}(t))\right]dt. \end{aligned} \tag{4.3}$$

Substituting (4.3) into the variational cost of Lemma 3.4 (using $z_T(P_T) = z_T(\mu_T\kappa(x))$) gives

$$J^{u^\delta} = E_{\mathcal{P}^y}\left\{\int_0^T\left[\nu_t\left(\mu_t\frac{\partial}{\partial u}L^{u^*}(t)u_t\mu_t^{-1}P_t\right) + \nu_t\mu_t\left(\frac{\partial}{\partial u}\pi^{u^*}(t)u_t\right)\right]dt\right\}.$$

Replacing $\nu_t\mu_t$ by ρ_t, and setting $\tilde{P}_t = \mu_t^{-1}P_t$, we recover (4.2) (since for each $t \in [0,T], u_t$ is $\mathcal{F}_t^y$-measurable). □

We now have the following necessary conditions of optimality.

THEOREM 4.3. *Suppose* u_t^* *is optimal for the control problem with cost function* (3.1) *and state* ρ_t *satisfying* (2.12). *Then there exists a process* $\overset{\wedge}{P_t}(x)$ *such that the condition*

$$\sum_{j=1}^{k}(u_t^j - u_t^{*j})\left\{\int_{R^n}\left[\frac{\partial}{\partial u^j}\pi(t,x,u_t^*) + \frac{\partial}{\partial u^j}L^{u^*}(t)\overset{\wedge}{P_t}(x)\right]\rho_t(x)dx\right\} \geq 0 \tag{4.4}$$

holds for all $u \in U_{ad}$ *a.e., on* $t \in [0,T]$ *a.s., where* $\overset{\wedge}{P_t}(x) = \mu_t^{-1}E_{\mathcal{P}^w}(P_t(x))$. *The Hamiltonian* $H_t(\rho_t, \overset{\wedge}{P_t}, u_t)$ *is given by*

$$\frac{\partial}{\partial u}H_t(\rho_t,\overset{\wedge}{P_t},u_t) = \frac{\partial}{\partial u}\int_{R^n}[\pi(t,x,u_t) + \tilde{X}^u(t)\overset{\wedge}{P_t}(x)]\rho_t(x)dx.$$

Proof. We start by using the conditional optimality given by Striebel [17, Chapter 4] which states that whenever $u \in U_{ad}$ is conditionally optimal, it is also optimal; see Bensoussan and van Schuppen [18, Definition 2.1]. Thus, by Lemma 4.2,

$$J^{u^\delta} = E_{\mathcal{P}^y}\left\{\int_0^T\int_{R^n}\left[\frac{\partial}{\partial u}L^{u^*}(t)\mu_t^{-1}P_t(x) + \frac{\partial}{\partial u}\pi^{u^*}(t)\right]\rho_t(x)dxu_tdt\right\}$$

and, by reconditioning on $\mathcal{F}_t^y$,

$$J^{u^\delta} = E_{\mathcal{P}^y}\left\{E_{\mathcal{P}^w}\int_0^T\int_{R^n}\left[\frac{\partial}{\partial u}L^{u^*}(t)\mu_t^{-1}P_t(x) + \frac{\partial}{\partial u}\pi^{u^*}(t,x)\right]\rho_t(x)dxu_tdt\right\},$$

where $E_{\mathcal{P}^w}$ is nothing more than the conditional expectation $E(\cdot|\mathcal{F}_t^y)$. However, since ρ_t and μ_t^{-1} are $\mathcal{F}_t^y$-adapted, we apply the result in Kunita [7, Lemma 4] twice and obtain

$$J^{u^\delta} = E_{\mathcal{P}^y}\left\{\int_0^T\int_{R^n}\left[\frac{\partial}{\partial u}L^{u^*}(t)\mu_t^{-1}E_{\mathcal{P}^w}(P_t(x)) + \frac{\partial}{\partial u}\pi^{u^*}(t)\right]\rho_t(x)dxu_tdt\right\}.$$

Since u^* is optimal and admissible for any other control $u \in U_{ad}$ and $\epsilon \in [0,1]$, we have $u_t^* + \epsilon(u_t - u_t^*) \in U$ (i.e., admissible) and, therefore,

$$J(u^* + \epsilon(u-u^*)) \geq J(u^*).$$

Thus, by the Gâteaux differentiability of $J(u)$ (as a function on the Hilbert space $L_y^2([0,T];R^k)$),

$$\begin{aligned}&\frac{d}{d\epsilon}J(u^* + \epsilon(u-u^*))|_{\epsilon=0}\\ &= E_{\mathcal{P}^y}\left\{\sum_{j=1}^d\int_0^T(u_t^j - u_t^{*j})\int_{R^n}\left[\frac{\partial}{\partial u^j}L^{u^*}(t)\mu_t^{-1}E_{\mathcal{P}^w}(P_t(x)) + \frac{\partial}{\partial u^j}\pi^{u^*}(t)\right]\rho_t(x)dxdt\right\}.\end{aligned}$$

Following the derivations presented in Fleming and Rishel [16, Theorem 11.2, p. 41], Bensoussan [10, Chapter VI, Theorem 1.2, pp. 232–234], or Bensoussan [1, Theorem 2.1], we deduce (4.4). □

Remark 4.1. Notice that the necessary condition of optimality established in Theorem 4.3 has as a special case the one given by Bensoussan [1, Theorem 2.1, equation (2.22)] or [2, Theorem 8.2.1, p. 279].

The next step is to determine a stochastic partial differential equation satisfied by the costate process $\overset{\wedge}{P_t}(x)$ identified in Theorem 4.3. To do so, we appeal to the martingale representation results of Bensoussan [1, Lemma 2.6]. Therefore, taking the expectation of (4.1) with respect to measure $\mathcal{P}^w$, we obtain

$$E_{\mathcal{P}^w}\frac{\partial}{\partial t}P_t(x) = E_{\mathcal{P}^w}\{-(\mu_tL^{u^*}(t)\mu_t^{-1})P_t(x) - \mu_t(\pi^{u^*}(t))\}.$$

But $E_{\mathcal{P}^w}[\mu_t(\pi^{u^*}(t))]$ is just $\mu_t(\pi^{u^*}(t))$, which follows from the representation

$$\mu_t\pi^{u^*}(t) = \pi(t, n_{0,t}(x), u_t^*)\exp\left\{\sum_{k=1}^d\int_0^t h_k(r, n_{0,r}(x))\circ dy_r^k\right\},$$

because $n_t(x)$ is the solution to (2.20) and is $\mathcal{F}_{0,t}^y$ measurable. Hence, one deduces

$$E_{\mathcal{P}^w}\frac{\partial}{\partial t}P_t(x) = -(\mu_tL^{u^*}(t)\mu_t^{-1})E_{\mathcal{P}^w}P_t(x) - \mu_t(\pi^{u^*}(t)). \tag{4.5}$$

Define $\bar{P}_t(x) = E_{\mathcal{P}^w}(P_t(x))$; from the martingale representation theorem given in Bensoussan [1, Lemma 2.6],

$$d\bar{P}_t(x) = [-(\mu_t L^{u^*}(t)\mu_t^{-1})\bar{P}_t(x) - \mu_t(\pi^{u^*}(t))]dt - \sum_{k=1}^{d} r_t^k dy_t^k, \tag{4.6}$$

where $\{r_s^k, 0 \leq s \leq t\}, 1 \leq k \leq d$, is a square integrable $\mathcal{F}_t^y$-adapted process. However, since μ_t is a one-to-one and onto map, we can define a new process $\tilde{r}_t^k \overset{\triangle}{=} \mu_t^{-1} r_t^k, 1 \leq k \leq d$, so that the martingale term in (4.6) can be written as

$$\sum_{k=1}^{d} r_t^k dy_t^k = \sum_{k=1}^{d} \mu_t(\tilde{r}_t^k) dy_t^k.$$

Suppose $\tilde{r}_t^k$ is in the domain of $M_k(t), 1 \leq k \leq d$, and that $\tilde{r}_t^k$ is absolutely continuous with respect to t, $1 \leq k \leq d$. Then we can represent the Ito stochastic integral in terms of the Stratonovich stochastic integral

$$\sum_{k=1}^{d} \mu_t(\tilde{r}_t^k) dy_t^k = \sum_{k=1}^{d} \mu_t(\tilde{r}_t^k) \circ dy_t^k - \frac{1}{2}\sum_{k=1}^{d} \mu_t(M_k(t)\tilde{r}_t^k)dt.$$

Therefore, (4.6) becomes

$$d\bar{P}_t(x) = -(\mu_t L^{u^*}(t)\mu_t^{-1})\bar{P}_t(x)dt - \mu_t(\pi^{u^*}(t))dt + \frac{1}{2}\sum_{k=1}^{d} \mu_t(M_k(t)\tilde{r}_t^k)dt - \sum_{k=1}^{d} \mu_t(\tilde{r}_t^k) \circ dy_t^k. \tag{4.7}$$

LEMMA 4.4. *Suppose $\tilde{r}_t^k$ is in the domain of $M_k(t), 1 \leq k \leq d$, and that $\tilde{r}_t^k$ is absolutely continuous with respect to t, $1 \leq k \leq d$. $(\overset{\wedge}{P_t}(x), \{\tilde{r}_t^k\}_{k=1}^d)$ is a solution to*

$$\begin{aligned} d\overset{\wedge}{P_t}(x) = -L^{u^*}(t)\overset{\wedge}{P_t}(x)dt - \pi^{u^*}(t)dt - \sum_{k=1}^{d} \tilde{r}_t^k \circ dy_t^k + \frac{1}{2}\sum_{k=1}^{d} M_k(t)\tilde{r}_t^k dt \\ - \sum_{k=1}^{d} M_k(t)\overset{\wedge}{P_t}(x) \circ dy_t, \end{aligned} \tag{4.8}$$

$$\lim_{t \uparrow T} \overset{\wedge}{P_t}(x) = \kappa(x), \tag{4.9}$$

if and only if $\bar{P}_t'(x) = \mu_t \overset{\wedge}{P_t}(x)$ is a solution of (4.7).

Proof. Here we shall follow the proof given by Kunita [11, Lemma 6.2.3, p. 307]. First set $\bar{P}_t' = \mu_t \overset{\wedge}{P_t}$, where $\overset{\wedge}{P_t}$ is the solution to (4.8). From (2.18),

$$\bar{P}_t'(x) = \overset{\wedge}{P_t}(n_{0,t}(x))\phi_{0,t}(x), \ \phi_{0,t}(x) = \exp\left\{\sum_{k=1}^{d}\int_0^t h_k(r, n_{0,r}(x)) \circ dy_r^k\right\},$$

where n_t is the solution to (2.20). By the Stratonovich version of the extended Ito formula

$$d[\overset{\wedge}{P_t}(n_{0,t}(x))\phi_{0,t}(x)] = \Bigg\{ - L^{u^*}(t)\overset{\wedge}{P_t}(n_{0,t}(x))dt - \pi^{u^*}(t, n_{0,t}(x))dt$$

$$
\begin{aligned}
&-\sum_{k=1}^{d}\tilde{r}_t^k(n_{0,t}(x))\circ dy_t^k\\
&+\frac{1}{2}\sum_{k=1}^{d}M_k(t)\tilde{r}_t^k(n_{0,t}(x))dt-\sum_{k=1}^{d}M_k(t)\hat{P}_t(n_{0,t}(x))\circ dy_t^k\\
&+\sum_{i=1}^{n}\frac{\partial}{\partial x^i}\hat{P}_t(n_{0,t}(x))\circ dn_{0,t}^i(x)\\
&+\hat{P}_t(n_{0,t}(x))\sum_{k=1}^{d}h_k(t,n_{0,t}(x))\circ dy_t^k\Bigg\}\phi_{0,t}(x).
\end{aligned}
$$

Substituting

$$
\sum_{i=1}^{n}\frac{\partial}{\partial x^i}\hat{P}_t(n_{0,t}(x))\circ dn_{0,t}^i(x)=\sum_{k=1}^{d}\sum_{i=1}^{n}Y_k^i(t,n_{0,t}(x))\frac{\partial}{\partial x^i}\hat{P}_t(n_{0,t}(x))\circ dy_t^k
$$

into the previous equation, the fifth term cancels the sixth and seventh terms. As a consequence,

$$
\begin{aligned}
d[\hat{P}_t(n_{0,t}(x))\phi_{0,t}(x)]&=d\bar{P}_t'(x)\\
&=\Bigg\{-L^{u^*}(t)\hat{P}_t(n_{0,t}(x))dt-\pi^{u^*}(t,n_{0,t}(x))dt\\
&\quad-\sum_{k=1}^{d}\tilde{r}_t^k(n_{0,t}(x))\circ dy_t^k+\frac{1}{2}\sum_{k=1}^{d}M_k(t)\tilde{r}_t^k(n_{0,t}(x))dt\Bigg\}\phi_{0,t}(x)\\
&=-(\mu_tL^{u^*}(t)\mu_t^{-1})\bar{P}_t'(x)dt-\mu(\pi^{u^*}(t))dt\\
&\quad-\sum_{k=1}^{d}\mu_t(\tilde{r}_t^k)\circ dy_t^k+\frac{1}{2}\sum_{k=1}^{d}\mu_t(M_k(t)\tilde{r}_t^k)dt.
\end{aligned}
$$

Therefore, $\bar{P}_t'(x)$ satisfies (4.7). Conversely, suppose $\bar{P}_t'(x)$ is a solution to (4.7). We can show similarly that

$$
\begin{aligned}
d[\mu_t^{-1}(\bar{P}_t')]=&-\sum_{k=1}^{d}M_k(t)\mu_t^{-1}(\bar{P}_t')\circ dy_t^k-L^{u^*}(t)\mu_t^{-1}(\bar{P}_t')dt-\pi^{u^*}(t)dt\\
&+\frac{1}{2}\sum_{k=1}^{d}M_k(t)\tilde{r}_t^kdt-\sum_{k=1}^{d}\tilde{r}_t^k\circ dy_t^k,
\end{aligned}
$$

and so $\mu_t^{-1}\bar{P}_t'(x)$ satisfies (4.8). □

We shall next derive the equation satisfied by the costate process $\hat{P}_t(x)$ when the stochastic integral is expressed in terms of an Ito stochastic integral, starting with (4.6), avoiding the transformation of Ito integrals to Stratonovich integrals, and using the assumptions of Lemma 4.4.

THEOREM 4.5. *The pair* $(\hat{P}_t(x),\{\tilde{r}_t^k\}_{k=1}^d)$ *describes the costate process which is a solution of the Ito equation*

$$
d\hat{P}_t(x)=-A^{u^*}(t)\hat{P}_t(x)dt+\sum_{k=1}^{d}M_k(t)^2\hat{P}_t(x)dt-\pi^{u^*}(t)dt+\sum_{k=1}^{d}M_k(t)\tilde{r}_t^kdt
$$

$$-\sum_{k=1}^{d} M_k(t) \hat{P}_t (x) dy_t^k - \sum_{k=1}^{d} \tilde{r}_t^k dy_t^k, \ \lim_{t \uparrow T} \hat{P}_t (x) = \kappa(x). \tag{4.10}$$

Furthermore, there exists one and only one pair

$$\hat{P} (\cdot) \in L_y^2((0,T); H^1(R^n)) \cap L^2(\Omega, \mathcal{F}_{0,t}, \mathcal{P}^y; C((0,T); H)),$$

$$\tilde{r}_t \in (L_y^2((0,T); H^{-1}(R^n)))^d,$$

such that (4.10) *holds.*

Proof. Using the definition of the adjoint process as given in Theorem 4.3, one way to show the validity of (4.10) is to apply the extended Ito formula to the composition $\mu_t \hat{P}_t (x)$, where $\hat{P}_t (x)$ satisfies (4.10), to show that (4.6) is recovered. Since the correlated case requires substantial algebra, let us first consider the uncorrelated case. Thus

$$\mu_t \hat{P}_t (x) = \exp \left\{ \sum_{k=1}^{d} \int_0^t h_k(s) dy_s^k \right\} \hat{P}_t (x),$$

where $\hat{P}_t$ is the solution to (4.10) when M_k is replaced by h_k. But the differential rule yields

$$\begin{aligned} d(\mu_t \hat{P}_t)(x) = \Bigg\{ & \left[-L^{u^*}(t) \hat{P}_t (x) + \frac{1}{2} \sum_{k=1}^{d} h_k(t)^2 \hat{P}_t (x) - \pi^{u^*}(t) + \sum_{k=1}^{d} h_k(t) \tilde{r}_t^k \right] dt \\ & - \sum_{k=1}^{d} h_k(t) \hat{P}_t \, dy_t^k - \sum_{k=1}^{d} \tilde{r}_t^k dy_t^k \\ & + \left[\sum_{k=1}^{d} h_k(t) dy_t^k \hat{P}_t (x) + \frac{1}{2} \sum_{k=1}^{d} h_k(t)^2 \hat{P}_t (x) \right] \\ & + \left[- \sum_{k=1}^{d} h_k(t)^2 \hat{P}_t (x) - \sum_{k=1}^{d} \tilde{r}_t^k h_k(t) \right] \Bigg\} \mu_t \end{aligned}$$

and, after cancellations, we obtain

$$d(\mu_t \hat{P}_t)(x) = -\mu_t (L^{u^*}(t) \hat{P}_t (x) - \pi^{u^*}(t)) dt - \sum_{k=1}^{d} \mu_t \tilde{r}_t^k dy_t^k.$$

Replacing $\hat{P}_t$ by $\mu_t^{-1} P_t$ and using $\tilde{r}_t^k = \mu_t^{-1} r_t^k$ we recover (4.6).

For the correlated case, notice also that μ_t satisfies the stochastic partial differential equation

$$d\mu_t(\tilde{f}) = \sum_{k=1}^{d} \mu_t(M_k(t)\tilde{f}) dy_t^k + \frac{1}{2} \sum_{k=1}^{d} \mu_t(M_k(t)^2 \tilde{f}) dt, \ \lim_{t \downarrow 0} \mu_t(\tilde{f}) = \tilde{f}(x),$$

which is the Ito form of (2.18). The solution of the above equation can be represented as

$$\begin{aligned}\mu_t(\tilde{f}) &= \tilde{f}(n_{0,t}(x)) \exp\left\{\sum_{k=1}^{d}\int_0^t h_k(r, n_{0,r}(x))dy_r^k + \frac{1}{2}\sum_{k=1}^{d}\int_0^t Y_k(r)h_k(r, n_{0,r}(x))dr\right\}\\ &= \tilde{f}(n_{0,t}(x))\phi_{0,t},\end{aligned}$$

where $n_{0,t}$ corresponds to the solution of

$$dn_t = \sum_{k=1}^{d} Y_k(t)dy_t^k + \sum_{k=1}^{d}\sum_{i=1}^{n} Y_k^i(t)\frac{\partial}{\partial x^i}Y_k(t)dt,\ n_0 = x.$$

Therefore,

$$\begin{aligned}d[\mu_t \hat{P}_t (x)] &= d[\hat{P}_t (n_{0,t}(x))\phi_{0,t}]\\ &= d \hat{P}_t (n_{0,t}(x))\phi_{0,t}+ \hat{P}_t (n_{0,t}(x))d\phi_{0,t} + d\langle \hat{P} (n(x)), \phi\rangle_t\\ &= \Bigg\{-A^{u^*}(t) \hat{P}_t (n_{0,t}(x))dt + \sum_{k=1}^{d} M_k(t)^2 \hat{P}_t (n_{0,t}(x))dt - \pi^{u^*}(t, n_{0,t}(x))dt\\ &\quad + \sum_{k=1}^{d} M_k(t)\tilde{r}_t^k(n_{0,t}(x))dt - \sum_{k=1}^{d} M_k(t) \hat{P}_t (n_{0,t}(x))dy_t^k\\ &\quad - \sum_{k=1}^{d} \tilde{r}_t^k(n_{0,t}(x))dy_t^k + \sum_{i=1}^{n}\frac{\partial}{\partial x^i} \hat{P}_t (n_{0,t}(x))dn_{0,t}^i(x)\\ &\quad + \frac{1}{2}\sum_{i,j=1}^{n}\frac{\partial^2}{\partial x^i\partial x^j} \hat{P}_t (n_{0,t}(x))d\langle n^i(x), n^j(x)\rangle_t\\ &\quad + \sum_{j=1}^{n} d\left\langle \frac{\partial}{\partial x^j} \hat{P} (n(x)), n^j(x)\right\rangle\Bigg\}\phi_{0,t}\\ &\quad + \Bigg\{\sum_{k=1}^{d} h_k(t) \hat{P}_t (n_{0,t}(x))dy_t^k + \frac{1}{2}\sum_{k=1}^{d} h_k(t)^2 \hat{P}_t (n_{0,t}(x))dt\\ &\qquad + \frac{1}{2}\sum_{k=1}^{d} Y_k(t)h_k(t) \hat{P}_t (n_{0,t}(x))dt\Bigg\}\phi_{0,t} + d\langle \hat{P} (n(x)), \phi\rangle_t.\end{aligned}$$

Furthermore,

$$A^{u^*}(t) = L^{u^*}(t) + \frac{1}{2}\sum_{k=1}^{d} M_k(t)^2,\ M_k(t)^2 = Y_k(t)^2 + h_k(t)^2 + Y_k(t)h_k(t) + 2h_k(t)Y_k(t),$$

$$\begin{aligned}&\sum_{i=1}^{n}\frac{\partial}{\partial x^i} \hat{P}_t (n_{0,t}(x))dn_{0,t}^i(x)\\ &\quad = \sum_{k=1}^{d}\left(Y_k(t) \hat{P}_t (n_{0,t}(x))dy_t^k + \frac{1}{2}\sum_{i,j=1}^{n} Y_k^i(t)\frac{\partial}{\partial x^i} \hat{P}_t (n_{0,t}(x))\frac{\partial}{\partial x^j}Y_k^j(t)dt\right),\end{aligned}$$

$$\frac{1}{2}\sum_{i,j=1}^{n}\frac{\partial^2}{\partial x^i \partial x^j}\overset{\wedge}{P_t}(n_{0,t}(x))d\langle n^i(x), n^j(x)\rangle_t$$

$$=\frac{1}{2}\sum_{k=1}^{d}\sum_{i,j=1}^{n}Y_k^i(t)Y_k^j(t)\frac{\partial^2}{\partial x^i \partial x^j}\overset{\wedge}{P_t}(n_{0,t}(x))dt,$$

$$\sum_{j=1}^{n}d\left\langle\frac{\partial}{\partial x^j}\hat{P}(n(x)), n^j(x)\right\rangle_t=-\sum_{k=1}^{d}Y_k(t)\tilde{r}_t^k(n_{0,t}(x))dt$$

$$-\sum_{k=1}^{d}(Y_k(t)^2+h_k(t)Y_k(t)+Y_k(t)h_k(t))\overset{\wedge}{P_t}(n_{0,t}(x))dt,$$

$$\phi_{0,t}^{-1}d\langle\hat{P}(n(x)),\phi\rangle_t=-\sum_{k=1}^{d}h_k(t)M_k(t)\overset{\wedge}{P_t}(n_{0,t}(x))dt$$

$$-\sum_{k=1}^{d}h_k(t)\tilde{r}_t^k(n_{0,t}(x))dt+\sum_{k=1}^{d}h_k(t)Y_k(t)\overset{\wedge}{P_t}(n_{0,t}(x))dt,$$

and, after some cancellations, we obtain

$$d[\overset{\wedge}{P_t}(n_{0,t}(x))\phi_{0,t}]=\left\{-L^{u^*}(t)\overset{\wedge}{P_t}(n_{0,t}(x))dt-\pi^{u^*}(t,n_{0,t}(x))dt\right.$$
$$\left.-\sum_{k=1}^{d}\tilde{r}_t^k(n_{0,t}(x))dy_t^k\right\}\phi_{0,t}$$
$$=-(\mu_t L^{u^*}(t)\mu_t^{-1})\bar{P}_t(x)dt-\mu_t\pi^{u^*}(t)dt-\sum_{k=1}^{d}\mu_t(\tilde{r}_t^k)dy_t^k.$$

Finally, by the definition of $\tilde{r}_t^k$, $\mu_t\tilde{r}_t^k=\mu_t\mu_t^{-1}r_t^k=r_t^k$, we arrive at (4.6) as desired. Uniqueness of the processes $\overset{\wedge}{P_t}$ and $\tilde{r}_t^k, 1\le k\le d$, follows if, with the help of the inner product $\langle\overset{\wedge}{P_t},\rho_t^{\mathcal{B}}\rangle$, the variational cost of Lemma 3.4 can be expressed as in (4.4); for a proof, see Bensoussan [2, Theorem 8.2.3]. □

Remark 4.2. Notice that the minimum principle given in Theorem 4.3 and the stochastic partial differential equation given in Theorem 4.5 ((4.10), costate equation) can be obtained through a more simplified approach that does not require the conditional optimality of Striebel [17, Chapter 20] and the martingale representation results of Bensoussan [1, Lemma 2.6]. To see this, suppose the process defined in (4.1) is replaced by q_t satisfying

$$dq_t(x)=-(\mu_t L^{u^*}(t)\mu_t^{-1})q_t(x)dt-\mu_t(\pi^{u^*}(t))dt-\sum_{k=1}^{d}\mu_t(\tilde{r}_t^k)dy_t^k,\ \lim_{t\uparrow T}q_t(\kappa(x))=\mu_T\kappa(x),$$
(4.11)

where, as before, $\bar{r}_t^k, 1 \leq k \leq d$, is an $\mathcal{F}_t^y$-adapted process. Similarly, as in (4.7), we can rewrite (4.11) in terms of the Stratonovich integral, which in differential form becomes

$$dq_t(x) = (\mu_t L^{u^*}(t)\mu_t^{-1})q_t(x)dt - \mu_t(\pi^{u^*}(t))dt - \sum_{k=1}^{d} \mu_t(\bar{r}_t^k) \circ dy_t^k + \frac{1}{2}\sum_{k=1}^{d} \mu_t(M_k(t)\bar{r}_t^k)dt \tag{4.12}$$

and is the same as (4.7). Therefore, if we define $\overset{\wedge}{q_t}(x) \triangleq \mu_t^{-1}(q_t)(x)$, then $\overset{\wedge}{q_t}(x)$ satisfies (4.8). It remains to show that if $\overset{\wedge}{q_t}(x)$ rather than $\overset{\wedge}{P_t}(x)$ is used in (4.4), the minimum principle of Theorem 4.3 can be established, thus implying that $\overset{\wedge}{q_t}(x)$ is the costate process. By the Ito formula,

$$\begin{aligned} z_T(q_T(x)) &= z_T\mu_T(\kappa(x)) \\ &= \int_0^T \left[\nu_t\mu_t\left(\frac{\partial}{\partial u}L^{u^*}(t)u_t\mu_t^{-1}q_t\right)dt - z_t\mu_t\pi^{u^*}(t)dt - \sum_{k=1}^{d} z_t\mu_t(\bar{r}_t^k)\circ dy_t^k \right. \\ &\qquad \left. + \frac{1}{2}\sum_{k=1}^{d} z_t\mu_t(M_k(t)\bar{r}_t^k)dt\right], \end{aligned}$$

and by substituting the above expression into the variational cost of Lemma 3.4, we obtain

$$\begin{aligned} J^{u^\delta} = E_{\mathcal{P}^y}\Bigg\{&\int_0^T \left[\nu_t\mu_t\left(\frac{\partial}{\partial u}L^{u^*}(t)u_t\mu_t^{-1}q_t\right) + \nu_t\mu_t\left(\frac{\partial}{\partial u}\pi^{u^*}(t)u_t\right)\right]dt \\ &+ \sum_{k=1}^{d}\int_0^T\left[-z_t\mu_t(\bar{r}_t^k)\circ dy_t^k + \frac{1}{2}\sum_{k=1}^{d} z_t\mu_t(M_k(t)\bar{r}_t^k)dt\right]\Bigg\}. \end{aligned}$$

Since $\rho_t^{\mathcal{B}} = z_t\mu_t$, then, by (3.3), the last two components of the right side of the previous expression correspond to an Ito integral; hence, using $\overset{\wedge}{q_t}(x) \triangleq \mu_t^{-1}q_t(x)$ and $\rho_t = \nu_t\mu_t$,

$$J^{u^\delta} = E_{\mathcal{P}^y}\left\{\int_0^T \left[\rho_t\left(\frac{\partial}{\partial u}L^{u^*}(t)u_t\overset{\wedge}{q_t}\right) + \rho_t\left(\frac{\partial}{\partial u}\pi^{u^*}(t)u_t\right)\right]dt - \sum_{k=1}^{d}\int_0^T \rho_t^{\mathcal{B}}(\bar{r}_t^k)dy_t^k\right\}.$$

But the martingale term of J^{u^δ} has zero expectation with respect to measure $\mathcal{P}^y$; therefore,

$$J^{u^\delta} = E_{\mathcal{P}^y}\left\{\int_0^T \left[\left\langle\frac{\partial}{\partial u}L^{u^*}(t)\overset{\wedge}{q_t}(x), \rho_t(x)\right\rangle + \left\langle\frac{\partial}{\partial u}\pi^{u^*}(t), \rho_t(x)\right\rangle\right]u_t dt\right\}. \tag{4.13}$$

However, from (4.12), the optimality condition (4.4) can be established, and the relation between $\bar{r}_t^k$ and r_t^k is given by $\bar{r}_t^k = \tilde{r}_t^k = \mu_t^{-1}r_t^k, 1 \leq k \leq d$.

5. Relation to previous work. Let us consider the stochastic control problem investigated by Bensoussan [1, 2], where $\sigma(x) \in R^n \otimes R^n, M_k = h_k, 1 \leq k \leq d$; that is, no correlation between the observation process and state process is allowed. Thus, the stochastic equation describing the unnormalized conditional density of nonlinear filtering becomes

$$d\rho_t + A_0 \rho_t dt = B^{u^*} \rho_t dt + \sum_{k=1}^{d} \rho_t h_k dy_t^k, \tag{5.1}$$

where

$$A_0 \tilde{f} = \frac{1}{2} \sum_{i,j=1}^{n} \frac{\partial}{\partial x^i} \left(a^{ij}(x) \frac{\partial \tilde{f}}{\partial x^j} \right),$$

$$B^{u^*} \tilde{f} = \sum_{i=1}^{n} \frac{\partial}{\partial x^i} (\alpha^i(x, u^*) \tilde{f}), \ \alpha^i(x, u^*) = -f^i(x, u^*) + \frac{1}{2} \sum_{j=1}^{n} \frac{\partial}{\partial x^j} a^{ij}(x).$$

It is easily seen that

$$L^{u^*,*} \tilde{f} = B^{u^*} \tilde{f} - A_0 \tilde{f} - \frac{1}{2} \sum_{k=1}^{d} h_k^2 \tilde{f}$$

and

$$\begin{aligned} \left\langle \overset{\wedge}{P_t} (x), \frac{\partial}{\partial u} L^{u^*,*} \rho_t(x) \right\rangle &= \left\langle \overset{\wedge}{P_t} (x), - \sum_{i=1}^{n} \frac{\partial}{\partial x^i} \left(\frac{\partial}{\partial u} f^i(x, u^*) \rho_t \right) \right\rangle \\ &= \left\langle \sum_{i=1}^{n} \frac{\partial}{\partial x^i} \overset{\wedge}{P_t} (x) \frac{\partial}{\partial u} f^i(x, u^*), \rho_t(x) \right\rangle, \end{aligned} \tag{5.2}$$

which implies the following.

THEOREM 5.1. *If the unnormalized conditional density of nonlinear filtering satisfies* (5.1)*, then the minimum principle of Theorem* 4.3 *is given by*

$$\sum_{j=1}^{k} (u_t^j - u_t^{*j}) \left\{ \int_{R^n} \left[\frac{\partial}{\partial u^j} \pi(x, u_t^*) + \sum_{i=1}^{n} \frac{\partial}{\partial x^i} \overset{\wedge}{P_t} (x) \frac{\partial}{\partial u^j} f^i(x, u_t^*) \right] \rho_t(x) dx \right\} \geq 0,$$

which is exactly the minimum principle presented in Bensoussan [1, *Theorem* 2.1]; [2, *Theorem* 8.2.1, *p.* 279].

Proof. The proof follows from (5.1), (5.2), and the result of Theorem 4.3. □

There is also a connection between the costate process of (4.10) and the costate process given in Bensoussan [1, Theorem 2.2], which is recognized as follows. Suppose we set $\tilde{r}_t^k = -\exp\{-\sum_{j=1}^{d} y_t^j h_j\} \hat{r}_t^k$; then by Lemma 4.4, we have the following.

THEOREM 5.2. *If the unnormalized conditional density of nonlinear filtering satisfies* (5.1)*, then the costate process identified in Theorem* 4.3 *satisfies* (4.10) *and solves*

$$d \overset{\wedge}{P_t} (x) = A_0 \overset{\wedge}{P_T} (x) dt + \sum_{i=1}^{n} a^i(x, u^*) \frac{\partial}{\partial x^i} \overset{\wedge}{P_t} (x) dt - \pi^{u^*} dt + \sum_{k=1}^{d} h_k^2 \overset{\wedge}{P_t} (x) dt$$

$$
+\exp\left\{-\sum_{j=1}^{d} y_t^j h_j\right\}\hat{r}_t^k h_k dt-\sum_{k=1}^{d} h_k \hat{P}_t\,(x)dy_t^k
$$
$$
+\sum_{k=1}^{d}\exp\left\{-\sum_{j=1}^{d} y_t^j h_j\right\}\hat{r}_t^k dy_t^k,\ \lim_{t\uparrow T}\hat{P}_t\,(x)=k(x),
$$

which is exactly the equation satisfied by the costate process established in Bensoussan [1, *Theorem* 2.2].

Proof. Using (5.2) and substituting $\tilde{r}_t^k = -\exp\{-\sum_{j=1}^{d} y_t^j h_j\}\hat{r}_t^k, 1 \le k \le d$, in the costate equation of Lemma 4.4, the result follows. □

6. Conclusion. In this paper we have presented a new approach based on measure-valued decompositions to derive necessary conditions of optimality for partially observed stochastic control problems when correlation between the state process and observation process is present. Let us also note that our approach can be applied to the case when the correlation is zero, but the control appears in both the drift and the diffusion coefficients of the state process. However, in this case, the validity of Lemma 3.3 and Lemma 3.4 must be established. The approach discussed in this paper was first published in [19].

REFERENCES

[1] A. Bensoussan, *Maximum principle and dynamic programming approaches of the optimal control of partially observed diffusions*, Stochastics, 9 (1983), pp. 169–222.

[2] A. Bensoussan, *Stochastic Control of Partially Observable Systems*, Cambridge University Press, Cambridge, UK, 1992.

[3] W. Fleming and M. Nisio, *On stochastic relaxed control for partially observed diffusions*, Nagoya Math. J., 9 (1984), pp. 71–108.

[4] U. G. Haussmann, *The maximum principle for optimal control of diffusions with partial information*, SIAM J. Control Optim., 25 (1987), pp. 341–361.

[5] J. Baras, R. Elliott, and M. Kohlmann, *The partially observed stochastic minimum principle*, SIAM J. Control Optim., 27 (1989), pp. 1279–1292.

[6] R. Elliott and H. Yang, *Control of partially observed diffusions*, J. Optim. Theory Appl., 71 (1991), pp. 485–501.

[7] H. Kunita, *Densities of a measure-valued process governed by stochastic partial differential equation*, Systems Control Lett., 1 (1981), pp. 100–104.

[8] H. Kunita, *Stochastic partial differential equations connected with nonlinear filtering*, in Nonlinear Filtering and Stochastic Control, S. K. Mitter and A. Moro, eds., Lecture Notes in Math. 972, Springer-Verlag, Berlin, 1982, pp. 100–169.

[9] W. M. Wonham, *On the separation theorem of stochastic control*, SIAM J. Control, 6 (1968), pp. 312–326.

[10] A. Bensoussan, *Stochastic Control by Functional Analysis Methods*, North–Holland, Amsterdam, 1982.

[11] H. Kunita, *Stochastic Flows and Stochastic Differential Equations*, Cambridge University Press, New York, 1990.

[12] E. Pardoux, *Equations of nonlinear filtering and applications to stochastic control with partial information*, in Nonlinear Filtering and Stochastic Control, S. K. Mitter and A. Moro, eds., Lecture Notes in Math. 972, Springer-Verlag, Berlin, 1982, pp. 208–248.

[13] H. Kunita, *On the decomposition of solutions of stochastic differential equations*, in Stochastic Integrals, Proceedings of the Durham Symposium, D. Williams, ed., Lecture Notes in Math. 851, Springer-Verlag, Berlin, 1980, pp. 213–255.

[14] H. Kunita, *Cauchy problem for stochastic partial differential equations arising in nonlinear filtering theory*, Systems Control Lett., 1 (1981), pp. 37–41.

[15] H. Kunita, *On backward stochastic differential equations*, in Stochastics, 6 (1981/82), pp. 293–313.

[16] W. FLEMING AND R. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, 1975.

[17] C. STRIEBEL, *Optimal Control of Discrete Time Stochastic Systems*, Springer-Verlag, Berlin, 1975.

[18] A. BENSOUSSAN AND J. H. VAN SCHUPPEN, *Optimal control of partially observable stochastic systems with an exponential-of-integral performance index*, SIAM J. Control Optim., 23 (1985), pp. 599–613.

[19] C. CHARALAMBOUS, *Topics in Nonlinear Stochastic Control, Estimation, and Decision Using a Measure Transformation Approach,* Ph.D. thesis, Old Dominion University, Norfolk, VA, December 1992.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 6, pp. 1701–1725

SIMULTANEOUS STABILIZATION OF LINEAR AND NONLINEAR SYSTEMS BY MEANS OF NONLINEAR STATE FEEDBACK*

BERTINA HO-MOCK-QAI† AND WIJESURIYA P. DAYAWANSA‡

Abstract. The simultaneous stabilization (resp., asymptotic stabilization) of a countable family of control systems consists of finding a control which stabilizes (resp., asymptotically stabilizes) all the systems in the family. In this paper, we introduce a new method which enables us to show that, given any countable family of *stabilizable nonlinear* systems, there exists a continuous state feedback law which simultaneously stabilizes (not asymptotically) the family. Then, by enriching this method, we prove that any finite family of stabilizable linear time invariant (LTI) systems can be simultaneously exponentially stabilized by means of nonlinear time-varying state feedback. We also derive sufficient conditions for the simultaneous asymptotic stabilizability of countably infinite families of LTI systems. Finally, sufficient conditions for the simultaneous asymptotic stabilizability of finite families of *nonlinear* systems are provided and used for the simultaneous asymptotic stabilization of certain pairs of nonlinear homogeneous systems.

Key words. simultaneous stabilization, linear systems, nonlinear systems, continuous state feedback, time-varying state feedback

AMS subject classifications. 93C05, 93C10, 93C60, 93D09

PII. S0363012997315610

1. Introduction.

To the best of our knowledge, the simultaneous asymptotic stabilization of *countably infinite* families of systems and of families of *nonlinear* systems has never been directly addressed in the literature. Actually, *finite families of linear time invariant (LTI) systems* have been the main focus of the literature related to simultaneous stabilization.

Although the simultaneous asymptotic stabilization of two LTI systems by means of LTI feedback has been completely solved for almost 15 years [13, 14], the simultaneous asymptotic stabilization of more than two LTI systems by LTI feedback laws remains a challenging issue. Indeed, necessary and sufficient conditions for the simultaneous asymptotic stabilizability of three LTI systems are provided in Blondel et al. [3], Ghosh and Byrnes [4], and Vidyasagar and Viswanadham [14], but none of them yields an algorithm to decide whether three LTI systems can be simultaneously stabilized by means of LTI feedback. It is therefore not clear whether the simultaneous stabilizability of more than two LTI systems by means of LTI feedback can be computationally decided, and investigation on this matter continues [1, 2].

To overcome the limitations of LTI controllers, the use of time-varying feedback laws and merely continuous feedback laws for simultaneous stabilization has been investigated: Kabamba and Yang [9] established the simultaneous asymptotic stabilizability of finite families of LTI systems by means of time-varying feedback laws which involve both the sampled output of the system and a periodic function of time.

*Received by the editors January 27, 1997; accepted for publication (in revised form) August 5, 1998; published electronically September 1, 1999. This research was partially supported by NSF grants ECS 92-12126 and EEC 94-02384. Part of this research was performed while the authors were with the Electrical Engineering Department and the Institute for Systems Research at the University of Maryland.

http://www.siam.org/journals/sicon/37-6/31561.html

†Ecole Nationale de l'Aviation Civile, 7 av. Edouard Belin, BP 4005, 31055 Toulouse Cédex, France (Bertina_Ho-Mock-Qai@pcmail.enac.fr).

‡Department of Mathematics, Texas Technical University, Lubbock, TX 79409 (daya@math.ttu.edu).

On the other hand, Zhang and Blondel [16] obtained sufficient conditions for the simultaneous asymptotic stabilizability of such families by controllers based on LTI feedback laws combined with zeroth order hold functions and samplers. While both of these design procedures are based on some discretization strategy, Khargonekar, Pascoal, and Ravi [10] introduced a method that does not involve any discretization scheme and proved that any finite family of stabilizable LTI systems can be simultaneously asymptotically stabilized by a periodic linear time-varying feedback law which is *piecewise continuous*. Finally, Petersen [11] derived a necessary and sufficient condition for the simultaneous quadratic asymptotic stabilizability of single-input LTI systems by means of continuous nonlinear feedback.

We stress that to the best of our knowledge, there do not exist any published results on the simultaneous stabilization of *nonlinear systems*, and our goal in this paper is to address this issue. In order to achieve this goal, we are led to solve several intermediate problems.

First, by introducing a new method to interpolate feedback laws, we prove that given any *countable* family of general nonlinear systems, there exists a merely continuous state feedback law which simultaneously stabilizes (not asymptotically) the family if each system of the family is asymptotically stabilizable by means of *continuous* state feedback. We actually find two feedback laws that solve the simultaneous stabilization problem. Although the first law cannot really be used in practical problems, it yields more insight into the construction than the second; however, the second law is more explicit and is used to prove further results. We then show that if each system of the family is globally asymptotically stabilizable, then there exists a continuous state feedback law which not only simultaneously stabilizes the family but also yields boundedness of all the solutions of the corresponding closed-loop systems. Finally, if the systems of the family are LTI and asymptotically stabilizable by means of LTI feedback, we provide a simple procedure to construct a simultaneous stabilizer.

Next, we introduce *time-varying* state feedback laws and we modify the construction used to derive the previous results. We then prove that given any *finite* family of LTI systems that are individually stabilizable by means of LTI state feedback, there exists a nonlinear time-varying state feedback law which is continuous and which simultaneously globally exponentially stabilizes the family. Because our approach does not involve any discretization strategy as in [9, 16], it should be compared to that of Khargonekar, Pascoal, and Ravi [10]. Our approach is actually very different from that of Khargonekar, Pascoal, and Ravi, and the controller that we derive is *nonlinear* and *continuous*, while theirs is *linear* and *piecewise continuous*. We also provide sufficient conditions for the simultaneous asymptotic stabilizability of countably infinite families of LTI systems.

Finally, by extending the previous approach to the nonlinear setting, we are able to derive sufficient conditions for the simultaneous local and global asymptotic stabilizability of finite families of *nonlinear* systems by means of continuous time-varying state feedback laws. We then use these conditions in order to establish the simultaneous asymptotic stabilizability of certain pairs of homogeneous nonlinear control systems.

The paper is organized as follows. In section 2, we review some definitions. We discuss the simultaneous stabilization of countable families of nonlinear systems in section 3. The simultaneous asymptotic stabilization of LTI systems is then discussed in section 4. We provide, in section 5, some sufficient conditions for the simultaneous asymptotic stabilization of nonlinear systems. These sufficient conditions are used in section 6 for the simultaneous asymptotic stabilization of certain pairs of homogeneous systems in $\mathbb{R}^n$. Finally, section 7 contains our conclusions.

2. Definitions.

Before introducing our results, we need some definitions and notation. Let X be a subset of $\mathbb{R}^n$ containing the origin, and let k be a positive integer. A mapping $u : X \to \mathbb{R}^m$ is said to be almost C^k on X, if it is C^k on $X\backslash\{0\}$. Further, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^k$, and for r positive $B_r(0)$ denotes the set $B_r(0) \triangleq \{x \in \mathbb{R}^k,\ \|x\| < r\}$. Let S denote an autonomous (resp., a time-varying) system $\dot{x} = f(x)$ (resp., $\dot{x} = f(t,x)$), where x lies in $\mathbb{R}^n$. We let $x(\cdot, x_0)$ (resp., $x(\cdot, x_0, t_0)$) denote any one of the solutions of S that starts at x_0 (resp., at x_0 at time t_0), for each x_0 in $\mathbb{R}^n$ and each $t_0 \geq 0$. We will adopt the usual convention that the infimum of a real-valued mapping over the empty set is $+\infty$. Finally, for a given symmetric positive definite matrix P, we let $\lambda_{min}(P)$ and $\lambda_{max}(P)$ denote, resp., the smallest and largest eigenvalues of P.

Definition 2.1 (stability of time-varying systems). *Let $f : [0,\infty) \times \mathbb{R}^n \to \mathbb{R}^n$ be a continuous mapping such that $f(t,0) = 0$ for each $t \geq 0$. The system $S : \dot{x} = f(t,x)$ is*

(i) *stable, if for each $\varepsilon > 0$ and each $t_0 \geq 0$, there exists $\delta(\varepsilon, t_0) > 0$ such that we have $\|x(t, x_0, t_0)\| < \varepsilon$ for each $t \geq t_0$ and each x_0 in $B_{\delta(\varepsilon,t_0)}(0)$;*

(ii) *locally asymptotically stable if it is stable according to case* (i) *and for each $t_0 \geq 0$, there exists $\bar{\delta}(t_0) > 0$ such that* $\lim_{t\to+\infty} x(t, x_0, t_0) = 0$*, for each x_0 in $B_{\bar{\delta}(t_0)}(0)$;*

(iii) *locally exponentially stable if there exist some positive reals γ, $\bar{\delta}$, and L such that*

$$\|x(t, x_0, t_0)\| \ \leq\ L\|x_0\|e^{-\gamma(t-t_0)}, \quad t \geq t_0, \quad t_0 \geq 0, \quad x_0 \in B_{\bar{\delta}}(0);$$

(iv) *locally uniformly stable with exponential (uniform in t_0) convergence if there exist some positive reals γ and $\bar{\delta}$, and a mapping $h : (0,\infty) \to (0,\infty)$ such that* $\lim_{r\to0^+} h(r) = 0$ *and*

$$\|x(t, x_0, t_0)\| \ \leq\ h(\|x_0\|)e^{-\gamma(t-t_0)}, \quad t \geq t_0, \quad t_0 \geq 0, \quad x_0 \in B_{\bar{\delta}}(0).$$

Note that whenever the mapping f does not explicitly depend on t, the positive reals δ and $\bar{\delta}$ that appear in cases (i) *and* (ii) *are independent of t_0.*

Definition 2.1, case (iv) is more general than that of exponential stability but reduces to this concept if the mapping h satisfies $h(r) \leq \alpha r$ for some constant $\alpha > 0$ and for each $r > 0$ close to 0. Although we allow general mappings h, the fundamental idea of uniform (in t_0) exponential convergence is preserved. Furthermore, it is plain that the requirement $\lim_{r\to0^+} h(r) = 0$ yields uniform stability.

Throughout this paper, the words *stable* and *stabilize* are used in the basic sense and refer to the concept of stability as given in Definition 2.1, case (i), while in the control theory literature they usually refer to the concept of *asymptotic stability* according to Definition 2.1, case (ii). Further, we will often omit the term *locally* and unless otherwise stated *asymptotically stable* will mean *locally asymptotically stable.*

Finally, a feedback law *simultaneously* stabilizes (resp., asymptotically stabilizes) a countable family of control systems if it stabilizes (resp., asymptotically stabilizes) *all* the systems in the family.

Definition 2.2. *Let D be a neighborhood of the origin in $\mathbb{R}^n$, and let $f : D \to \mathbb{R}^n$ be continuous. A positive definite mapping $V : D \to [0,\infty)$ is a Lyapunov function* *for the system $\dot{x} = f(x)$ if V is C^0 on D, C^1 on $D\backslash\{0\}$, and*

$$\nabla V(x) f(x) \;\equiv\; \left[\frac{\partial V(x)}{\partial x_1}, \cdots, \frac{\partial V(x)}{\partial x_n}\right] f(x) \;<\; 0, \quad x \in D\backslash\{0\}.$$

Definition 2.3. *Let $\{x_m\}_{m\in\mathbb{Z}}$ be a sequence of positive integers. Further, for each $i = 1, \dots, x_n$ and each n in $\mathbb{Z}$, let Q_i^n belong to a given class of mathematical objects. Then, $\{Q_i^n,\ i = 1, \dots, x_n\}_{n=1}^{\infty}$, $\{Q_i^n,\ i = 1, \dots, x_n\}_{n=-1}^{-}\infty$, and $\{Q_i^n,\ i = 1, \dots, x_n\}_{n\in\mathbb{Z}}$ denote, resp., the three sequences*

$$\{Q_1^1, \dots, Q_{x_1}^1, Q_1^2, \dots, Q_{x_2}^2, Q_1^3, \dots\},$$
$$\{\dots, Q_{x_{-3}}^{-3}, Q_1^{-2}, \dots, Q_{x_{-2}}^{-2}, Q_1^{-1}, \dots, Q_{x_{-1}}^{-1}\},$$
$$\{\dots, Q_1^{-1}, \dots, Q_{x_{-1}}^{-1}, Q_1^0, \dots, Q_{x_0}^0, Q_1^1, \dots, Q_{x_1}^1, \dots\}.$$

3. Simultaneous stabilization.

Throughout this section, we consider a countable family $\{S_i,\ i \in \mathcal{I}\}$ of systems

$$S_i: \quad \dot{x} \;=\; f_i(x, u), \quad i \in \mathcal{I}, \tag{1}$$

where the set $\mathcal{I}$ contains more than one element, the state x lies in $\mathbb{R}^n$, the input u belongs to $\mathbb{R}^m$, and for each i in $\mathcal{I}$ the mapping $f_i : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is continuous on a neighborhood of the origin with $f_i(0, 0) = 0$.

Under these assumptions, we show that if each system S_i can be asymptotically stabilized then there exists a continuous feedback law which simultaneously stabilizes (not asymptotically) the family $\{S_i,\ i \in \mathcal{I}\}$.

3.1. Main result.

In this subsection, we establish the following theorem.

Theorem 3.1. *Let k be in $\{0, 1, \dots\}$. Assume that for each i in $\mathcal{I}$, there exists a state feedback law $u_i : \mathbb{R}^n \to \mathbb{R}^m$ such that $u_i(0) = 0$, u_i is continuous and almost C^k on some neighborhood of the origin, and u_i locally asymptotically stabilizes the system S_i. Then, there exists a state feedback law $v : \mathbb{R}^n \to \mathbb{R}^m$, such that $v(0) = 0$, v is continuous and almost C^k on some neighborhood of the origin, and v simultaneously stabilizes (not asymptotically) the family $\{S_i,\ i \in \mathcal{I}\}$.*

We prove the theorem in the case where the set $\mathcal{I}$ is countably infinite. Thus, we may assume that $\mathcal{I} = \{1, 2, \dots\}$. The result for finite families of systems will clearly follow from the fact that simultaneous stabilizability of the countable family $\{S_1, S_2, \dots\}$ yields simultaneous stabilizability of any finite family $\{S_1, \dots, S_n\}$.

The main lines of the proof are as follows: For each $i = 1, 2, \dots$, we let V_i denote a Lyapunov function for the system $\dot{x} = f_i(x, u_i(x))$. We define a sequence of neighborhoods of the origin $\{U_i^n, i = 1, 2, \dots\}_{n=i}^{\infty}$ such that for each $i = 1, 2, \dots$ the boundaries of the sets U_i^n, $n = i, i + 1, \dots$, are level sets of V_i. We then design a continuous feedback law v which is equal to u_i on the boundaries of the sets U_i^n, $n = i, i + 1, \dots$, for each $i = 1, 2, \dots$. It follows that for each $i = 1, 2, \dots$ and each $n = 1, 2, \dots$, the set $\overline{U}_i^n$ is invariant with respect to the system $\dot{x} = f_i(x, v(x))$. We conclude that for each $i = 1, 2, \dots$, the feedback law v stabilizes S_i upon noting that the family $\{U_i^n\}_{n=i}^{\infty}$ is a topological base at the origin.

Proof. Throughout the proof, we assume that the set $\mathcal{I}$ is equal to $\{1, 2, \dots\}$. For each $i = 1, 2, \dots$, we let D_i be a bounded neighborhood of the origin and $V_i : \overline{D}_i \to [0, \infty)$ be a Lyapunov function such that the mappings u_i and $f_i(\cdot, u_i(\cdot))$ are continuous on $\overline{D}_i$, the mapping u_i is almost C^k on $\overline{D}_i$, and

$$\nabla V_i(x)\, f_i(x, u_i(x)) \;<\; 0, \quad x \in \overline{D}_i\backslash\{0\}. \tag{2}$$

For each $i = 1, 2, \ldots$, we let W_i^β denote the set $W_i^\beta \triangleq \{x \in D_i : V_i(x) < \beta\}$ for each $\beta > 0$, and we define the mapping $\bar{u}_i : D_1 \to \mathbb{R}^m$ by setting

$$\bar{u}_i(x) = \begin{cases} u_i(x), & x \in D_1 \cap D_i, \\ 0, & x \in D_1 \backslash D_i. \end{cases}$$

Let $\theta > 0$. Using the continuity of the mapping u_i on D_i and the fact that $u_i(0) = 0$ for each $i = 1, 2, \ldots$, it is not hard to construct a sequence of positive reals $\{\beta_i^n,\ i = 1, \ldots, n\}_{n=1}^\infty$ [5, Lemma 3.3, p. 51] such that

$$\beta_n^n < \inf_{x \in \partial D_n} V_n(x), \quad n = 1, 2, \ldots,$$

$$\beta_i^n \to 0 \text{ as } n \to \infty, \quad i = 1, 2, \ldots, \tag{3}$$

$$D_n \supset W_{n-1}^{\beta_{n-1}^{n-1}} \supset \overline{W}_1^{\beta_1^n}, \quad n = 2, 3, \ldots, \tag{4}$$

$$W_{i-1}^{\beta_{i-1}^n} \supset \overline{W}_i^{\beta_i^n}, \ i = 2, \ldots, n, \quad n = 2, 3, \ldots, \tag{5}$$

and

$$\|u_k(x)\| < \frac{\theta}{n}, \quad x \in D_k \cap W_n^{\beta_n^n}, \quad k = 1, \ldots, n+2, \quad n = 1, 2, \ldots. \tag{6}$$

Next, we set $U_i^n \triangleq W_i^{\beta_i^n}$ for each $i = 1, \ldots, n$ and each $n = 1, 2, \ldots$. From the inequality $\beta_1^1 < \inf_{x \in \partial D_1} V_1(x)$, we get the inclusion $D_1 \supset U_1^1$. Thus, (4) and (5) yield a sequence of nested neighborhoods

$$\begin{array}{ccccccc} D_1 & \supset & U_1^1 & \supset & & & \\ & & U_1^2 & \supset & U_2^2 & \supset & \\ & & U_1^3 & \supset & U_2^3 & \supset & U_3^3 \quad \supset \\ & & U_1^4 & \supset & \ldots & \ldots & \ldots \\ & & \vdots & \vdots & & \vdots & \end{array} \tag{7}$$

such that each neighborhood contains the closure of the neighborhood that follows.

For each $n = 1, 2, \ldots$ and each $i = 1, \ldots, n$, we define the sets Δ_i^n by setting

$$\Delta_1^1 \triangleq D_1 \backslash \overline{U}_1^2, \qquad \text{with} \qquad \Delta_1^n \triangleq U_{n-1}^{n-1} \backslash \overline{U}_2^n, \quad n = 2, 3, \ldots,$$
$$\Delta_i^n \triangleq U_{i-1}^n \backslash \overline{U}_{i+1}^n, \quad i = 2, \ldots, n-1, \ n = 3, 4, \ldots,$$
$$\Delta_n^n \triangleq U_{n-1}^n \backslash \overline{U}_1^{n+1}, \quad n = 2, 3, \ldots.$$

Because the family $\{U_i^n,\ i = 1, \ldots, n\}_{n=1}^\infty$ is a topological base at the origin (this follows from (3)), we obtain from (7) that the family $\{\Delta_i^n,\ i = 1, \ldots, n\}_{n=1}^\infty$ is an open cover of $D_1 \backslash \{0\}$. We let $\{q_i^n,\ i = 1, \ldots, n\}_{n=1}^\infty$ be a partition of unity subordinate to $\{\Delta_i^n,\ i = 1, \ldots, n\}_{n=1}^\infty$, such that the support of q_i^n is included in Δ_i^n for each $i = 1, \ldots, n$ and each $n = 1, 2, \ldots$ [15, p. 10]. Finally, we define the feedback law $v : D_1 \to \mathbb{R}^m$ by setting

$$v(x) = \begin{cases} 0, & x = 0, \\ \displaystyle\sum_{n=1}^\infty \sum_{i=1}^n \bar{u}_i(x)\, q_i^n(x), & x \in D_1 \backslash \{0\}. \end{cases}$$

The mapping v is almost C^k. Let x be in $D_1\backslash\{0\}$ and let r be in $(0, \|x\|)$ with $\overline{B_r(0)} \subset D_1$. Because $\{U_i^n,\ i = 1, \dots, n\}_{n=1}^{\infty}$ is a topological base at the origin composed of nested neighborhoods, there exists an integer N such that

$$U_N^N \cup \{U_i^n,\ i = 1, \dots, n\}_{n=N+1}^{\infty} \ \subset\ \overline{B_r(0)}.$$

This and the fact that the support of each mapping q_i^n is included in Δ_i^n yield

$$v(y) = \sum_{n=1}^{N} \sum_{i=1}^{n} \bar{u}_i(y) q_i^n(y), \quad y \in D_1\backslash\overline{B_r(0)}. \tag{8}$$

We now show that v is C^k on the neighborhood $D_1\backslash\overline{B_r(0)}$ of x by proving that the mapping $\bar{u}_i\, q_i^n : D_1\backslash\{0\} \to \mathbb{R}^m$ is C^k on $D_1\backslash\{0\}$ for each $i = 1, \dots, n$ and each $n = 1, 2, \dots$.

We fix $n = 2, 3, \dots$ and $i = 1, \dots, n$. By definition of $\bar{u}_i$ we have $\bar{u}_i = u_i$ on D_i, so that

$$\bar{u}_i(y)\, q_i^n(y) \ = \ u_i(y)\, q_i^n(y), \quad y \in (D_1 \cap D_i)\backslash\{0\}. \tag{9}$$

Further, as the support of q_i^n is included in Δ_i^n, we get

$$\bar{u}_i(y)\, q_i^n(y) \ = \ 0, \quad y \in \left(D_1\backslash\overline{\Delta}_i^n\right)\backslash\{0\}. \tag{10}$$

Because q_i^n is smooth on $D_1\backslash\{0\}$ and u_i is C^k on $D_i\backslash\{0\}$, it follows from (9) and (10) that $\bar{u}_i\, q_i^n$ is C^k on

$$\left(\left(D_1\backslash\overline{\Delta}_i^n\right) \cup (D_1 \cap D_i)\right)\backslash\{0\}. \tag{11}$$

Next, by combining the definition of the sets Δ_i^n with (4) and (7), we obtain for each $n = 2, 3, \dots$ the inclusions

$$\begin{aligned}
&\overline{\Delta}_1^n \subset \overline{U}_{n-1}^{n-1} \ \subset\ D_1, \\
&\overline{\Delta}_i^n \subset \overline{U}_{i-1}^{n} \ \subset\ U_i^{n-1} \ \subset\ D_i, \quad i = 1, \dots, n-1\,, \\
&\overline{\Delta}_n^n \subset \overline{U}_{n-1}^{n} \ \subset\ D_n \qquad\qquad \text{(follows from (4))},
\end{aligned}$$

and because we also have $\overline{\Delta}_i^n \subset D_1\backslash\{0\}$, we get $\overline{\Delta}_i^n \ \subset\ (D_1 \cap D_i)\backslash\{0\}$. This implies that the set in (11) is equal to $D_1\backslash\{0\}$, and it follows that the mapping $\bar{u}_i\, q_i^n$ is C^k on $D_1\backslash\{0\}$. Furthermore, the mapping $\bar{u}_1 q_1^1$ is C^k on $D_1\backslash\{0\}$, since by definition of $\bar{u}_1$ we have $\bar{u}_1 q_1^1 = u_1 q_1^1$ on $D_1\backslash\{0\}$. In view of (8), we conclude that for each x in $D_1\backslash\{0\}$, there exists a neighborhood U_x of x included in $D_1\backslash\{0\}$ such that the mapping v is C^k on U_x. In short, the mapping v is C^k on $D_1\backslash\{0\}$.

Continuity of v. We fix $n = 2, 3, \dots$. From the definition of the sets Δ_j^m, it is easily checked that for each $m = n+2, n+3, \dots$, we have

$$\left(U_{n-1}^{n-1}\backslash\overline{U}_{n+1}^{n+1}\right) \cap \Delta_j^m = \emptyset, \quad j = 1, \dots, m\,,$$

and because the support of each function q_j^m is included in Δ_j^m, we get

$$v(x) \ = \ \sum_{m=1}^{n+1} \ \sum_{j=1}^{m} \bar{u}_j(x)\, q_j^m(x), \quad x \in U_{n-1}^{n-1}\backslash\overline{U}_{n+1}^{n+1}.$$

As the functions of $\{q_i^n, i = 1, \ldots, n\}_{n=1}^{\infty}$ sum up to 1, the previous equality implies that

$$\|v(x)\| \;\le\; \max(\|\bar{u}_1(x)\|, \ldots, \|\bar{u}_{n+1}(x)\|), \quad x \in U_{n-1}^{n-1} \backslash \overline{U}_{n+1}^{n+1},$$

so that (6) combined with the definition of the mappings $\bar{u}_i,\ i = 1, 2, \ldots,$ yield

$$\|v(x)\| \;<\; \frac{\theta}{n-1}, \quad x \in U_{n-1}^{n-1} \backslash \overline{U}_{n+1}^{n+1}. \tag{12}$$

Further, upon noting that the family $\{U_l^l\}_{l=1}^{\infty}$ is a topological base at the origin composed of nested neighborhoods such that each neighborhood contains the closure of the neighborhood that follows, we deduce that

$$U_{l-1}^{l-1} \backslash \{0\} \;=\; \bigcup_{n=l}^{\infty} \left(U_{n-1}^{n-1} \backslash \overline{U}_{n+1}^{n+1} \right), \quad l = 2, 3, \ldots,$$

and (12) implies that

$$\|v(x)\| \;\le\; \frac{\theta}{l-1}, \quad x \in U_{l-1}^{l-1} \backslash \{0\}, \quad l = 2, 3, \ldots.$$

Because $\frac{\theta}{l-1} \to 0 \;\; as \;\; l \to \infty$, continuity of v at the origin follows.

Simultaneous stability. From the definitions of the sets U_i^n and Δ_i^n, it is not hard to see that for each $i = 1, 2, \ldots$ and each $n = 1, 2, \ldots$, the boundary ∂U_i^n is included in Δ_i^n and does not intersect with any other set Δ_j^m. Thus, because the support of the mapping q_i^n is included in Δ_i^n for each $i = 1, 2, \ldots$ and each $n = 1, 2, \ldots$, it follows from the definition of v that

$$v(x) = u_i(x), \quad x \in \partial U_i^n, \quad i = 1, \ldots, I, \quad n = 1, 2, \ldots.$$

This, together with (2) and the fact that the sets $\overline{U}_i^n$ are included in D, yields

$$\nabla V_i(x)\, f_i(x, v(x)) \;<\; 0, \quad x \in \partial U_i^n, \quad i = 1, \ldots, I, \quad n = 1, 2, \ldots. \tag{13}$$

For each $i = 1, \ldots, I$ and each $n = 1, 2, \ldots$, by combining (13) with Lemma 8.1 applied with $D,\ V = V_i,\ f(\cdot, \cdot) = f_i(\cdot, v(\cdot))$, and $b(\cdot) \equiv \beta_i^n$, we obtain that the set $\overline{U}_i^n$ is invariant with respect to the system $\dot{x} = f_i(x, v(x))$. Thus, for each $i = 1, \ldots, I$, the family $\{U_i^n\}_{n=1}^{\infty}$ is a topological base at the origin such that $\overline{U}_i^n$ is positively invariant with respect to the system $\dot{x} = f_i(x, v(x))$ for each $n = 1, 2, \ldots$. It follows that v stabilizes (not asymptotically) the system S_i for each $i = 1, \ldots, I$. □

3.2. A more explicit simultaneous stabilizer. In this subsection, we show that we can circumvent the computation of the partition of unity that appears in the expression of the previous simultaneous stabilizer, and we provide a more explicit simultaneous stabilizer which can actually be used in *practical* problems. Although this simultaneous stabilizer is more explicit than the one presented in the previous subsection, it somehow yields less insight into the construction. Furthermore, we will use this more explicit stabilizer, in this paper, to prove further results.

For the sake of simplicity, we assume throughout this subsection that the set $\mathcal{I}$ is finite and equal to $\{1, \ldots, I\}$ (with $I \ge 2$). However, the construction that we provide can be easily transposed to the case of a countably infinite family of systems (see [5, p. 40] for further details).

In that case, a simpler proof of Theorem 3.1 may be given (we refer the interested reader to [8] for a proof). The main difference between this proof and the one presented in the previous subsection lies in the structure of the topological base that is considered. We actually define a sequence of neighborhoods of the origin $\{U_i^n, i = 1, \ldots, I\}_{n=1}^\infty$ such that for each $i = 1, \ldots, I$ the boundaries of the sets U_i^n, $n = 1, 2, \ldots$ are level sets of V_i. Using a partition of unity $\{q_i^n,\ i = 1, \ldots, n\}_{n=1}^\infty$ subordinate to $\{U_i^n, i = 1, \ldots, I\}_{n=1}^\infty$, we then design a continuous feedback law v which is equal to u_i on the boundaries of the sets U_i^n, $n = 1, 2, \ldots,$ for each $i = 1, 2, \ldots$. By using an argument similar to that used in the proof of Theorem 3.1, we obtain that v stabilizes the system S_i for each $i = 1, \ldots, I$.

We now show that we can circumvent the computation of the partition of unity $\{q_i^n, i = 1, \ldots, I\}_{n=1}^\infty$ that appears in the expression of the obtained simultaneous stabilizer.

Derivation of a more explicit simultaneous stabilizer. Under the enforced assumptions, there exist a bounded neighborhood of the origin D and a C^k (where $k \geq 0$ is an integer) Lyapunov function $V_i : \overline{D} \to [0, \infty)$ such that

$$\nabla V_i(x)\, f_i(x, u_i(x)) \ < \ 0, \quad x \in \overline{D} \backslash \{0\}, \tag{14}$$

for each $i = 1, \ldots, I$. Without loss of generality, we may assume that for each $i = 1, \ldots, I$, the mappings $f_i(\cdot, u_i(\cdot))$ and u_i are continuous on $\overline{D}$ with u_i almost C^k on $\overline{D}$.

For each $i = 1, \ldots, I$, and each $\beta > 0$, we set $W_i^\beta \ \triangleq \ D \cap V_i^{-1}(\,[0, \beta)\,)$. It is not hard to obtain [5, Lemma 2.5, p. 32] three sequences of positive reals $\{\alpha_i^n,\ i = 1, \ldots, I\}_{n=1}^\infty$, $\{\beta_i^n,\ i = 1, \ldots, I\}_{n=1}^\infty$, and $\{\gamma_i^n,\ i = 1, \ldots, I\}_{n=1}^\infty$ converging to the origin, such that for each $n = 1, 2, \ldots$, we have

$$\inf_{x \in \partial D} V_i(x) \ > \ \gamma_i^n \ > \ \beta_i^n \ > \ \alpha_i^n, \quad i = 1, \ldots, I, \tag{15}$$

$$W_I^{\alpha_I^n} \ \supset \ \overline{W_1^{\gamma_1^{n+1}}} \qquad \text{and} \qquad W_{i-1}^{\alpha_{i-1}^n} \ \supset \ \overline{W_i^{\gamma_i^n}}, \quad i = 2, \ldots, I. \tag{16}$$

For each $i = 1, \ldots, I$ and each $n = 1, 2, \ldots,$ we define the mappings $\overline{q}_i^n : D \to [0, 1]$ by setting

$$\overline{q}_i^n(x) = \begin{cases} e^{\frac{\left(V_i(x) - \beta_i^n\right)^2}{\left(V_i(x) - \beta_i^n\right)^2 - \left(\beta_i^n - \alpha_i^n\right)^2}} & \text{if } \ V_i(x) \in (\alpha_i^n, \beta_i^n), \\ e^{\frac{(V_i(x) - \beta_i^n)^2}{\left(V_i(x) - \beta_i^n\right)^2 - \left(\gamma_i^n - \beta_i^n\right)^2}} & \text{if } \ V_i(x) \in [\beta_i^n, \gamma_i^n), \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

for each x in D. It can be checked [5, Lemma B.6, p. 159] that the mapping $\overline{q}_i^n$ is C^k on D for each $i = 1, \ldots, I$ and each $n = 1, 2, \ldots$. Finally, we let $\bar{v} : D \to \mathbb{R}^m$ be given by

$$\bar{v}(x) \ = \ \sum_{n=1}^{\infty} \sum_{i=1}^{I} \overline{q}_i^n(x) u_i(x), \quad x \in D. \tag{18}$$

PROPOSITION 3.2. *The state feedback law $\bar{v}$ as given in* (18) *is continuous and almost C^k on D and simultaneously stabilizes (not asymptotically) the family $\{S_i,\ i = 1, \ldots, I\}$.*

Proof. First, we note that (15) combined with (16) yields a sequence of nested neighborhoods

$$D \supset W_1^{\gamma_1^1} \supset W_1^{\beta_1^1} \supset W_1^{\alpha_1^1} \supset W_2^{\gamma_2^1} \supset \cdots \supset W_I^{\alpha_I^1} \supset W_1^{\gamma_1^2} \supset \cdots \tag{19}$$

such that each neighborhood contains the closure of the neighborhood that follows. It is not hard to see from the definition of the mappings $\overline{q}_i^n$ that

$$\{x \in D : \overline{q}_i^n(x) \neq 0\} \;=\; W_i^{\gamma_i^n} \backslash \overline{W}_i^{\alpha_i^n}, \quad i = 1, \ldots, I, \quad n = 1, 2, \ldots, \tag{20}$$

so that these sets are disjoint. Now, let x be in $D\backslash\{0\}$ and let r be in $(0, \|x\|)$ such that $\overline{B_r(0)} \subset D$. Further, let the integer n_r be such that $W_1^{\gamma_1^n} \subset \overline{B_r(0)}$, $n = n_r + 1, n_r + 2, \ldots$. It follows from the definition of $\bar{v}$, together with (20), that

$$\bar{v}(y) \;=\; \sum_{n=1}^{n_r} \sum_{i=1}^{I} u_i(y)\overline{q}_i^n(y), \quad y \in D\backslash\overline{B_r(0)}. \tag{21}$$

Because the mappings u_i and $\overline{q}_i^n$ are C^k on $D\backslash\{0\}$ for each $i = 1, \ldots, I$ and each $n = 1, 2, \ldots$ (see [5, Lemma B.6, p. 159] for further details), we easily obtain from (21) that $\bar{v}$ is C^k on $D\backslash\{0\}$. Furthermore, (20) implies that

$$\|\bar{v}(x)\| \;\leq\; \max\left(\|u_1(x)\|, \ldots, \|u_I(x)\|\right), \quad x \in D,$$

and continuity of $\bar{v}$ at the origin follows from that of the mappings $u_i, i = 1, \ldots, I$.

Simultaneous stability. From (20) and the definition of the mappings $\overline{q}_i^n$, we deduce that for each $i = 1, \ldots, I$ and each $n = 1, 2, \ldots$, we have

$$\overline{q}_i^n(x) \;=\; 1 \quad \text{with} \quad \overline{q}_j^m(x) \;=\; 0, \quad x \in \partial W_i^{\beta_i^n}, \quad (j, m) \neq (i, n),$$

and the definition of $\bar{v}$ yields $\bar{v}(x) = u_i(x)$, $x \in \partial W_i^{\beta_i^n}$. By an argument similar to that used in the proof of Theorem 3.1 to establish simultaneous stability, it follows from (14) and Lemma 8.1 that for each $i = 1, \ldots, I$ the mapping $\bar{v}$ stabilizes S_i, which completes the proof. □

We now assume that there exists a continuous and almost C^k feedback law $u_i : \mathbb{R}^n \to \mathbb{R}^m$ which *globally* asymptotically stabilizes S_i and that the mapping $f_i(\cdot, u_i(\cdot))$ is continuous on $\mathbb{R}^n$ for each $i = 1, \ldots, I$. In that case, the previous construction can be slightly modified in order to yield a feedback law $\widehat{v}$ that simultaneously stabilizes the family $\{S_i,\ i = 1, \ldots, I\}$ and in such a way that all the solutions of the closed-loop system $\dot{x} = f_i(x, \widehat{v}(x))$ are bounded for each $i = 1, \ldots, I$.

Indeed, we may assume that the Lyapunov function V_i is radially unbounded for all $i = 1, \ldots, I$. It can easily be seen [5, Lemma 2.2, p. 28] that there exist some two-sided sequences of positive reals $\{\alpha_i^n,\ i = 1, \ldots, I\}_{n\in\mathbb{Z}}$, $\{\beta_i^n,\ i = 1, \ldots, I\}_{n\in\mathbb{Z}}$, and $\{\gamma_i^n,\ i = 1, \ldots, I\}_{n\in\mathbb{Z}}$ converging to 0 and $+\infty$ as n tends to $+\infty$ and $-\infty$, resp., and satisfying

$$W_I^{\alpha_I^n} \supset \overline{W}_1^{\gamma_1^{n+1}} \qquad \text{and} \qquad W_{i-1}^{\alpha_{i-1}^n} \supset \overline{W}_i^{\gamma_i^n}, \quad i = 2, \ldots, I, \quad n \in \mathbb{Z}. \tag{22}$$

For each $i = 1, \ldots, I$ and each n in $\mathbb{Z}$, we let the mapping $\overline{q}_i^n : \mathbb{R}^n \to [0, 1]$ be given by the formula (17) with α_i^n, β_i^n, and γ_i^n as defined here, and we let the mapping $\widehat{v} : \mathbb{R}^n \to \mathbb{R}^m$ be given by

$$\widehat{v}(x) \;=\; \sum_{n\in\mathbb{Z}} \sum_{i=1}^{I} \overline{q}_i^n(x) u_i(x), \quad x \in \mathbb{R}^n. \tag{23}$$

By arguments similar to those used in the proof of Proposition 3.2, it can be shown that $\widehat{v}$ is continuous on $\mathbb{R}^n$, C^k on $\mathbb{R}^n \backslash \{0\}$, and that for each $i = 1, \dots, I$ and each n in $\mathbb{Z}$, the solutions of $\dot{x} = f_i(x, \widehat{v}(x))$ starting in $\overline{W}_i^{\beta_i^n}$ remain in this set forever. As $\{W_i^{\beta_i^n}\}_{n \in \mathbb{Z}}$ covers $\mathbb{R}^n$ for each $i = 1, \dots, I$, we obtain the following proposition.

PROPOSITION 3.3. *The state feedback law $\widehat{v}$ as given in* (23) *is continuous and almost C^k on $\mathbb{R}^n$ and simultaneously stabilizes (not asymptotically) the family $\{S_i,\ i = 1, \dots, I\}$. Further, the solutions of the system $\dot{x} = f_i(x, \widehat{v}(x))$, starting at x_0, are bounded for each x_0 in $\mathbb{R}^n$ and each $i = 1, \dots, I$.*

In order to obtain three sequences $\{\gamma_i^n,\ i = 1, \dots, I\}_{n \in \mathbb{Z}}$, $\{\beta_i^n,\ i = 1, \dots, I\}_{n \in \mathbb{Z}}$, and $\{\alpha_i^n,\ i = 1, \dots, I\}_{n \in \mathbb{Z}}$, one actually needs to find a condition on the two positive reals α and β that yields the inclusion $D \cap V_{i-1}^{-1}([0, \alpha)) \supset D \cap V_i^{-1}([0, \beta])$, for each $i = 2, \dots, I$, as well as a condition for the inclusion $D \cap V_I^{-1}([0, \alpha)) \supset D \cap V_1^{-1}([0, \beta])$ to be satisfied. In the case where the system S_i and the feedback law u_i are linear for all $i = 1, \dots, I$, these conditions are well known and we provide, in the following paragraph, a simple procedure that yields the desired sequences of reals. We will then use these sequences in section 4.

3.3. Application to families of linear systems.

We now consider a countable family $\{S_i,\ i \in \mathcal{I}\}$ of LTI systems

$$S_i: \quad \dot{x} \;=\; A_i x \;+\; B_i u, \quad i \in \mathcal{I}, \tag{24}$$

where the state x lies in $\mathbb{R}^n$, the input u is in $\mathbb{R}^m$, and for each $i \in \mathcal{I}$, the matrices A_i and B_i belong to $\mathbb{R}^{n \times n}$ and $\mathbb{R}^{n \times m}$, respectively. Finally, for each $i \in \mathcal{I}$, we assume that there exists K_i in $\mathbb{R}^{n \times m}$ such that the linear feedback law $u_i : \mathbb{R}^n \to \mathbb{R}^m$ given by $u_i(x) = K_i x,\ x \in \mathbb{R}^n$, asymptotically stabilizes S_i.

For the sake of simplicity, we assume that $\mathcal{I}$ is finite and equal to $\{1, \dots, I\}$ where $I \geq 2$ is an integer. However, the construction that we provide below can be transposed easily to the case of countably infinite families of LTI systems (see [5, Chapter 3] for further details).

We now let $V_i : \mathbb{R}^n \to [0, \infty)$ be a Lyapunov function for the system $\dot{x} = f_i(x, u_i(x))$ given by $V_i(x) = x^t P_i x,\ x \in \mathbb{R}^n$, where P_i is a positive definite matrix. As u_i globally asymptotically stabilizes the system S_i for each $i = 1, \dots, I$, a simultaneous stabilizer for the family $\{S_i, i = 1, \dots, I\}$ may be given by the formula (23). In this particular case, the sequences of real $\{\alpha_i^n,\ i = 1, \dots, I\}_{n \in \mathbb{Z}}$, $\{\beta_i^n,\ i = 1, \dots, I\}_{n \in \mathbb{Z}}$, and $\{\gamma_i^n,\ i = 1, \dots, I\}_{n \in \mathbb{Z}}$, used in the expression of the mappings $\overline{q}_i^n$, can be defined as follows: We let π_1 be in $(0, \lambda_{min}(P_1)/\lambda_{max}(P_I))$ and π_i be in $(0, \lambda_{min}(P_i)/\lambda_{max}(P_{i-1}))$ for each $i = 2, \dots, I$. Further, we let θ_i in $(0, 1)$ for each $i = 1, \dots, I$ be such that

$$(\pi_1 \cdots \pi_I)\,(\theta_1^2 \cdots \theta_I^2) \;<\; 1, \tag{25}$$

and we let $\widehat{\gamma}_1^0$ be an arbitrary positive real. Finally, we let the sequences of positive reals $\{\alpha_i^n,\ i = 1, \dots, I\}_{n \in \mathbb{Z}}$, $\{\beta_i^n,\ i = 1, \dots, I\}_{n \in \mathbb{Z}}$, and $\{\gamma_i^n,\ i = 1, \dots, I\}_{n \in \mathbb{Z}}$ be defined by setting, on one hand,

$$\gamma_1^0 = \widehat{\gamma}_1^0, \quad \beta_i^n \;=\; \theta_i \gamma_i^n, \quad \alpha_i^n \;=\; \theta_i \beta_i^n, \quad i = 1, \dots, I, \quad n = 0, 1, \dots \tag{26}$$

with

$$\gamma_1^{n+1} \;=\; \pi_1 \alpha_I^n \quad \text{and} \quad \gamma_i^n = \pi_i \alpha_{i-1}^n, \quad i = 2, \dots, I, \quad n = 0, 1, \dots, \tag{27}$$

and, on the other hand,

$$(28)\qquad \alpha_I^n = \frac{\gamma_1^{n+1}}{\pi_1} \quad \text{and} \quad \alpha_i^n = \frac{\gamma_{i+1}^n}{\pi_{i+1}}, \quad i = I-1, \dots, 1, \quad n = -1, -2, \dots$$

with

$$(29)\qquad \beta_i^n = \frac{\alpha_i^n}{\theta_i} \quad \text{and} \quad \gamma_i^n = \frac{\beta_i^n}{\theta_i}, \quad i = I, \dots, 1, \quad n = -1, -2, \dots.$$

It is plain that the obtained sequences of reals converge to 0 (resp., $+\infty$) as n tends to $+\infty$ (resp., $-\infty$), and satisfy (22) for each n in $\mathbb{Z}$. Thus, these sequences can be used in the expression of the mappings q_i^n as given in (17) in order to produce a feedback law $\widehat{v}$ [given by (23)] which simultaneously stabilizes $\{S_i,\ i = 1, \dots, I\}$.

4. Simultaneous asymptotic stabilization of LTI systems. The results of the previous section are primarily derived using a stability criterion based on the positive invariance of sets. In order to construct feedback laws which achieve asymptotic stability, we now enrich this criterion and introduce time-varying feedback laws. We obtain an invariance criterion for time-varying sets that we use for the simultaneous *asymptotic* stabilization of families of LTI systems.

Throughout this section, we consider a countable family $\{S_i,\ i \in \mathcal{I}\}$ of LTI systems as defined in (24). Further, for each i in $\mathcal{I}$, we assume that there exists K_i in $\mathbb{R}^{n\times m}$ such that the linear feedback law $u_i : \mathbb{R}^n \to \mathbb{R}^m$ given by $u_i(x) = K_i x$ asymptotically stabilizes S_i.

Further, for each i in $\mathcal{I}$, we let $V_i : \mathbb{R}^n \to [0, \infty)$ denote a Lyapunov function for the system $\dot{x} = (A_i + B_i K_i)\, x$ given by $V_i(x) = x^t\, P_i\, x$ for each x in $\mathbb{R}^n$, where P_i is a positive definite matrix in $\mathbb{R}^{n\times n}$. Finally, we let Q_i be the positive definite matrix defined by

$$\nabla V_i(x)\, ((A_i + B_i K_i)\, x) \;=\; -\, x^t Q_i x, \quad x \in \mathbb{R}^n, \quad i \in \mathcal{I}.$$

The main result of this section is the following theorem.

THEOREM 4.1. *Assume that, in addition to the enforced assumptions, the set $\mathcal{I}$ is finite with $\mathcal{I} = \{1, \dots, I\}$ and $I \geq 2$. Then, there exists a time-varying state feedback law $v : [0,\infty) \times \mathbb{R}^n \to \mathbb{R}^m$, continuous on $[0,\infty) \times \mathbb{R}^n$ and C^∞ on $[0,\infty) \times (\mathbb{R}^n \backslash \{0\})$, which simultaneously globally exponentially stabilizes the family $\{S_i,\ i = 1, \dots, I\}$.*

The general lines of the proof of this theorem are as follows: We introduce a sequence $\{b_i^n(\cdot),\ i = 1, \dots, I\}_{n\in\mathbb{Z}}$ of mappings defined from $[0,\infty)$ into $(0,+\infty)$, decreasing to 0 as t tends to $+\infty$, and such that for each $t \geq 0$, the sequence of neighborhoods $\{V_i^{-1}([0, b_i^n(t)))\}_{n\in\mathbb{Z}}$ is a topological base at the origin. We then design a time-varying feedback law $v(t,x)$ such that for each $t \geq 0$, each $i = 1, \dots, I$, and each n in $\mathbb{Z}$, we have $v(t,x) = u_i(x)$ for all x in $V_i^{-1}(b_i^n(t))$. Finally, we show that for each $t_0 \geq 0$, each $i = 1, \dots, I$, and each n in $\mathbb{Z}$, each solution of the system $\dot{x} = f_i(x, v(t,x))$, which starts in the set $V_i^{-1}([0, b_i^n(t_0)])$ at time t_0, remains in the set $V_i^{-1}([0, b_i^n(t)])$ for all $t \geq t_0$. For each $i = 1, \dots, I$, we conclude that v asymptotically stabilizes the system S_i, upon noting that the mapping b_i^n converges to 0 as t tends to $+\infty$ for each n in $\mathbb{Z}$.

Proof. For each $i = 1, \dots, I$ and each $\beta > 0$, we let W_i^β denote the set $W_i^\beta \triangleq V_i^{-1}(\,[0, \beta)\,)$. We then define three sequences of positive reals $\{\alpha_i^n,\ i = 1, \dots, I\}_{n\in\mathbb{Z}}$, $\{\beta_i^n,\ i = 1, \dots, I\}_{n\in\mathbb{Z}}$, and $\{\gamma_i^n,\ i = 1, \dots, I\}_{n\in\mathbb{Z}}$ exactly as we did in subsection 3.3. In other words, we let π_1 be in $(0, \lambda_{min}(P_1)/\lambda_{max}(P_I))$ and π_i be in $(0, \lambda_{min}(P_i)/\lambda_{max}(P_{i-1}))$ for each $i = 2, \dots, I$. Further, we let $\widehat{\gamma}_1^0$ be an arbitrary positive real, and we let θ_i in $(0,1)$ for each $i = 1,\ldots,I$ be such that (25) holds. Finally, we apply the formulas (26), (27), (28), and (29).

Construction of the simultaneous stabilizer. We now seek a C^1 mapping $h : [0,\infty) \to (0,\infty)$ such that the mapping $b_i^n : [0,\infty) \to (0,\infty)$, given by

$$b_i^n(t) \;=\; \beta_i^n\, h(t), \quad t \geq 0,$$

satisfies

$$\nabla V_i(x)\,(\,A_i x + B_i u_i(x)\,) \;<\; \dot{b}_i^n(t), \quad x \in V_i^{-1}(b_i^n(t)), \quad t \geq 0, \tag{30}$$

or, equivalently,

$$-x^t Q_i\, x \;<\; \dot{b}_i^n(t), \quad x \in V_i^{-1}(b_i^n(t)), \quad t \geq 0, \tag{31}$$

for each $i = 1,\ldots,I$ and each n in $\mathbb{Z}$. In what follows, we fix $i = 1,\ldots,I$, n in $\mathbb{Z}$, and $t \geq 0$. Let x be such that $x^t P_i\, x = b_i^n(t)$. Then, by elementary linear algebra, we get $-x^t x \leq -b_i^n(t)/\lambda_{max}(P_i)$, and because we also have $-x^t Q_i\, x \leq -\lambda_{min}(Q_i)\, x^t x$, inequality (31) will be satisfied if

$$-\frac{\lambda_{min}(Q_i)}{\lambda_{max}(P_i)}\, b_i^n(t) \;<\; \dot{b}_i^n(t). \tag{32}$$

Because we require that $h(t) > 0$ and $b_i^n(t) = \beta_i^n h(t)$, inequality (32) will hold if

$$\frac{\dot{h}(t)}{h(t)} \;>\; -\,\frac{\lambda_{min}(Q_i)}{\lambda_{max}(P_i)}. \tag{33}$$

We now set $\rho \triangleq \min_{i=1,\ldots,I}\,(\lambda_{min}(Q_i)/\lambda_{max}(P_i))$ and we deduce from (33) that the desired assertion (30) will be satisfied for each $i = 1,\ldots,I$ and each n in $\mathbb{Z}$, if

$$\frac{\dot{h}(t)}{h(t)} \;>\; -\,\rho, \quad t \geq 0. \tag{34}$$

Let η be a fixed constant in $(1,\infty)$, and let the mapping $h : [0,\infty) \to (0,\infty)$ be given by

$$h(t) \;=\; e^{-\frac{\rho}{\eta}t}, \quad t \geq 0.$$

It is plain that h satisfies (34) so that for each $i = 1,\ldots,I$ and each n in $\mathbb{Z}$, the mapping $b_i^n : [0,\infty) \to (0,\infty)$ given by

$$b_i^n(t) \;=\; \beta_i^n h(t) \;=\; \beta_i^n e^{-\frac{\rho}{\eta}t}, \quad t \geq 0,$$

satisfies the desired assertion (30).

Next, for each $i = 1,\ldots,I$ and each n in $\mathbb{Z}$, we define the mappings $a_i^n, c_i^n : [0,\infty) \to (0,\infty)$ by setting

$$c_i^n(t) \;=\; \gamma_i^n\, h(t) \quad \text{and} \quad a_i^n(t) \;=\; \alpha_i^n\, h(t), \quad t \geq 0.$$

Because we have

$$b_i^n(t) \;=\; \theta_i c_i^n(t) \quad \text{with} \quad a_i^n(t) \;=\; \theta_i b_i^n(t), \quad t \geq 0, \quad i = 1,\ldots,I, \quad n \in \mathbb{Z},$$

and

$$c_1^{n+1}(t) \;=\; \pi_1 a_I^n(t) \quad \text{with} \quad c_i^n(t) \;=\; \pi_i a_{i-1}^n(t), \quad t \geq 0, \quad i = 2, \dots, I, \quad n \in \mathbb{Z},$$

it follows from the definition of θ_i and π_i for each $i = 1, \dots, I$ that for each $t \geq 0$, we have a two-sided sequence of neighborhoods

$$\begin{array}{cccccccccccc}
\vdots & & & & & & & & & & & \vdots \\
W_1^{c_1^{-1}(t)} & \supset & & & \cdots & & & & & & W_I^{a_I^{-1}(t)} & \supset \\
W_1^{c_1^{0}(t)} & \supset & W_1^{b_1^{0}(t)} & \supset & W_1^{a_1^{0}(t)} & \supset & W_2^{c_2^{0}(t)} & \supset & \cdots & \supset & W_I^{a_I^{0}(t)} & \supset \\
W_1^{c_1^{1}(t)} & \supset & W_1^{b_1^{1}(t)} & \supset & W_1^{a_1^{1}(t)} & \supset & W_2^{c_2^{1}(t)} & \supset & \cdots & \supset & W_I^{a_I^{1}(t)} & \supset \\
\vdots & & & & & & \vdots & & & & \vdots &
\end{array} \tag{35}$$

such that each neighborhood contains the closure of the neighborhood that follows. Next, for each $i = 1, \dots, I$ and each n in $\mathbb{Z}$, we define the mapping $q_i^n : [0, \infty) \times \mathbb{R}^n \to [0, 1]$ by setting

$$q_i^n(t, x) = \begin{cases} e^{\frac{(V_i(x) - b_i^n(t))^2}{\left(V_i(x) - b_i^n(t)\right)^2 - \left(b_i^n(t) - a_i^n(t)\right)^2}} & \text{if } V_i(x) \in (a_i^n(t), b_i^n(t)], \\ e^{\frac{\left(V_i(x) - b_i^n(t)\right)^2}{\left(V_i(x) - b_i^n(t)\right)^2 - (c_i^n(t) - b_i^n(t))^2}} & \text{if } V_i(x) \in (b_i^n(t), c_i^n(t)), \\ 0 & \text{otherwise} \end{cases} \tag{36}$$

for each (t, x) in $[0, \infty) \times \mathbb{R}^n$, and we let the mapping $v : [0, \infty) \times \mathbb{R}^n \to \mathbb{R}^m$ be given by

$$v(t, x) = \sum_{i=1}^{I} \sum_{n \in \mathbb{Z}} u_i(x) q_i^n(t, x), \quad (t, x) \in [0, \infty) \times \mathbb{R}^n.$$

The feedback law v is continuous on $[0, \infty) \times \mathbb{R}^n$ and C^∞ on $[0, \infty) \times (\mathbb{R}^n \backslash \{0\})$. Let (t, x) be in $[0, \infty) \times \mathbb{R}^n \backslash \{0\}$. It is easily checked from (35) combined with the continuity of the mappings V_j, a_j^m, and c_j^m for each (j, m) in $\{1, \dots, I\} \times \mathbb{Z}$ that there exist a neighborhood U of (t, x) in $[0, \infty) \times (\mathbb{R}^n \backslash \{0\})$ and a pair (i, n) in $\{1, \dots, I\} \times \mathbb{Z}$ such that

$$v(\tau, y) \;=\; u_i(y)\, q_i^n(\tau, y), \quad (\tau, y) \in U. \tag{37}$$

Because q_i^n is C^∞ on $[0, \infty) \times \mathbb{R}^n$ [5, Lemma B.6, p. 159] and u_i is C^∞ on $\mathbb{R}^n$ for each $i = 1, \dots, I$ and each n in $\mathbb{Z}$, the equality (37) implies that v is C^∞ on $[0, \infty) \times \mathbb{R}^n \backslash \{0\}$. Further, because the mappings q_i^n take values in $[0, 1]$, the equality (37) yields

$$\|v(t, x)\| \;\leq\; \max\left(\|u_1(x)\|, \dots, \|u_I(x)\|\right), \quad (t, x) \in [0, \infty) \times \mathbb{R}^n,$$

and continuity of u_i for each $i = 1, \dots, I$ combined with the fact that $v(t, 0) = 0,\ t \geq 0$ implies that v is continuous at each point $(t, 0), t \geq 0$. Therefore, v is continuous on $[0, \infty) \times \mathbb{R}^n$.

Global exponential stability. Throughout the rest of the proof, we fix $i = 1, \dots, I$. From the definition of v, it is not hard to see that

$$v(t, x) = u_i(x), \quad t \geq 0, \quad x \in V_i^{-1}(\, b_i^n(t)\,), \quad n \in \mathbb{Z}.$$

Therefore, from (30) we get

$$\nabla V_i(x)\,(A_i\,x + B_i v(t,x)) \;<\; \dot{b}_i^n(t), \quad x \in V_i^{-1}(b_i^n(t)), \quad t \geq 0, \quad n \in \mathbb{Z},$$

and because we have $\partial W_i^\beta = V_i^{-1}(\beta)$ and $\overline{W}_i^\beta = V_i^{-1}([0,\beta])$ for each $\beta > 0$, Lemma 8.1 implies that for each $t_0 \geq 0$ and each n in $\mathbb{Z}$, the solution $x(\cdot, x_0, t_0)$ of $\dot{x} = A_i x + B_i v(t,x)$ starting from x_0 at time t_0 satisfies

$$V_i(x(t,x_0,t_0)) \;\leq\; b_i^n(t), \quad t \geq t_0, \quad x_0 \in \overline{W}_i^{b_i^n(t_0)}. \tag{38}$$

From the definition of the sequence $\{\beta_i^n\}_{n\in\mathbb{Z}}$, it is easily seen that we have

$$\beta_i^{n+k} \;=\; (\pi_1 \cdots \pi_I\, \theta_1^2 \cdots \theta_I^2)^k \beta_i^n, \quad n \in \mathbb{Z}, \quad k = 0, 1, \ldots.$$

Upon setting $T \triangleq -\frac{\eta}{\rho}\,\log(\pi_1 \ldots \pi_I\, \theta_1^2 \ldots \theta_I^2)$, the previous equality translates to

$$\beta_i^n e^{-\frac{\rho}{\eta}kT} \;=\; \beta_i^{n+k}, \quad n \in \mathbb{Z}, \quad k = 0, 1, \ldots, \tag{39}$$

which in turn easily yields

$$b_i^n(t + kT) \;=\; b_i^{n+k}(t), \quad t \geq 0, \quad n \in \mathbb{Z}, \quad k = 0, 1, \ldots. \tag{40}$$

Let x_0 be in $\mathbb{R}^n$. As the sequence $\{b_i^n(0) = \beta_i^n\}_{n\in\mathbb{Z}}$ is strictly decreasing and converges to 0 and $+\infty$ as n tends to $+\infty$ and $-\infty$, resp., there exists an integer $\bar{n}$ such that

$$b_i^{\bar{n}+1}(0) \;<\; V_i(x_0) \;\leq\; b_i^{\bar{n}}(0). \tag{41}$$

Fix $t_0 \geq 0$, and let the integer k and t_0' in $[0,T)$ be such that $t_0 = kT + t_0'$. By combining (40) with the fact that the mapping b_i^n is decreasing for each n in $\mathbb{Z}$, we get

$$b_i^{\bar{n}}(0) \;=\; b_i^{\bar{n}-k-1}((k+1)T) \;\leq\; b_i^{\bar{n}-k-1}(t_0),$$

so that (41) yields $x_0 \in V_i^{-1}(\,[0, b_i^{\bar{n}-k-1}(t_0)]\,)$. Thus, the inequality (38) implies that

$$V_i(x(t,x_0,t_0)) \;\leq\; b_i^{\bar{n}-k-1}(t), \quad t \geq t_0,$$

and from the expression of $b_i^{\bar{n}-k-1}(t)$ we obtain that

$$V_i(x(t,x_0,t_0)) \;\leq\; \beta_i^{\bar{n}-k-1}\, e^{-\frac{\rho}{\eta}(t-t_0+t_0'+kT)}, \quad t \geq t_0. \tag{42}$$

Next, by combining (39) with the fact that $t_0' \geq 0$, we obtain from (42) that

$$V_i(x(t,x_0,t_0)) \;\leq\; \beta_i^{\bar{n}-1}\, e^{-\frac{\rho}{\eta}(t-t_0+t_0')} \;\leq\; \beta_i^{\bar{n}-1}\, e^{-\frac{\rho}{\eta}(t-t_0)}, \quad t \geq t_0. \tag{43}$$

The identity (39) yields $\beta_i^{\bar{n}-1} = e^{\frac{2\rho}{\eta}T}\beta_i^{\bar{n}+1}$, so that the inequality $\beta_i^{\bar{n}-1} < e^{\frac{2\rho}{\eta}T}V_i(x_0)$ follows from (41). Thus, (43) implies that

$$\sqrt{V_i(x(t,x_0,t_0))} \;\leq\; e^{\frac{\rho}{\eta}T}\sqrt{V_i(x_0)}\; e^{-\frac{\rho}{2\eta}(t-t_0)}, \quad t \geq t_0.$$

Because $e^{\frac{\rho}{\eta}T}$ is a constant and the mapping $x \mapsto \sqrt{V_i(x)}$ is a norm on $\mathbb{R}^n$, we obtain from the equivalence of all norms on $\mathbb{R}^n$ that v globally exponentially stabilizes S_i

with the rate of convergence $\frac{\rho}{2\eta}$. The proof of the theorem is complete upon noting that the previous argument holds for each $i = 1, \ldots, I$. $\square$

In case the set $\mathcal{I}$ is countably infinite, we obtain the following sufficient conditions for simultaneous asymptotic stabilizability.

THEOREM 4.2. *Assume that, in addition to the enforced assumptions, the set $\mathcal{I}$ is countably infinite and equal to $\{1, 2, \ldots\}$ and that the following statements hold.*

(i) *There exists a positive real M such that $\|K_i\| \leq M,\ i = 1, 2, \ldots.$*

(ii) *The real $\rho \triangleq \inf_{i=1,2,\ldots}(\frac{\lambda_{min}(Q_i)}{\lambda_{max}(P_i)})$ is strictly positive.*

Then, there exists a time-varying state feedback law $v : [0, \infty) \times \mathbb{R}^n \to \mathbb{R}^m$, continuous on $[0, \infty) \times \mathbb{R}^n$ and C^∞ on $[0, \infty) \times (\mathbb{R}^n \backslash \{0\})$, which simultaneously globally asymptotically stabilizes the family $\{S_i,\ i = 1, 2, \ldots\}$.

We omit the proof of this theorem as it is based on arguments similar to those used for Theorem 4.1 (we refer the interested reader to [5, p. 65] or [7] for a proof).

5. Simultaneous asymptotic stabilization of nonlinear systems.

Because the asymptotic criterion introduced in the previous section is nonlinear we are now able to use it for the simultaneous stabilization of nonlinear systems. We actually obtain the following sufficient conditions for simultaneous stabilizability.

THEOREM 5.1. *Let $k \geq 0$ and $k' \geq 1$ be two integers and set $k'' \triangleq \min(k, k')$. Let D be a neighborhood of the origin, and assume that there exists a continuous and almost C^k state feedback law $u_i : D \to \mathbb{R}^m$ which locally asymptotically stabilizes S_i for each $i = 1, \ldots, I$. Further, assume that for each $i = 1, \ldots, I$, the mapping $f_i(\cdot, u_i(\cdot))$ is continuous on D and let $V_i : \overline{D} \to [0, \infty)$ be a $C^{k'}$ Lyapunov function, satisfying*

$$\nabla V_i(x)\, f_i(x, u_i(x)) \ < \ 0, \quad x \in D \backslash \{0\}.$$

(a) *Assume that there exists a sequence $\{b_i^n,\ i = 1, \ldots, I\}_{n=1}^\infty$ of $C^{k'}$ mappings $b_i^n : [0, \infty) \to (0, \infty)$, such that the following assertions hold.*

(i) $\sup_{t \geq t_0} b_i^n(t) \ \to\ 0$ *as* $n \to \infty, \quad i = 1, \ldots, I, \quad t_0 \geq 0.$

(ii) $b_i^n(t) < \inf_{x \in \partial D} V_i(x), \quad t \geq 0, \quad i = 1, \ldots, I, \quad n = 1, 2, \ldots.$

(iii) $b_i^n(t) \ \to\ 0$ *as* $t \to \infty, \quad i = 1, \ldots, I, \quad n = 1, 2, \ldots.$

(iv) *For each $n = 1, 2, \ldots,$ we have*

$$D \cap V_I^{-1}(\,[0, b_I^n(t))\,) \ \supset\ D \cap V_1^{-1}(\,[0, b_1^{n+1}(t)]\,), \quad t \geq 0, \tag{44}$$

and for each $n = 1, 2, \ldots$ and each $i = 1, \ldots, I - 1$ we have

$$D \cap V_i^{-1}(\,[0, b_i^n(t)]\,) \ \supset\ D \cap V_{i+1}^{-1}(\,[0, b_{i+1}^n(t))\,), \quad t \geq 0. \tag{45}$$

(v) *For each $n = 1, 2, \ldots$ and each $i = 1, \ldots, I$, we have*

$$\nabla V_i(x)\, f_i(x, u_i(x)) \ < \ \dot{b}_i^n(t), \quad x \in D \cap V_i^{-1}(b_i^n(t)), \quad t \geq 0.$$

Then, there exists a time-varying state feedback law $v : [0, \infty) \times D \to \mathbb{R}^m$, continuous on $[0, \infty) \times D$ and $C^{k''}$ on $[0, \infty) \times (D \backslash \{0\})$, which simultaneously locally asymptotically stabilizes the family $\{S_i,\ i = 1, \ldots, I\}$.

(b) *Now, assume that for each $i = 1, \ldots, I$, the feedback law u_i globally asymptotically stabilizes S_i and that V_i is a radially unbounded Lyapunov function, so that we may set $D \triangleq \mathbb{R}^n$. Further, assume that there exists a sequence $\{b_i^n,\ i = 1, \ldots, I\}_{n \in \mathbb{Z}}$*

of $C^{k'}$ *mappings* $b_i^n : [0,\infty) \to (0,\infty)$*, such that on one hand, assertion* (i) *holds and, on the other hand, assertions* (ii), (iii), (iv), and (v) *hold for each* n *in* $\mathbb{Z}$*. Finally, assume that we have*

$$b_i^n(t_0) \to +\infty \quad as \quad n \to -\infty, \quad i = 1, \ldots, I, \quad t_0 \geq 0. \tag{46}$$

Then there exists a time-varying state feedback law $v : [0,\infty) \times \mathbb{R}^n \to \mathbb{R}^m$*, continuous on* $[0,\infty) \times \mathbb{R}^n$ *and* $C^{k''}$ *on* $[0,\infty) \times (\mathbb{R}^n \backslash \{0\})$*, which simultaneously globally asymptotically stabilizes the family* $\{S_i,\ i = 1, \ldots, I\}$*.*

Proof. As the proof of assumption (b) is similar to that of assumption (a), we prove only (a) (see [5, p. 83] for further details). Further, because the main lines of the proof of assumption (a) are similar to those of the proof of Theorem 4.1, we will merely outline the main differences between these two proofs. For each $i = 1, \ldots, I$ and each $\beta > 0$, we set $W_i^\beta \triangleq \{x \in D : V_i(x) < \beta\}$.

We first construct two sequences $\{a_i^n,\ i = 1, \ldots, I\}_{n=1}^\infty$ and $\{c_i^n,\ i = 1, \ldots, I\}_{n=1}^\infty$ of mappings $a_i^n, b_i^n : [0,\infty) \to (0,\infty)$, satisfying for each $n = 1, 2, \ldots$

$$\inf_{x \in \partial D} V_i(x) \ > \ c_i^n(t) \ > \ b_i^n(t) \ > \ a_i^n(t) \ > \ 0, \quad t \geq 0, \quad i = 1, \ldots, I, \tag{47}$$

with

$$W_I^{a_I^n(t)} \supset \overline{W}_1^{c_1^{n+1}(t)}, \qquad \text{and} \qquad W_i^{a_i^n(t)} \supset \overline{W}_{i+1}^{c_{i+1}^n(t)}, \quad i = 1, \ldots, I-1, \tag{48}$$

for each $t \geq 0$ as follows: We define $c_1^1 : [0,\infty) \to (0,\infty)$ by setting

$$c_1^1(t) \ = \ \frac{b_1^1(t) + \inf\limits_{x \in \partial D} V_1(x)}{2}, \quad t \geq 0.$$

Then, for each $n = 1, 2, \ldots$ and each $i = 1, \ldots, I-1$, we define two C^∞ mappings $a_i^n, c_{i+1}^n : [0,\infty) \to (0,\infty)$ from b_i^n and b_{i+1}^n by applying Lemma 8.2 with $b_1 = b_i^n$, $b_2 = b_{i+1}^n$, and (45). Finally, for each $n = 1, 2, \ldots$, we define two C^∞ mappings $a_I^n, c_1^{n+1} : [0,\infty) \to (0,\infty)$ by applying Lemma 8.2 with $b_1 = b_I^n$, $b_2 = b_1^{n+1}$, and the inequality (44).

Next, for each $n = 1, 2, \ldots$ and each $i = 1, \ldots, I$, we define the mapping $q_i^n : [0,\infty) \times D \to [0,1]$ by applying the formula (36) for each (t,x) in $[0,\infty) \times D$ (with V_i, a_i^n, b_i^n, c_i^n as defined here). Finally, we define the feedback law $v : [0,\infty) \times D \to \mathbb{R}^m$ by setting

$$v(t,x) = \sum_{n=1}^{\infty} \sum_{i=1}^{I} u_i(x) q_i^n(t,x), \quad (t,x) \in [0,\infty) \times D. \tag{49}$$

Using an argument similar to that used in the proof of Theorem 4.1, it is not hard to check that v is continuous on $[0,\infty) \times D$ and $C^{k''}$ on $[0,\infty) \times (D \backslash \{0\})$.

We now fix $i = 1, \ldots, I$ and $n = 1, 2, \ldots$. Because $\partial W_i^{b_i^n(t)} = D \cap V_i^{-1}(b_i^n(t))$ and $q_i^n(t,x) = 1$ for each x in $V_i^{-1}(b_i^n(t))$ and each $t \geq 0$, we obtain that $v(t,x) = u_i(x)$ for each x in $\partial W_i^{b_i^n(t)}$ and each $t \geq 0$. It follows from the assumption v that

$$\nabla V_i(x)\, f_i(x, v(t,x)) \ = \ \nabla V_i(x)\, f_i(x, u_i(x)) \ < \ \dot{b}_i^n(t), \quad x \in \partial W_i^{b_i^n(t)}, \quad t \geq 0.$$

Thus, by an argument similar to that used in the proof of Theorem 4.2 to prove asymptotic stability, it can be shown that v simultaneously asymptotically stabilizes the family $\{S_i,\ i = 1, \ldots, I\}$. □

We are now going to use these sufficient conditions for the simultaneous stabilization of certain homogeneous systems.

6. Simultaneous asymptotic stabilization of homogeneous systems. In this section, we consider a pair of control systems,

$$S_1 : \ \dot{x} = f(x) - g(x)u \qquad \text{and} \qquad S_2 : \ \dot{x} = f(x) + g(x)u, \tag{50}$$

where the state x lies in $\mathbb{R}^n$ and the input u lies in $\mathbb{R}^m$. We assume that the mappings $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ are continuous with $f(0) = 0$ and that there exists a continuous state feedback law $u : \mathbb{R}^n \to \mathbb{R}^m$ which locally asymptotically stabilizes S_2, so that $-u$ locally asymptotically stabilizes S_1. We define the mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ by setting

$$F(x) \ = \ f(x) \ + \ g(x)\, u(x), \quad x \in \mathbb{R}^n, \tag{51}$$

and we assume that F is homogeneous, i.e., there exists s in $\mathbb{R}$ and $(r_1, \dots, r_n)$ in $(0, +\infty)^n$ such that for each x in $\mathbb{R}^n \backslash \{0\}$ and each $\lambda > 0$, we have

$$F(\lambda^{r_1} x_1, \dots, \lambda^{r_n} x_n) \ = \ \begin{pmatrix} \lambda^{s+r_1} F_1(x) \\ \vdots \\ \lambda^{s+r_n} F_n(x) \end{pmatrix}, \tag{52}$$

where F_i is the ith coordinate mapping of F for each $i = 1, \dots, n$ (note that this assumption clearly holds if the control system S_1 and the feedback law u are homogeneous). As we shall see in the following proposition, it turns out that there *always* exists a time-varying state feedback law which simultaneously stabilizes S_1 and S_2.

PROPOSITION 6.1. *Let the systems S_1 and S_2 and the mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ satisfy the assumptions* (50), (51), *and* (52). *Then, the following hold.*

(i) *If $s = 0$, then there exists a continuous time-varying state feedback law $v : [0, \infty) \times \mathbb{R}^n \to \mathbb{R}^m$ which simultaneously globally uniformly asymptotically stabilizes S_1 and S_2, with uniform exponential convergence of the corresponding closed-loop systems* (*according to Definition* 2.1 (iv)).

(ii) *If $s > 0$, then there exists a continuous time-varying state feedback law $v : [0, \infty) \times \mathbb{R}^n \to \mathbb{R}^m$ which simultaneously globally asymptotically stabilizes S_1 and S_2.*

(iii) *If $s < 0$, then there exists a continuous time-varying state feedback law $v : [0, \infty) \times \mathbb{R}^n \to \mathbb{R}^m$ which simultaneously locally asymptotically stabilizes S_1 and S_2.*

Proof. The main idea of the proof is to construct some sequences of mappings b_i^n satisfying the assumptions of Theorem 5.1. To this end, we first define a mapping $h_\beta : [0, \infty) \to (0, \infty)$ for each β in some subset of $(0, \infty)$ as follows.

We let $p = 1$ and we choose k satisfying

$$k \ \geq \ \max_{i=1,\dots,n} (r_i) \qquad \text{and} \qquad 1 + \frac{s}{k} > 0.$$

By Theorem 2 in Rosier [12], there exists a C^1 radially unbounded homogeneous Lyapunov function $V : \mathbb{R}^n \to [0, \infty)$ such that

$$\nabla V(x)\, F(x) \ < \ 0, \quad x \in \mathbb{R}^n \backslash \{0\}, \tag{53}$$

and

$$V(\lambda^{r_1} x_1, \dots, \lambda^{r_n} x_n) = \lambda^k V(x), \quad x \in \mathbb{R}^n \backslash \{0\}, \tag{54}$$

for each $\lambda > 0$. We stress that (53) implies that the system $\dot{x} = F(x)$ is globally asymptotically stable. Further, as mentioned in [12, p. 470], the identities (52) and (54) yield

$$\nabla V(\lambda^{r_1}x_1,\dots,\lambda^{r_n}x_n)\,F(\lambda^{r_1}x_1,\dots,\lambda^{r_n}x_n) \;=\; \lambda^{s+k}\,\nabla V(x)\,F(x)$$

for all x in $\mathbb{R}^n\backslash\{0\}$ and all $\lambda > 0$. Furthermore, from the homogeneity of V and F it can be checked [5, Lemma 5.1, p. 87] that there exists $\theta > 0$ such that we have

$$\nabla V(x)\,F(x) \;<\; -\,\theta\rho^{1+\frac{s}{k}}, \quad x \in V^{-1}(\rho), \quad \rho > 0. \tag{55}$$

If $s = 0$, then for each $\beta > 0$ we define $h_\beta : [0,\infty) \to (0,\infty)$ by setting

$$h_\beta(t) \;=\; \beta e^{-\theta t}, \quad t \ge 0. \tag{56}$$

If $s > 0$, then for each $\beta > 0$ we let $h_\beta : [0,\infty) \to (0,\infty)$ be given by

$$h_\beta(t) \;=\; \frac{1}{\left(\beta^{-\frac{s}{k}} + \dfrac{s}{k}\theta t\right)^{\frac{k}{s}}}, \quad t \ge 0. \tag{57}$$

Finally, if $s < 0$, we set

$$\bar{\beta} \;\triangleq\; \left[\theta\left(\frac{-s}{2k}\right)^{\frac{1}{2}} e^{\frac{1}{2}}\right]^{-\frac{k}{s}} \tag{58}$$

and for each β in $(0,\bar{\beta}]$ we define $h_\beta : [0,\infty) \to (0,\infty)$ by setting

$$h_\beta(t) \;=\; \begin{cases} \beta, & t \le \dfrac{\beta^{\frac{-s}{k}}}{\theta}, \\ \beta e^{-\left(t - \frac{\beta^{\frac{-s}{k}}}{\theta}\right)^2}, & t > \dfrac{\beta^{\frac{-s}{k}}}{\theta}. \end{cases} \tag{59}$$

It is not hard to check [5, Lemma B.13, p. 164] that for each β in $(0,\bar{\beta}]$ (resp., $\beta > 0$) the mapping h_β as defined in (59) (resp., (56) or (57)), satisfies $h_\beta(0) = \beta$ together with

$$-\theta h_\beta(t)^{1+\frac{s}{k}} \;\le\; \dot{h}_\beta(t), \quad t \ge 0. \tag{60}$$

If $s < 0$ (resp., $s \ge 0$), we fix β in $(0,\bar{\beta}]$ (resp., in $(0,\infty)$). Because the mapping h_β given by the formula (59) (resp., (56) or (57)) satisfies the inequality (60), it follows from (55) that

$$\nabla V(x)\,F(x) \;<\; \dot{h}_\beta(t), \quad x \in V^{-1}(h_\beta(t)), \quad t \ge 0. \tag{61}$$

Furthermore, it is plain that for each β in $(0,\bar{\beta}]$ (resp., in $(0,\infty)$), the mapping h_β is nonincreasing and converges to 0 as t tends to $+\infty$. Finally, for each β and each γ in $(0,\bar{\beta}]$ (resp., in $(0,\infty)$) with $\beta < \gamma$, we have $h_\beta(t) < h_\gamma(t)$ for all $t \ge 0$.

Simultaneous asymptotic stabilization in the case $s \ge 0$.

Throughout this paragraph, we assume $s \ge 0$. We fix $\widehat{\beta}$ in $(0,1)$, and we define the sequence of positive reals $\{\beta_i^n,\ i = 1,2\}_{n\in\mathbb{Z}}$ by setting

$$\beta_1^n \;\triangleq\; (\widehat{\beta})^{2n-1} \qquad \text{and} \qquad \beta_2^n \;\triangleq\; (\widehat{\beta})^{2n}, \quad n \in \mathbb{Z}.$$

Next, we choose the sequences of positive reals $\{\alpha_i^n,\ i = 1, 2, \}_{n\in\mathbb{Z}}$ and $\{\beta_i^n,\ i = 1, 2\}_{n\in\mathbb{Z}}$ such that we have

$$\gamma_1^n \ > \ \beta_1^n \ > \ \alpha_1^n \ > \ \gamma_2^n \ > \ \beta_2^n \ > \ \alpha_2^n \ > \ \gamma_1^{n+1}, \quad n \in \mathbb{Z},$$

and we define the mappings $a_i^n,\ b_i^n,\ c_i^n : [0,\infty) \to (0,\infty)$ by setting

$$a_i^n(t) \ = \ h_{\alpha_i^n}(t), \quad b_i^n(t) \ = \ h_{\beta_i^n}(t), \quad c_i^n(t) \ = \ h_{\gamma_i^n}(t)$$

for each $t \geq 0$, with h_β given by (56) if $s = 0$ and (57) if $s > 0$. Because h_β satisfies (61), it follows that

$$\nabla V(x)\, F(x) \ < \ \dot{b_i^n}(t), \quad x \in V^{-1}(b_i^n(t)), \quad t \geq 0, \tag{62}$$

for each $i = 1, 2$ and each n in $\mathbb{Z}$. For each $i = 1, 2$, and each n in $\mathbb{Z}$, we let the mapping $q_i^n : [0,\infty) \times \mathbb{R}^n \to [0, 1]$ be given by the formula (36) for each (t, x) in $[0,\infty) \times \mathbb{R}^n$, after substituting V for V_i, with a_i^n, b_i^n, and c_i^n as given here. It can be checked that the sequence of mappings $\{b_i^n,\ i = 1, 2, \}_{n\in\mathbb{Z}}$ satisfies each one of the assumptions of Theorem 5.1(b), and from the proof of this theorem we see that the mapping $v : [0,\infty) \times \mathbb{R}^n \to \mathbb{R}^m$ given by

$$v(t,x) = -\sum_{n\in\mathbb{Z}} u(x)\, q_1^n(t,x) \ + \ \sum_{n\in\mathbb{Z}} u(x)\, q_2^n(t,x), \quad (t,x) \in [0,\infty) \times \mathbb{R}^n, \tag{63}$$

simultaneously *globally* asymptotically stabilizes S_1 and S_2. In the case $s = 0$ we now show that the feedback law v uniformly stabilizes with exponential convergence both S_1 and S_2.

Exponential convergence in the case $s = 0$. Throughout this paragraph, we assume that $s = 0$, so that we have

$$b_1^n(t) \ = \ (\widehat{\beta})^{2n-1}\, e^{-\theta t}, \quad \text{and} \quad b_2^n(t) \ = \ (\widehat{\beta})^{2n}\, e^{-\theta t}, \quad t \geq 0, \quad n \in \mathbb{Z}.$$

Upon setting $T \triangleq -\frac{1}{\theta}\ln(\beta^2)$, these definitions yield

$$b_i^n(t + kT) \ = \ b_i^{n+k}(t), \quad t \geq 0, \quad i = 1, 2, \quad n \in \mathbb{Z}, \quad k = 0, 1, \ldots. \tag{64}$$

We now fix $i = 1, 2$ and we let x_0 be in $\mathbb{R}^n$. Because the sequence $\{b_i^n(0)\}_{n\in\mathbb{Z}}$ is strictly decreasing and converges to $+\infty$ and 0 as n tends to $-\infty$ and $+\infty$, resp., there exists $\bar{n}(x_0)$ in $\mathbb{Z}$ (for the sake of brevity, we will let $\bar{n}$ denote $\bar{n}(x_0)$) such that

$$b_i^{\bar{n}+1}(0) \ < \ V(x_0) \ \leq \ b_i^{\bar{n}}(0). \tag{65}$$

We fix $t_0 \geq 0$ and let the integer k and the real t_0' in $[0, T)$ be such that $t_0 = kT + t_0'$. By combining (64) with the fact that the mapping b_i^n is decreasing for each n in $\mathbb{Z}$, we get

$$b_i^{\bar{n}}(0) \ = \ b_i^{\bar{n}-k-1}((k+1)T) \ \leq \ b_i^{\bar{n}-k-1}(t_0),$$

so that (65) yields $V(x_0) \leq b_i^{\bar{n}-k-1}(t_0)$. Therefore, Lemma 8.1 combined with (62) implies that

$$V(x(t, x_0, t_0)) \ \leq \ b_i^{\bar{n}-k-1}(t), \quad t \geq t_0. \tag{66}$$

Next, for x in $\mathbb{R}^n$, Lemma 1 in [12] yields the existence of $(\lambda_x) > 0$ and y in $S^{n-1} \triangleq \{x \in \mathbb{R}^n : \|x\| = 1\}$ satisfying $x = (\,(\lambda_x)^{r_1} y_1, \dots, (\lambda_x)^{r_n} y_n\,)$, and upon setting $m_V = \min_{y \in S^{n-1}} V(y)$, we get

$$V(x) \;=\; (\lambda_x)^k\, V(y) \;\geq\; m_V (\lambda_x)^k. \tag{67}$$

We now define r and R by setting $r \triangleq \min_{i=1,\dots,n} r_i$ and $R \triangleq \max_{i=1,\dots,n} r_i$, and we obtain

$$\|x\|^2 \;=\; (\lambda_x)^{2r_1} y_1^2 + \cdots + (\lambda_x)^{2r_n} y_n^2 \;\leq\; \begin{cases} (\lambda_x)^{2R} & \text{if} \quad (\lambda_x) \geq 1, \\ (\lambda_x)^{2r} & \text{if} \quad (\lambda_x) < 1. \end{cases}$$

Thus, for each $t \geq t_0$, the inequalities (66) and (67) together with the expression of $b_i^{\bar{n}-k-1}(t)$ yield

$$\begin{aligned} \|x(t, x_0, t_0)\|^{\frac{k}{R}} \;&\leq\; \frac{\beta_i^{\bar{n}-1-k}}{m_V} e^{-\theta t} \quad \text{if } (\lambda_{x(t,x_0,t_0)}) \geq 1, \\ \|x(t, x_0, t_0)\|^{\frac{k}{r}} \;&\leq\; \frac{\beta_i^{\bar{n}-1-k}}{m_V} e^{-\theta t} \quad \text{if } (\lambda_{x(t,x_0,t_0)}) < 1. \end{aligned}$$

Because $t_0 = kT + t_0'$ with $t_0' \geq 0$, it is easily seen from (64) that we have

$$\begin{aligned} \|x(t, x_0, t_0)\| \;&\leq\; \left(\frac{\beta_i^{\bar{n}-1}}{m_V}\right)^{\frac{R}{k}} e^{-\theta \frac{R}{k}(t-t_0)} \quad \text{if } (\lambda_{x(t,x_0,t_0)}) \geq 1, \\ \|x(t, x_0, t_0)\| \;&\leq\; \left(\frac{\beta_i^{\bar{n}-1}}{m_V}\right)^{\frac{r}{k}} e^{-\theta \frac{r}{k}(t-t_0)} \quad \text{if } (\lambda_{x(t,x_0,t_0)}) < 1 \end{aligned}$$

for each $t \geq t_0$. Thus, we get

$$\|x(t, x_0, t_0)\| \;\leq\; \max \left[\left(\frac{\beta_i^{\bar{n}-1}}{m_V}\right)^{\frac{R}{k}}, \left(\frac{\beta_i^{\bar{n}-1}}{m_V}\right)^{\frac{r}{k}} \right] e^{-\theta \frac{r}{k}(t-t_0)}, \quad t \geq t_0.$$

We note that the real

$$\max \left[\left(\frac{\beta_i^{\bar{n}-1}}{m_V}\right)^{\frac{R}{k}}, \left(\frac{\beta_i^{\bar{n}-1}}{m_V}\right)^{\frac{r}{k}} \right]$$

depends solely on $\|x_0\|$ (uniformly in t_0). Further, by definition, the integer $\bar{n}(x_0)$ converges to $+\infty$ as $\|x_0\|$ tends to 0, so that $\beta_i^{\bar{n}}$ converges to 0 as $\|x_0\|$ tends to 0. We therefore obtain that v globally uniformly asymptotically stabilizes with exponential convergence the systems S_1 and S_2 (according to Definition 2.1 (iv)).

Simultaneous local asymptotic stabilization in the case $s < 0$.

We let

$$\beta_1^n \;\triangleq\; \frac{\bar{\beta}}{2n-1} \qquad \text{and} \qquad \beta_2^n \;\triangleq\; \frac{\bar{\beta}}{2n}, \quad n = 1, 2, \dots,$$

with $\bar{\beta}$ as given in (58), and we define $b_i^n : [0, \infty) \to (0, \infty)$ by setting $b_i^n(t) = h_{\beta_i^n}(t)$ for each $t \geq 0$, each $i = 1, 2$, and each $n = 1, 2, \dots$, with h_β as given in (59). It can easily be checked that the mappings b_i^n satisfy the assumptions of Theorem 5.1(a). Thus, there exists a continuous time-varying state feedback law $v : [0, \infty) \times V^{-1}(\,[0, \bar{\beta}{+}1)\,) \to \mathbb{R}^m$ which simultaneously locally asymptotically stabilizes S_1 and S_2. $\square$

7. Conclusion. In this paper, merely continuous and time-varying nonlinear state feedback have proved very useful for simultaneous stabilization of countable families of linear and nonlinear systems. We believe that the use of such feedback for robust stabilization of *uncountable* families of systems also should yield interesting results; we continue our investigation in that direction [6].

8. Appendix. The following lemma is the key argument in establishing the results of this paper. It is either implicitly or explicitly used in the proofs of all the theorems and propositions of this paper.

LEMMA 8.1. *Let D be a bounded neighborhood of the origin in $\mathbb{R}^n$ (resp., $D = \mathbb{R}^n$) and let $V : \overline{D} \to (0, \infty)$ be a Lyapunov function (resp., a radially unbounded Lyapunov function). Further, let the mapping $f : [0, \infty) \times \overline{D} \to \mathbb{R}^n$ be continuous and let the mapping $b : [0, \infty) \to (0, \inf_{x \in \partial D} V(x))$ be C^1. Finally, set $W^\beta \triangleq \{x \in D : \; V(x) < \beta\}$ for each $\beta > 0$, and assume that*

$$\nabla V(x)\, f(t, x) \; < \; \dot{b}(t), \quad x \in \partial W^{b(t)}, \quad t \geq 0. \tag{68}$$

Then, for each $t_0 \geq 0$ and each x_0 in $\overline{W}^{b(t_0)}$, each solution $x(\cdot, x_0, t_0)$ of $\dot{x} = f(t, x)$ satisfies

$$x(t, x_0, t_0) \; \in \; \overline{W}^{b(t)}, \quad t \geq t_0.$$

Proof. Fix $t_0 \geq 0$ and x_0 in $\overline{W}^{b(t_0)}$ and let $x(\cdot, x_0, t_0)$ denote the solution of $\dot{x} = f(t, x)$ that starts from x_0 at time t_0. In order to prove the lemma, the next two claims will be needed.

CLAIM 1. *Let $t_3 > t_0$ be such that $x(t, x_0, t_0)$ lies in D for each t in $[t_0, t_3)$. Then we have the inequality $V(x(t, x_0, t_0)) \leq b(t)$, $t \in [t_0, t_3)$.*

Assume that Claim 1 does not hold. Then there exists t_2 in $[t_0, t_3)$ such that we have $V(x(t_2, x_0, t_0)) > b(t_2)$. Because we have $V(x(t_0, x_0, t_0)) - b(t_0) \leq 0$, continuity of the mapping $V(x(\cdot, x_0, t_0)) - b(\cdot)$ yields the existence of t_1 in $[t_0, t_2)$ and h_1 in $(0, t_3 - t_1)$ such that

$$V(x(t_1, x_0, t_0)) - b(t_1) = 0, \tag{69}$$

and

$$V(x(t_1 + h, x_0, t_0)) - b(t_1 + h) > 0, \quad h \in (0, h_1). \tag{70}$$

By assumption $x(t_1, x_0, t_0)$ belongs to D, and we obtain from (69) that $x(t_1, x_0, t_0)$ lies in $\partial W^{b(t_1)}$. Thus, assumption (68) yields

$$\frac{d}{dt}|_{t=t_1} V(x(t, x_0, t_0)) = \nabla V(x(t_1, x_0, t_0))\, f(\, t_1, x(t_1, x_0, t_0)\,) \; < \; \dot{b}(t_1),$$

and from the continuity of the mappings $\nabla V(\cdot)$, $f(\cdot, \cdot)$, and $\dot{b}(\cdot)$, combined with (69) we get

$$V(x(t_1 + h, x_0, t_0)) \; < \; b(t_1 + h) \quad \text{for} \quad h > 0 \quad \text{small enough,}$$

which is a contradiction of (70). The proof of Claim 1 is thus complete.

CLAIM 2. *$x(t, x_0, t_0)$ lies in D for each $t \geq t_0$.*

Since Claim 2 clearly holds if $D = \mathbb{R}^n$, we assume that D is bounded. Suppose that the claim does not hold. Because x_0 belongs to D, it is easily seen [5, Lemma B.2.iii, p. 152] that there exists $t_1 > t_0$ such that

$$x(t_1, x_0, t_0) \in \partial D, \qquad \text{and} \qquad x(t, x_0, t_0) \in D, \quad t \in [t_0, t_1). \tag{71}$$

Thus, Claim 1 yields

$$V(x(t, x_0, t_0)) \leq b(t), \quad t \in [t_0, t_1). \tag{72}$$

Further, from (71) combined with the definition of b, we obtain

$$V(x(t_1, x_0, t_0)) \geq \inf_{x \in \partial D} V(x) > b(t_1),$$

and continuity of the mapping $V(x(\cdot, x_0, t_0)) - b(\cdot)$ at t_1 yields the existence of h_1 in $(0, t_1 - t_0)$ satisfying

$$V(x(t_1 - h, x_0, t_0)) > b(t_1 - h), \quad h \in (0, h_1),$$

which contradicts (72). Hence, we have Claim 2.

We now let $t_3 > t_0$. By Claim 2, the point $x(t, x_0, t_0)$ lies in D for each $t \geq t_0$, so that Claim 1 applied with t_3 yields

$$V(x(t, x_0, t_0)) \leq b(t), \quad t \in [t_0, t_3).$$

The proof of the lemma is completed upon noting that this last argument holds for all $t_3 > t_0$, all x_0 in $\overline{W}^{b(t_0)}$, and all $t_0 \geq 0$. □

The following lemma is used in the proof of Theorem 5.1 in order to establish robust asymptotic stabilizability of certain families of nonlinear systems.

LEMMA 8.2. *Let D_1 and D_2 be two bounded neighborhoods of the origin (resp., Let $D_1 = D_2 = \mathbb{R}^n$), and let $V_1 : \overline{D}_1 \to [0, \infty)$ and $V_2 : \overline{D}_2 \to [0, \infty)$ be two Lyapunov functions (resp., two radially unbounded Lyapunov functions). Further, let b_1 and b_2 be two continuous mappings from $[0, \infty)$ into $(0, \inf_{x \in \partial D_1} V_1(x))$ and $(0, \inf_{x \in \partial D_2} V_2(x))$, resp., such that*

$$D_1 \cap V_1^{-1}([0, b_1(t))) \supset D_2 \cap V_2^{-1}([0, b_2(t)]), \quad t \geq 0. \tag{73}$$

Then, there exist two C^∞ mappings $a_1 : [0, \infty) \to (0, \inf_{x \in \partial D_1} V_1(x))$ and $c_2 : [0, \infty) \to (0, \inf_{x \in \partial D_2} V_2(x))$ such that for each $t \geq 0$ we have

$$b_1(t) > a_1(t) \qquad \text{and} \qquad c_2(t) > b_2(t), \tag{74}$$

together with

$$D_1 \cap V_1^{-1}([0, a_1(t))) \supset D_2 \cap V_2^{-1}([0, c_2(t)]). \tag{75}$$

Proof. For each $t \geq 0$ and each δ_t, either in $(0, t)$ if $t > 0$ or in $(0, \infty)$ if $t = 0$, we set

$$I(t, \delta_t) \triangleq \begin{cases} (t - \delta_t, t + \delta_t) & \text{if} \quad t > 0, \\ [0, \delta_t) & \text{if} \quad t = 0. \end{cases}$$

As usual, $\overline{I}(t, \delta_t)$ denotes the closure of $I(t, \delta_t)$. We first construct a mapping $a_1 : [0, \infty) \to (0, \inf_{x \in \partial D_1} V_1(x))$ such that $b_1(t) > a_1(t)$ for each $t \geq 0$ and

$$D_1 \cap V_1^{-1}([0, a_1(t))) \supset D_2 \cap V_2^{-1}([0, b_2(t)]), \quad t \geq 0. \tag{76}$$

Construction of a_1. Fix t in $[0,\infty)$. The continuity of the mapping b_1 and b_2 combined with (73) yields the existence of $\delta_t > 0$ (with δ_t in $(0,t)$ if $t>0$) [5, Lemma B.12, p. 163] such that

$$D_1 \cap V_1^{-1}\left(\left[0, \min_{\tau\in\bar{I}(t,\delta_t)} b_1(\tau)\right)\right) \supset D_2 \cap V_2^{-1}\left(\left[0, \max_{\tau\in\bar{I}(t,\delta_t)} b_2(\tau)\right]\right).$$

It is easily seen that there exists α_t satisfying

$$\alpha_t \in \left(0, \min_{\tau\in\bar{I}(t,\delta_t)} b_1(\tau)\right) \tag{77}$$

and

$$D_1 \cap V_1^{-1}([0,\alpha_t)) \supset D_2 \cap V_2^{-1}\left(\left[0, \max_{\tau\in\bar{I}(t,\delta_t)} b_2(\tau)\right]\right). \tag{78}$$

We now extract from the family $\{I(t,\delta_t)\}_{t\in[0,\infty)}$ a countable subcover $\{I(t_k,\delta_{t_k})\}_{k=0}^{\infty}$ such that $t_k < t_{k+1}$ for all $k=0,1,\dots$. We let δ'_{t_k} and δ''_{t_k} in $(0,\delta_{t_k}]$ be such that the sets

$$I_0 \triangleq [0,\delta_0) \qquad \text{and} \qquad I_k \triangleq (t_k-\delta'_{t_k}, t_k+\delta''_{t_k}), \quad k=1,2,\dots,$$

form an open cover of $[0,\infty)$ and any t in $[0,\infty)$ lies in at most two successive sets I_k and I_{k+1}. We let $\{\bar{p}_k\}_{k=0}^{\infty}$ be a partition of unity subordinate to $\{I_k\}_{k=0}^{\infty}\backslash\{0\}$ such that for each $k=1,2,\dots$, the support of $\bar{p}_k$ is included in I_k and the support of $\bar{p}_0$ is included in $(0,\delta t_0)$ [15, p. 10]. For each $k=0,1,\dots$, we define the C^∞ mapping $p_k : [0,\infty) \to [0,1]$ by setting

$$\begin{aligned} &p_k(t) = \bar{p}_k(t), \quad t>0, \quad k=0,1,\dots,\\ &p_0(0)=1 \qquad \text{and} \qquad p_k(0)=0, \quad k=1,2,\dots. \end{aligned}$$

Next, we let the mapping $a_1 : [0,\infty) \to (0,\infty)$ be given by

$$a_1(t) = \sum_{k=0}^{\infty} \alpha_{t_k} p_k(t), \quad t\geq 0.$$

Let t be in $[0,\infty)$. Then, there exists an open neighborhood U_t of t in $[0,\infty)$ that intersects with at most two sets I_k and I_{k+1} of the family $\{I_k\}_{k=0}^{\infty}$. It follows that we have $a_1(\tau) = \alpha_{t_k}p_k(\tau) + \alpha_{t_{k+1}}p_{k+1}(\tau)$ for each τ in U_t. Thus, a_1 is C^∞ on $[0,\infty)$. Further, in view of (77), we get that $0 < a_1(t) < b_1(t)$ for each $t \geq 0$.

We now show that (76) holds. Let t be in $[0,\infty)$. If t lies in a unique set I_k, we get $a_1(t) = \alpha_{t_k}$, so that the inclusion (76) follows from (78). On the other hand, if t lies in two sets I_k and I_{k+1}, we get

$$a_1(t) = \alpha_{t_k} p_k(t) + \alpha_{t_{k+1}} p_{k+1}(t). \tag{79}$$

As (78) yields

$$D_1 \cap V_1^{-1}([0,\alpha_{t_j})) \supset D_2 \cap V_2^{-1}([0,b_2(t)]), \quad j=k,k+1,$$

and since we have $a_1(t) \geq \min(\alpha_{t_k},\alpha_{t_{k+1}})$ (this follows from (79)), we obtain (76).

Construction of c_2. The construction of a mapping c_2 satisfying (74) and (75) is similar to that of a_1. From the continuity of the mappings a_1 and b_2, we obtain from (76) a real $\delta_t > 0$ [5, Lemma B.12, p. 163] such that

$$D_1 \cap V_1^{-1}\left(\left[0, \min_{\tau \in \overline{I}(t,\delta_t)} a_1(\tau)\right)\right) \supset D_2 \cap V_2^{-1}\left(\left[0, \max_{\tau \in \overline{I}(t,\delta_t)} b_2(\tau)\right]\right).$$

Thus, there exists γ_t [5, Lemma B.11, p. 162] satisfying

$$\gamma_t \in \left(\max_{\tau \in \overline{I}(t,\delta_t)} b_2(\tau), \inf_{x \in \partial D_2} V_2(x) \right)$$

and

$$D_1 \cap V_1^{-1}\left(\left[0, \min_{\tau \in \overline{I}(t,\delta_t)} a(\tau)\right)\right) \supset D_2 \cap V_2^{-1}([0, \gamma_t]).$$

From the new family $\{I(t,\delta_t)\}_{t \geq 0}$, we construct some sequences of sets $\{I(t_k, \delta_{t_k})\}_{k=0}^{\infty}$ and $\{I_k\}_{k=0}^{\infty}$, and a family of mappings $\{p_k\}_{k=0}^{\infty}$ exactly as we did in the construction of a_1. Next, we let the mapping $c_2 : [0,\infty) \to (0,\infty)$ be given by

$$c_2(t) = \sum_{k=0}^{\infty} \gamma_{t_k}\, p_k(t), \quad t \geq 0.$$

Using arguments similar to those used for a_1, it can be shown that c_2 is C^∞ on $[0,\infty)$, and that both (74) and (75) hold. Hence, we have the lemma. □

Acknowledgments. The authors would like to acknowledge the valuable comments made by the anonymous referees.

REFERENCES

[1] D. Bertilsson and V. Blondel, *Semialgebraic sets, stabilization, and computability*, in Proceedings of the 34th Conference on Decision and Control, New Orleans, LA, Vol. 4, IEEE, New York, 1995.

[2] V. Blondel and M. Gevers, *The simultaneous stabilizability question of three linear systems is undecidable*, Math. Control Signals Systems, 6 (1994), pp. 135–145.

[3] V. Blondel, M. Gevers, R. Mortini, and R. Rupp, *Simultaneous stabilization of three or more plants: Conditions on the positive real axis do not suffice*, SIAM J. Control Optim., 32 (1994), pp. 572–590.

[4] B. Ghosh and C. Byrnes, *Simultaneous stabilization and pole-placement by nonswitching dynamic compensation*, IEEE Trans. Automat. Control, 28 (1983), pp. 735–741.

[5] B. Ho-Mock-Qai, *Simultaneous and Robust Stabilization of Nonlinear Systems by Means of Continuous and Time-Varying Feedback*, Ph.D. thesis, University of Maryland, College Park, MD, 1996; Institute for Systems Research, Tech. report 96-8.

[6] B. Ho-Mock-Qai, *Robust exponential stabilization of parameterized families of LTI systems by means of nonlinear time-varying state feedback*, in Proceedings of the Conference on Decision and Control, San Diego, CA, IEEE, New York, 1997, pp. 2633–2638.

[7] B. Ho-Mock-Qai, *Simultaneous stabilization of LTI systems by means of nonlinear time-varying feedback*, in Proceedings of the American Control Conference, Albuquerque, NM, IEEE, New York, 1997, pp. 3145–3149.

[8] B. Ho-Mock-Qai and W. P. Dayawansa, *Simultaneous stabilization of countable families of nonlinear systems by means of continuous feedback*, in Proceedings of the 35th Conference on Decision and Control, Kobe, Japan, IEEE, New York, 1996, pp. 2630–2635.

[9] P. T. Kabamba and C. Yang, *Simultaneous controller design for linear time-invariant systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 106–111.

[10] P. P. Khargonekar, A. M. Pascoal, and R. Ravi, *Strong, simultaneous and reliable stabilization of finite-dimensional linear time-varying plants*, IEEE Trans. Automat. Control, 33 (1988), pp. 1158–1161.

[11] I. R. Petersen, *A procedure for simultaneously stabilizing a collection of single input linear systems using non-linear state feedback control*, Automatica J. IFAC, 23 (1987), pp. 33–40.

[12] L. Rosier, *Homogeneous Lyapunov function for homogeneous continuous vector field*, Systems Control Lett., 19 (1992), pp. 467–473.

[13] R. Saeks and J. Murray, *Fractional representation, algebraic geometry, and the simultaneous stabilization problem*, IEEE Trans. Automat. Control, 27 (1982), pp. 895–903.

[14] M. Vidyasagar and N. Viswanadham, *Algebraic design techniques for reliable stabilization*, IEEE Trans. Automat. Control, 27 (1982), pp. 1085–1095.

[15] F. W. Warner, *Foundations of Differentiable Manifolds and Lie Groups*, Springer-Verlag, New York, 1983.

[16] C. Zhang and V. Blondel, *Simultaneous stabilization using LTI compensator with a sampler and hold*, Internat. J. Control, 57 (1993), pp. 293–308.

SIAM J. Control Optim.
Vol. 37, No. 6, pp. 1726–1741

OPTIMAL CONTROL PROBLEMS WITH PARTIALLY POLYHEDRIC CONSTRAINTS[*]

J. FRÉDÉRIC BONNANS[†] AND HOUSNAA ZIDANI[‡]

Abstract. This paper discusses a class of state constrained optimal control problems, for which it is possible to formulate second-order necessary or sufficient conditions for local optimality or quadratic growth that do not involve all curvature terms for the constraints. This kind of result is classical in the case of polyhedric control constraints. Our theory of optimization problems with partially polyhedric constraints allows to extend these results to the case when the control constraints are polyhedric, in the presence of state constraints satisfying some specific hypotheses. The analysis is based on the assumption that some strict semilinearized qualification condition is satisfied. We apply the theory to some optimal control problems of elliptic equations with state and control constraints.

AMS subject classifications. Primary, 49K40; Secondary, 49K20, 35B30, 35J60, 90C31

Key words. optimal control, elliptic systems, sensitivity analysis, expansion of solutions, second-order optimality conditions, quadratic growth, Legendre forms, polyhedricity

PII. S0363012998333724

1. Introduction.

This paper discusses a class of optimal control problems that have local control constraints and a finite number of state constraints. The problem was considered recently in [15], where second-order necessary optimality conditions were obtained. The aim of this paper is to generalize this type of result to more general optimal control problems that have two types of constraints, the first of them being polyhedric. We also prove that this type of second-order condition allows us to state second-order sufficient conditions, and in fact that they allow us to characterize quadratic growth. Our basic tools are the second-order necessary conditions based on second-order tangent sets and polyhedricity theory.

The approach of second-order necessary conditions based on second-order tangent sets was renewed in [22], where the computation of the contribution of the curvature of the feasible set to second-order necessary conditions was done in the case of nonnegative continuous functions of time. This approach was extended in [16, 20, 21, 31] for abstract optimization problems and more recently in [29, 30] for optimal control problems.

Polyhedricity theory for convex sets is a classical tool for obtaining formulas for the directional derivative of the projection over a convex set [18, 28], was applied to nonlinear control problems [36, 26], and has been linked to the recent work on sensitivity analysis for abstract optimization problems [6, 8, 10].

The paper is organized as follows. Section 2 presents a theory of second-order necessary or sufficient optimality conditions for abstract optimization problems that satisfy the strict semilinearized qualification condition. In the case corresponding to an optimal control problem with polyhedric control constraints and a finite number of additional inequality constraints, the theory is complete in the sense that there is no gap between the necessary and sufficient conditions. More precisely, we obtain a

[*] Received by the editors February 4, 1998; accepted for publication (in revised form) September 15, 1998; published electronically September 30, 1999.
http://www.siam.org/journals/sicon/37-6/33372.html

[†] INRIA Rocquencourt, B.P. 105, 78153 Le Chesnay, France (frederic.bonnans@inria.fr).

[‡] UFR Math, Lab MIP, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse, France. Current address: INRIA Rocquencourt, B.P. 105, 78153 Le Chesnay, France (housnaa.zidani@inria.fr).

characterization of the quadratic growth condition.

In section 3, assuming a weak second-order sufficient condition and the strict semilinearized qualification condition, we provide a formula for computing the directional derivative of the optimal control (as well as a second-order expansion of the value function) with respect to a perturbation.

Section 4 discusses the application of the previous results to some optimal control problems of elliptic equations. We consider the case of nonnegative control subject to a finite number of state constraints.

Notations. Let (P) be an optimization problem. By $F(\mathrm{P})$, ε-$S(\mathrm{P})$, and val(P), we denote the feasible set, set of ε solutions, and value of problem (P), respectively.

2. Second-order abstract optimality conditions.

In this section we discuss the theory of second-order optimality conditions for optimization problems of the following type:

$$\text{(AP)} \qquad \mathrm{Min}_x f(x);\ x \in K_X;\ G(x) \in K_Y.$$

Here X and Y are Banach spaces, K_X and K_Y are closed convex subsets of X and Y, respectively, and f and G are twice continuously differentiable mappings from X into $\mathbb{R}$ and Y. We note that, if $\mathcal{K}$ is a convex subset of a Banach space $\mathcal{X}$ and $x \in \mathcal{K}$, then the tangent and normal cones $T_{\mathcal{K}}$ and $N_{\mathcal{K}}$ and the cone of feasible directions $\mathcal{R}_{\mathcal{K}}$ are defined as

$$\begin{aligned} T_{\mathcal{K}}(x) &:= \{y \in \mathcal{X};\ \exists\, x(\sigma) = x + \sigma y + o(\sigma) \in \mathcal{K},\ \sigma \geq 0\},\\ N_{\mathcal{K}}(x) &:= \{x^* \in \mathcal{X}^*;\ \langle x^*, y\rangle \leq 0, \forall y \in T_{\mathcal{K}}(x)\},\\ \mathcal{R}_{\mathcal{K}}(x) &:= \{y \in \mathcal{X};\ \exists\, t > 0;\ x + ty \in \mathcal{K}\}, \end{aligned}$$

with the convention that these sets are empty if $x \notin \mathcal{K}$. An interesting case is when K_X is *polyhedric* in the following sense [28, 18].

DEFINITION 2.1. *Let $x_0 \in K_X$ and $x^* \in N_{K_X}(x_0)$. We say that K_X is polyhedric at x_0 for the direction x^* if*

$$T_{K_X}(x_0) \cap (x^*)^{\perp} = \overline{\mathcal{R}_{K_X}(x_0) \cap (x^*)^{\perp}}. \tag{2.1}$$

If K_X is polyhedric at each $x_0 \in K_X$ $\forall$ $x^* \in N_{K_X}(x_0)$, we say that K_X is polyhedric.

By setting

$$\mathcal{K} := K_X \times K_Y, \quad \mathcal{Y} := X \times Y, \quad \mathcal{G}(x) := (x, G(x)),$$

we can write the abstract optimization problem (AP) under the form

$$\text{(AP2)} \qquad \mathrm{Min}_x f(x)(\text{such that})\mathcal{G}(x) \in \mathcal{K},$$

with $\mathcal{G}(x)$, twice continuously differentiable mapping from X into $\mathcal{Y}$, and $\mathcal{K}$ closed convex subset of $\mathcal{Y}$. We will use the relationship between the two formats several times, in order to use the results that were derived for problem (AP2). For instance, the standard constraint qualification condition for $x_0 \in F(\mathrm{AP2})$, due to Robinson [32], is as follows:

$$0 \in \ \mathrm{int}\{D\mathcal{G}(x_0)X - (\mathcal{K} - \mathcal{G}(x_0))\}. \tag{2.2}$$

LEMMA 2.2 (Robinson [32]). *Let $x_0 \in F(\mathrm{AP2})$ satisfy (2.2). Then the following metric regularity property holds. There exist $\varepsilon > 0$ and $\alpha > 0$ such that $\forall$ $x \in B(x_0, \varepsilon)$ $\exists$ $\hat{x} \in \mathcal{G}^{-1}(\mathcal{K})$ satisfying*

$$\|\hat{x} - x\| \leq \alpha \mathrm{dist}(\mathcal{G}(x), \mathcal{K}).$$

It is easy to show (e.g., [10]) that the qualification condition for a problem of the form (AP) (after it has been put under the form (AP2)) is equivalent to

$$0 \in \operatorname{int}\{DG(x_0)(K_X - x_0) - (K_Y - G(x_0))\}. \tag{2.3}$$

The *critical cone* at $x_0 \in F(\mathrm{AP})$ is defined as the set of directions of nonincrease of the cost function that are tangent to the feasible set. More precisely,

$$C(x_0) := \{h \in T_{K_X}(x_0); Df(x_0)h \le 0;\ DG(x_0)h \in T_{K_Y}[G(x_0)]\}.$$

The Lagrangian function and the set of Lagrange multipliers are defined as

$$\begin{aligned}\mathcal{L}(x,\lambda) &:= f(x) + \langle \lambda, G(x)\rangle,\\ \Lambda(x) &:= \{(q,\lambda) \in N_{K_X}(x) \times N_{K_Y}[G(x)];\ D_x\mathcal{L}(x,\lambda) + q = 0\}.\end{aligned}$$

LEMMA 2.3 (Zowe and Kurcyusz [37]). *Let x_0 be a local solution of* (AP) *satisfying the qualification hypothesis* (2.3)*. Then with x_0 is associated a nonempty and bounded set of Lagrange multipliers.*

It is convenient to use the following well-known characterization of the critical cone.

LEMMA 2.4. *Let $\Lambda(x) \neq \emptyset$, say contains (q,λ). Then $Df(x)h = 0$ whenever $h \in C(x)$, and*

$$C(x) = \{h \in T_{K_X}(x) \cap q^\perp;\ DG(x)h \in T_{K_Y}[G(x)] \cap \lambda^\perp\}. \tag{2.4}$$

Proof. Let $h \in X$ be tangent to the feasible set of (AP), in the sense that $h \in T_{K_X}(x)$ and $DG(x)h \in T_{K_Y}[G(x)]$. By definition of $\Lambda(x)$, we have

$$0 = \langle D_x\mathcal{L}(x,\lambda) + q, h\rangle = Df(x)h + \langle \lambda, DG(x)h\rangle + \langle q, h\rangle.$$

Since the last two terms are nonpositive, we have $Df(x)h \ge 0$, and $Df(x)h \le 0$ iff the last two terms are zero. The result follows. □

Let $x \in F(\mathrm{AP})$. Using the above lemma, we may view the critical cone as a linearization of the following set:

$$A(q,\lambda) := \{h \in (K_X - x) \cap q^\perp;\ DG(x)h \in T_{K_Y}[G(x)] \cap \lambda^\perp\}. \tag{2.5}$$

Note that in this expression we chose to "linearize" the constraint $G(x) \in K_Y$ but not the relation $x \in K_X$. The set $A(q,\lambda)$ is the inverse image, through the linear continuous mapping $h \to (h, DG(x)h)$, of the closed convex set

$$\left((K_X - x) \cap q^\perp\right) \times \left(T_{K_Y}[G(x)] \cap \lambda^\perp\right).$$

We will use the associated qualification condition which we will call the *strict semilinearized qualification condition* (we justify this terminology below). From the above discussion, it follows that the expression of the strict semilinearized qualification condition is

$$0 \in \operatorname{int}\left\{DG(x)\left((K_X - x) \cap q^\perp\right) - T_{K_Y}[G(x)] \cap \lambda^\perp\right\}. \tag{CQA}$$

We may compare this condition to the more classical *strict qualification condition*, introduced in [34] (see also [4]), whose expression for problem (AP) is

$$0 \in \operatorname{int}\left\{DG(x)\left((K_X - x) \cap q^\perp\right) - (K_Y - G(x)) \cap \lambda^\perp\right\}. \tag{2.6}$$

LEMMA 2.5. (i) *Condition* (2.6) *implies* (CQA), *and both conditions are equivalent if K_Y is a polyhedron in the Banach space Y.*

(ii) *Assume that the standard constraint qualification* (2.3) *holds. Then condition* (CQA) *implies existence and uniqueness of the Lagrange multiplier.*

Proof. (i) Since $(K_Y - G(x)) \subset T_{K_Y}(G(x))$, we obviously have that (2.6) implies (CQA). Assume now that K_Y is a polyhedron in Y and that (CQA) holds. Since $\mathbb{R}_+(K_Y - G(x)) = \mathcal{R}_{K_Y}(x)$ is equal to $T_{K_Y}(G(x))$, conditions (CQA) and (2.6) are obviously equivalent.

(ii) It is known that the strict qualification condition (2.6) implies existence and uniqueness of the Lagrange multiplier, see [34]. Since (CQA) is nothing but the strict qualification condition after linearization of the second constraint (that leaves invariant the set of Lagrange multipliers), we obtain that the set of Lagrange multipliers that by (2.2) is nonempty is in fact a singleton. □

It is possible to express a second-order necessary optimality condition for problem (AP), using the result of [16], in terms of the *second-order tangent set* to $K_Y \subset Y$ at $y \in K_Y$ in direction $z \in T_{K_Y}(y)$ that is defined as

$$T^2_{K_Y}(y,z) := \left\{ w \in Y;\ y + tz + \frac{t^2}{2}w + o(t^2) \in K_Y, t \geq 0 \right\}.$$

Let x_0 be a local minimum of (AP) satisfying (2.3). Set

$$\mathcal{T}(h) := T^2_{K_Y}[G(x_0), DG(x_0)h].$$

In all that follows, we shall use this definition of a support function.

DEFINITION 2.6. *Let $\mathcal{K}$ a subset of a Banach space $\mathcal{X}$, and let x^* be in $\mathcal{X}^*$. The support function of $\mathcal{K}$ at x^* is $\sigma(x^*, \mathcal{K}) := \sup\{\langle x^*, x\rangle \ :\ x \in \mathcal{K}\}$.*

The following theorem is obtained by combining the result of [16] with some polyhedricity properties. Note that if $\mathcal{T}(h) = \emptyset$, then $\sigma(\cdot, \mathcal{T}(h))$ is identically equal to $-\infty$; therefore, in that case, the conclusion of case (i) of Lemma 2.5 is trivially satisfied.

THEOREM 2.7. *Assume that K_X is polyhedric. Let x_0 be a local minimum of* (AP) *satisfying* (2.3) *and the strict semilinearized qualification condition* (CQA). *Then*

(i) *$C(x_0) \cap \mathcal{R}_{K_X}(x_0)$ is a dense subset of $C(x_0)$, and each $h \in C(x_0) \cap \mathcal{R}_{K_X}(x_0)$ satisfies*

$$D^2_{x^2}\mathcal{L}(x_0, \lambda)(h,h) - \sigma(\lambda, \mathcal{T}(h)) \geq 0, \tag{2.7}$$

where λ is the Lagrange multiplier.

(ii) *If in addition $h \to \sigma(\lambda, \mathcal{T}(h))$ is lower semicontinuous over $C(x_0)$, then* (2.7) *holds for all critical direction h.*

(iii) *If K_Y is a polyhedron, then for all critical direction h we have*

$$D^2_{x^2}\mathcal{L}(x_0, \lambda)(h,h) \geq 0. \tag{2.8}$$

Proof. (i), step a. We claim that $C(x_0) \cap \mathcal{R}_{K_X}(x_0)$ is a dense subset of $C(x_0)$. Let $h \in C(x_0)$, and fix $\varepsilon > 0$. Since K_X is polyhedric, $\exists\ \hat{h}_\varepsilon \in \mathcal{R}_{K_X}(x_0) \cap q^\perp$ such that $\|\hat{h}_\varepsilon - h\| \leq \varepsilon$. Let $t_\varepsilon > 0$ be such that $x_0 + t_\varepsilon \hat{h}_\varepsilon \in K_X$. We use the metric regularity property that, by Lemma 2.2, follows from (CQA):

$\exists\ \gamma > 0$ and $\alpha > 0$ such that, if $\hat{w} \in X$, and $\|\hat{w}\| \leq \gamma$,
then $\exists\ w \in (K_X - x_0) \cap q^\perp$ such that
$DG(x_0)w \in T_{K_Y}[G(x_0)] \cap \lambda^\perp$, and
$\|w - \hat{w}\| \leq \alpha[\mathrm{dist}(\hat{w}, (K_X - x_0) \cap q^\perp) + \mathrm{dist}(DG(x_0)\hat{w}, T_{K_Y}[G(x_0)] \cap \lambda^\perp)]$.

Reducing t_ε if necessary, we have that $\hat{w}_\varepsilon := t_\varepsilon \hat{h}_\varepsilon$ is such that $\|\hat{w}_\varepsilon\| \le \gamma$ and $\hat{w}_\varepsilon \in (K_X - x_0) \cap q^\perp$. Since

$$\operatorname{dist}(DG(x_0)\hat{w}_\varepsilon, T_{K_Y}[G(x_0)] \cap \lambda^\perp) = O(\varepsilon t_\varepsilon),$$

it follows that $\exists\ w_\varepsilon \in (K_X - x_0) \cap q^\perp$ such that

$$\begin{aligned} &DG(x)w_\varepsilon \in T_{K_Y}[G(x_0)] \cap \lambda^\perp, \\ &\|w_\varepsilon - \hat{w}_\varepsilon\| = \alpha \operatorname{dist}(DG(x_0)\hat{w}_\varepsilon, T_{K_Y}[G(x_0)] \cap \lambda^\perp) = O(t_\varepsilon \varepsilon). \end{aligned}$$

Set $h_\varepsilon := t_\varepsilon^{-1} w_\varepsilon$. Then $h_\varepsilon \in C(x_0) \cap \mathcal{R}_{K_X}(x_0)$, and $\|h_\varepsilon - h\| = O(\varepsilon)$. This proves our claim.

(i), step b. Since x_0 is a qualified local solution of (AP) and the Lagrange multiplier is unique, by [16, Thm. 4.2] the following second-order necessary condition holds. For any critical direction h, we have

$$D^2_{x^2}\mathcal{L}(x_0, \lambda)(h, h) - \sigma(q, T^2_{K_X}(x_0, h)) - \sigma(\lambda, \mathcal{T}(h)) \ge 0. \tag{2.9}$$

Note that we have used here the fact that the support function of a set in product form is the sum of the corresponding support function.

On the other hand, since $q \in N_{K_X}(x_0)$, we have $\sigma(q, T^2_{K_X}(x_0, h)) \le 0$ [16, Sect. 4]. If $h \in \mathcal{R}_{K_X}(x_0)$, then $0 \in T^2_{K_X}(x_0, h)$, so that $\sigma(q, T^2_{K_X}(x_0, h)) = 0$. Point (i) follows.

(ii) Let $h \in C(x_0)$. By (i) $\exists\ h_k \to h$, $h_k \in C(x_0) \cap \mathcal{R}_{K_X}(x_0)$, and

$$D^2_{x^2}\mathcal{L}(x_0, \lambda)(h_k, h_k) \ge \sigma(\lambda, \mathcal{T}(h_k)).$$

Since the right-hand side is lower semicontinuous (l.s.c.) by hypothesis, and $D^2_{x^2}\mathcal{L}(x_0, \lambda)(\cdot, \cdot)$ is a continuous function, we may pass to the limit in this inequality. Point (ii) follows.

(iii) If K_Y is a polyhedron, then it is well known that $0 \in \mathcal{T}(h)$ (e.g., [10]), whence $\sigma(\lambda, \mathcal{T}(h)) = 0$ for all critical direction h. The result then follows from (ii). □

In order to formulate second-order sufficient conditions, we need the following concept.

Definition 2.8 (see, e.g., [19]). *We say that a quadratic form Q on a Hilbert space X is a Legendre form if Q is weakly l.s.c. and, whenever a sequence $\{x_k\} \subset X$ satisfies $x_k \xrightarrow{w} x$ and $Q(x_k) \to Q(x)$, then $x_k \to x$.*

The function $x \to \|x\|^2$ is the simplest example of a Legendre form. More generally, if $N > 0$ and Q is a weakly continuous quadratic form, it is easy to check that $x \to N\|x\|^2 + Q(x)$ is a Legendre form; see, e.g., [11].

Definition 2.9. *We say that x_0 is a local solution of* (AP) *satisfying the* quadratic growth condition *if*

$$\exists \alpha > 0; \quad f(x) \ge f(x_0) + \alpha\|x - x_0\|^2 + o(\|x - x_0\|^2) \quad \forall x \in F(\text{AP}), \tag{2.10}$$

where F(AP) *is the feasible set of the problem* (AP).

Theorem 2.10. *Assume that K_X is a polyhedric subset of the Hilbert space X. Let x_0 be a qualified local minimum of* (AP) *satisfying the strict semilinearized qualification condition* (CQA)*, and let (q_0, λ_0) be the unique associated Lagrange multiplier. If $Q_0(h) := D^2_{x^2}\mathcal{L}(x_0, \lambda_0)(h, h)$ is a Legendre form and K_Y is a polyhedron, then the following condition is necessary and sufficient for quadratic growth:*

$$D^2_{x^2}\mathcal{L}(x_0, \lambda_0)(h, h) > 0 \quad \forall\ h \in C(x_0)\backslash\{0\}. \tag{2.11}$$

Proof. Let x_0 satisfy the quadratic growth condition. Then $\exists\ \alpha > 0$ such that x_0 is a local solution of the problem

$$(\mathrm{AP}_\alpha) \qquad \mathrm{Min}_x f(x) - \frac{\alpha}{2}\|x - x_0\|^2;\ x \in K_X;\ G(x) \in K_Y.$$

Since K_Y is a polyhedron, and therefore $\sigma(\lambda, \mathcal{T}(h)) = 0$, (2.11) follows from Theorem 2.7.

Conversely, assume that (2.11) holds, while the quadratic growth condition is not satisfied. Then $\exists\ x_k \to x_0$ such that

$$(2.12) \qquad f(x_k) \leq f(x_0) + o(\|x_k - x_0\|^2).$$

Set $t_k := \|x_k - x_0\|$. Extracting a subsequence if necessary, we may assume that $x_k = x_0 + t_k h_k$, $\|h_k\| = 1$, and $h_k \overset{w}{\to} \bar{h}$. Also $\bar{h} \in T_{K_X}(x_0)$ since $h_k \in \mathcal{R}_{K_X}(x_0)$, (2.12) implies that $Df(x_0)\bar{h} \leq 0$, and from $G(x_k) \in K_Y$ we deduce that $DG(x_0)\bar{h} \in T_{K_Y}(G(x_0))$. It follows that $\bar{h}$ is a critical direction.

By the first-order optimality condition we have

$$\langle q_0, x_k - x_0\rangle \leq 0 \text{ and } \langle \lambda_0, G(x_k) - G(x_0)\rangle \leq 0.$$

Combining with $D_x\mathcal{L}(x_0, \lambda_0) + q_0 = 0$, we deduce that

$$\begin{aligned} f(x_k) - f(x_0) &\geq \mathcal{L}(x_k, \lambda_0) - \mathcal{L}(x_0, \lambda_0) + \langle q_0, x_k - x_0\rangle \\ &= \frac{t_k^2}{2} Q_0(h_k) + o(t_k^2). \end{aligned}$$

Combining with (2.12), it follows that $Q_0(h_k) \leq o(1)$. Since $Q_0(\cdot)$ is l.s.c., we have $Q_0(\bar{h}) \leq 0$. Since $\bar{h}$ is critical, with (2.11) this implies that $\bar{h} = 0$. It follows that $Q_0(h_k) \to Q_0(\bar{h})$. Due to $\|h_k\| = 1$ and $\bar{h} = 0$, this contradicts the fact that $Q_0(h_k)$ is a Legendre form. □

3. Abstract sensitivity analysis.

This section is devoted to the study of the family of perturbed optimization problems

$$(\mathrm{AP}_u) \qquad \mathrm{Min}_x f(x, u) \text{ s.t. } x \in K_X;\ \ G(x, u) \in K_Y.$$

Here u belongs to a Banach space U, K_X is a closed convex subset of the Hilbert space X, K_Y is a polyhedron included in the finite dimensional space Y, so that (CQA) is equivalent to the strict qualification condition (2.6), f and G are twice continuously differentiable mappings from $X \times U$ into $\mathbb{R}$ and Y. The Lagrangian of this problem is

$$\mathcal{L}(x, \lambda, u) := f(x, u) + \langle \lambda, G(x, u)\rangle.$$

We perform a sensitivity analysis along a path of perturbation variables of the form

$$u(t) := u_0 + tu_1 + \frac{t^2}{2}u_2 + o(t^2) \quad \text{with } u_i(t) \in U, i = 0, 1, 2.$$

Let x_0 be a local solution of (AP_{u_0}). The following problems may be interpreted as the linearization and the second-order expansion of problem (AP_u) at (x_0, u_0) along the path $u(t)$, respectively:

$$(\mathrm{LP}) \qquad \begin{aligned} \mathrm{Min}_{h \in X} Df(x_0, u_0)(h, u_1) \text{ s.t. } h \in T_{K_X}(x_0);& \\ DG(x_0, u_0)(h, u_1) \in T_{K_Y}[G(x_0, u_0)],& \end{aligned}$$

and, (q_0, λ_0) being the Lagrange multiplier associated with x_0:

$$\text{(SP)} \qquad \text{Min}_{h \in S(LP)} D_u \mathcal{L}(x_0, \lambda_0, u_0)u_2 + D^2_{(x,u)^2}\mathcal{L}(x_0, \lambda_0, u_0)((h, u_1), (h, u_1)).$$

LEMMA 3.1. *Let x_0 satisfy* (CQA). *Then*

(i) S(LP) *is nonempty, and*

$$\text{val (LP)} = D_u \mathcal{L}(x_0, \lambda_0, u_0)u_1, \tag{3.1}$$

where (q_0, λ_0) is the unique Lagrange multiplier associated with x_0,

(ii) *The set* $S(\text{LP}) \cap \mathcal{R}_{K_X}(x_0)$ *is a dense subset of* S(LP).

Proof. (i) The dual, in the sense of convex analysis, to the linearized problem (LP), is known to be (e.g., [8])

$$\text{(LD)} \qquad \text{Max}_{(q,\lambda)} D_u \mathcal{L}(x_0, \lambda, u_0)u_1; \quad (q, \lambda) \in \Lambda(x_0).$$

By Lemma 2.5, we know that there exists a unique Lagrange multiplier (q_0, λ_0), and that the primal and dual values are equal. This proves (3.1). It follows that $h \in X$ is a solution of (LP) iff $h \in F$(LP) and the complementarity conditions

$$\langle q_0, h \rangle = \langle \lambda_0, DG(x_0, u_0)(h, u_1) \rangle = 0$$

are satisfied. In other words, $h \in S$(LP) iff

$$h \in T_{K_X}(x_0) \cap (q_0)^\perp; \quad DG(x_0, u_0)(h, u_1) \in T_{K_Y}[G(x_0, u_0)] \cap (\lambda_0)^\perp.$$

By (CQA) the set of such h is not empty; hence S(LP) is not empty.

(ii) This is a consequence of Theorem 2.7(i) applied to problem (LP), once we have checked that problem (LP) itself satisfies the strict semilinearized qualification condition. The expression of the latter (for problem (LP)) is

$$0 \in \text{ int} \left\{ DG(x)[(T_{K_X(x_0)} - h) \cap q_0^\perp] - (T_{T_{K_Y}[G(x_0)]} DG(x_0)h) \cap \lambda_0^\perp \right\}.$$

Since $(K_X - x_0) \subset T_{K_X(x_0)} - h$ and $T_{K_Y}[G(x_0)] \subset T_{T_{K_Y}[G(x_0)]} DG(x_0)h$, this is an obvious consequence of (CQA). □

THEOREM 3.2. *Assume that*

(i) *For small enough $t > 0$, $\exists\, x(t)$, $o(t^2)$-solution of* $(\text{AP}_{u(t)})$ *such that $x(t) \to x_0$.*

(ii) *The point x_0 is the unique solution of* (AP_{u_0}) *and satisfies* (CQA) *and the second-order sufficient optimality condition* (2.11),

(iii) *The Hessian $Q_0(h) := D^2_{x^2}\mathcal{L}(x_0, \lambda_0, u_0)(h, h)$ is a Legendre form over the Hilbert space X.*

Then

(a) *The following expansion for the value function of* $(\text{AP}_{u(t)})$ *holds:*

$$\text{val } (\text{AP}_{u(t)}) = \text{val } (\text{AP}_{u_0}) + t\text{val (LP)} + \frac{t^2}{2}\text{val (SP)} + o(t^2). \tag{3.2}$$

(b) *One has $x(t) = x_0 + O(t)$. Any weak limit-point of $t^{-1}(x(t) - x_0)$ is a strong limit-point and is a solution of* (SP). *In particular, if* (SP) *has a unique solution $\bar{h}$, then $x(t) = x_0 + t\bar{h} + o(t)$.*

Proof. (a) Let $Q_{u_1}(h) := D^2_{x^2}\mathcal{L}(x_0, \lambda_0)((h, u_1), (h, u_1))$. (Note that this notation is coherent with the definition of $Q_0(\cdot)$ given before.) Consider the subproblem

$$(\text{SP}_\sigma) \qquad \text{Min}_{h \in S(\text{LP})} D_u \mathcal{L}(x_0, \lambda, u_0)u_2 + Q_{u_1}(h) - \sigma(q_0, T^2_{K_X}(x_0, h)).$$

Since K_Y is a polyhedron, we have $\sigma(\lambda_0, T^2_{K_Y}[G(x_0,u_0), DG(x_0,u_0)(h,u_1)]) = 0$. It follows from [8, Prop. 2.1] that

$$\mathrm{val}(\mathrm{AP}_{u(t)}) \le \mathrm{val}(\mathrm{AP}_{u_0}) + t\mathrm{val}(\mathrm{LP}) + \frac{t^2}{2}\mathrm{val}(\mathrm{SP}_\sigma) + o(t^2),$$

while [8, Prop. 4.3] implies that the right-hand side of (3.2) is a lower estimate of $\mathrm{val}(\mathrm{AP}_{u(t)})$. We now prove (3.2) by checking that $\mathrm{val}(\mathrm{SP}) \ge \mathrm{val}(\mathrm{SP}_\sigma)$. By Lemma 3.1, the set $S(\mathrm{LP}) \cap \mathcal{R}_{K_X}(x_0)$ is a dense subset of $S(\mathrm{LP})$. Also on $S(\mathrm{LP}) \cap \mathcal{R}_{K_X}(x_0)$ the cost functions of (SP) and (SP_σ) coincide. Since $\sigma(q_0, T^2_{K_X}(x_0,h)) \le 0$, it follows that

$$\begin{aligned}\mathrm{val}(\mathrm{SP}) &= \inf_{h\in S(\mathrm{LP})\cap\mathcal{R}_{K_X}(x_0)} \{D_u\mathcal{L}(x_0,\lambda_0,u_0)u_2 + Q_{u_1}(h)\}\\ &= \inf_{h\in S(\mathrm{LP})\cap\mathcal{R}_{K_X}(x_0)} \{D_u\mathcal{L}(x_0,\lambda_0,u_0)u_2 + Q_{u_1}(h) - \sigma(q_0, T^2_{K_X}(x_0,h))\}\\ &\ge \mathrm{val}(\mathrm{SP}_\sigma),\end{aligned}$$

as was to be proved.

(b) By [8, Prop 5.3], we have $x(t) = x_0 + O(t)$. Let us prove that any weak limit-point of $t^{-1}(x(t) - x_0)$ is a strong limit-point. Let $t_k \to 0^+$, $x_k := x(t_k)$, and $h_k := t_k^{-1}(x_k - x_0)$ be such that $h_k \overset{w}{\to} \bar{h}$. By [8, Prop. 5.3], we know that

$$Q_{u_1}(h_k) \to Q_{u_1}(\bar{h}).$$

Since $Q_0(\cdot)$ is a Legendre form, we have $h_k \to \bar{h}$, as was to be proved. Finally if (SP) has a unique solution $\bar{h}$, it follows that $t^{-1}(x(t) - x_0)$ converges to $\bar{h}$. The conclusion follows. □

4. Application to state constrained optimal control problems.

4.1. General results.

In this section we apply the results of the previous sections to some optimal control problems for semilinear elliptic equations. In the rest of this paper, we denote by Ω a bounded open subset of $\mathbb{R}^n$ $(n \le 3)$ with Lipschitz boundary Γ. Given a function $u \in L^2(\Omega)$ (we take in this section the standard notations for optimal control problems), we consider the following boundary value problem:

$$-\Delta y + \phi(x,y) = u \ \text{ in } \Omega, \qquad y(x) = 0 \ \text{ on } \Gamma, \tag{4.1}$$

where $\phi : \overline{\Omega} \times \mathbb{R} \longrightarrow \mathbb{R}$ is a continuous function which is of class C^2 and such that $\phi'_y(x,\cdot) \ge 0 \ \forall\, x \in \Omega$.

From now on, the weak solution of (4.1) associated with u will be denoted y_u. Under the above assumption, we can prove the existence and uniqueness of a solution of (4.1).

THEOREM 4.1. *For every $u \in L^2(\Omega)$, (4.1) admits a unique weak solution y_u in $H^1_0(\Omega) \cap C(\overline{\Omega})$, which is Hölder continuous, and we have*

$$\|y_u\|_{C(\overline{\Omega})} \le C_1(1 + \|u\|_{L^2(\Omega)})$$

where $C_1 = C_1(\Omega)$ is independent of u. Moreover, if we denote by $\mathcal{A} : L^2(\Omega) \longrightarrow C(\overline{\Omega})$ the mapping which associates with every control u the weak solution y_u of (4.1), then

$\mathcal{A}$ is twice continuously Fréchet differentiable, and for every $u, h \in L^2(\Omega)$, if we denote $y_u = \mathcal{A}(u)$ and $z_h = \mathcal{A}'(u)h$, then z_h is the weak solution of

$$-\Delta z_h + \phi'_y(x, y_u) z_h = h \quad in\ \Omega, \qquad z_h = 0 \quad on\ \Gamma. \tag{4.2}$$

Proof. The above theorem is a collection of known results for semilinear elliptic equations (see [7, 6, 15] and the general references [1, 2, 5]). □

Consider the following control constraints:

$$L^2_+(\Omega) := \{u \in L^2(\Omega) \mid u(x) \geq 0 \text{ almost everywhere (a.e.) on } x \in \Omega\}.$$

Let us also consider a family of functions G_j of class $C^2 : L^2(\Omega) \to \mathbb{R}$ for $1 \leq j \leq m$. We consider the following optimal control problem:

$$\text{Min}\{F(u) \mid u \in L^2_+(\Omega),\ G_j(u) \leq 0 \text{ for } 1 \leq j \leq m\} \tag{$\mathcal{P}$}$$

where

$$F(u) := \frac{1}{2}\int_\Omega (y_u(x) - y_d(x))^2\, dx + \frac{N}{2}\int_\Omega u(x)^2\, dx, \tag{4.3}$$

with y_d is a given function in $L^2(\Omega)$, and $N > 0$. The adjoint state p^0_u associated with u is defined as the unique solution in $H^2(\Omega)$ of the system

$$-\Delta p^0_u + \phi'_y(x, y_u)p^0_u = y_u - y_d \ \text{ in } \Omega, \quad p^0_u = 0 \ \text{ on } \Gamma.$$

It is known that $u \to F(u)$ is a C^2 mapping with derivative

$$DF(u) = Nu + p^0_u.$$

We will detail later the cases when $G_j(u)$ are some punctual or integral functions of the state.

Let $\bar{u}$ be an optimal solution of problem $(\mathcal{P})$. Set

$$\begin{aligned} J_+ &= \{j \in \{1, \dots, m\} \mid G_j(\bar{u}) < 0\}, \\ J_0 &= \{j \in \{1, \dots, m\} \mid G_j(\bar{u}) = 0\}, \\ J_- &= \{j \in \{1, \dots, m\} \mid G_j(\bar{u}) = 0,\ \bar{\lambda}_j > 0\}. \end{aligned}$$

Then $J_0 \cup J_- \cup J_+ = \{1, \dots, m\}$. Problem $(\mathcal{P})$ can be written as

$$\text{Min}\{F(u) \mid u \in L^2_+(\Omega), G(u) \in \mathbb{R}^m_-\}.$$

In addition (see, e.g., [10, 3]), Robinson's constraint qualification assumption is equivalent to

$$\exists v \in L^2_+(\Omega), \quad G_j(\bar{u}) + DG_j(\bar{u})(v - \bar{u}) < 0. \tag{4.4}$$

Therefore we obtain the following (classical) expression of the first order optimality system.

THEOREM 4.2. *Assume that $\bar{u}$ is a local solution of* (P) *satisfying* (4.4). *Denote by $\bar{y}$ and $\bar{p}$ the state and adjoint state associated with $\bar{u}$. Then there exist Lagrange multipliers $(\bar{q}, \bar{\lambda}) \in L^2(\Omega) \times \mathbb{R}^m$ such that*

$$\bar{\lambda}_j \geq 0,\ 1 \leq j \leq m,\ \textit{and } \bar{\lambda}_j = 0 \qquad \textit{if } G_j(\bar{u}) < 0, \tag{4.5}$$

$$N\bar{u} + \bar{p} + \sum_{j=1}^m \bar{\lambda}_j DG_j(\bar{u}) + \bar{q} = 0; \qquad \langle \bar{q}, u - \bar{u}\rangle \leq 0 \ \ \forall u \in L^2_+(\Omega). \tag{4.6}$$

Since $L^2_+(\Omega)$ is polyhedric in $L^2(\Omega)$ (see, e.g., [18, 28]), $(\mathcal{P})$ is of the form (AP) with

$$X = L^2(\Omega), \quad Y = \mathbb{R}^m, \quad K_X = L^2_+(\Omega), \quad K_Y = \mathbb{R}^m_-.$$

We now discuss the strict semilinearized qualification condition (CQA). We need a notation for the contact set of $\bar{u}$ and its complement (defined up to a null measure set):

$$\Omega_-(\bar{u}) := \{x \in \Omega;\ \bar{q}(x) < 0\}, \qquad \Omega_0(\bar{u}) := \{x \in \Omega;\ \bar{u}(x) = 0\},$$

$$\Omega_+(\bar{u}) := \{x \in \Omega;\ \bar{u}(x) > 0\}.$$

Since $\Omega_-(\bar{u}) \subset \Omega_0(\bar{u})$, we have

$$T_{L^2_+(\Omega)}(\bar{u}) \cap \bar{q}^\perp = \left\{h \in L^2(\Omega);\ h \geq 0 \ \text{ on } \ \Omega_0(\bar{u});\ \ h = 0 \ \text{ on } \ \Omega_-(\bar{u})\right\}.$$

Therefore, the strict qualification condition is identical to the qualification condition of the problem

$$\operatorname{Min}\{F(u) \mid u \in L^2(\Omega),\ u = 0 \text{ on } \Omega_-(\bar{u}),\ G_j(u) = 0 \text{ for } j \in J_-\}.$$

Lemma 4.3. *Let $\bar{u} \in F(\mathcal{P})$, with associated Lagrange multiplier $(\bar{q}, \bar{\lambda})$. Then the three conditions below are equivalent:*

(i) *The strict semilinearized qualification condition* (CQA) *is satisfied.*

(ii) *The following conditions hold:*

$$\left\{\begin{array}{l} \text{(a)}\ \ \{DG_i(\bar{u})h;\ \ i \in J_-;\ \ h \in (T_{L^2_+(\Omega)}(\bar{u}) \cap \bar{q}^\perp)\} \textit{ is onto,} \\ \text{(b)}\ \ \exists h \in (\mathcal{R}_{L^2_+(\Omega)}(\bar{u}) \cap \bar{q}^\perp);\ DG_i(\bar{u})h = 0, i \in J_-;\ DG_i(\bar{u})h < 0, i \in J_0 \backslash J_-. \end{array}\right. \tag{4.7}$$

(iii) *There exists no $(\tilde{q}, \tilde{\lambda}) \in L^2(\Omega) \times \mathbb{R}^m$ with $\tilde{\lambda} \neq 0$, satisfying the following relations:*

$$\left\{\begin{array}{l} \text{(a)}\ \ \tilde{\lambda}_i = 0, \qquad i \in J_+; \qquad \tilde{\lambda}_i \geq 0, \qquad i \in J_0 \backslash J_-, \\ \text{(b)}\ \ \tilde{q}(x) = 0 \quad \textit{on } \Omega_+(\bar{u}), \quad \tilde{q}(x) \leq 0 \quad \textit{on } \Omega_0(\bar{u}) \backslash \Omega_-(\bar{u}); \\ \text{(c)}\ \ \tilde{q} + \sum_{1 \leq i \leq m} \tilde{\lambda}_i DG_i(\bar{u}) = 0. \end{array}\right. \tag{4.8}$$

Proof. By the definition, (CQA) holds iff for any $z \in \mathbb{R}^m$ close enough to 0, $\exists$ $h \in \left(L^2_+(\Omega) - \bar{u}\right) \cap \bar{q}^\perp$ satisfying the following relations:

$$\text{(i)}\ \ DG_i(\bar{u})h = z_i,\ i \in J_-; \quad \text{(ii)}\ \ DG_i(\bar{u})h \leq z_i,\ i \in J_0 \backslash J_-. \tag{4.9}$$

It follows from (4.9(i)) that the set

$$\{DG_i(\bar{u})h;\ \ i \in J_-;\ \ h \in (\mathcal{R}_{L^2_+(\Omega)}(\bar{u}) \cap \bar{q}^\perp)\} \text{ is onto.}$$

This implies that (4.7(i)) is a necessary condition for (CQA). Then taking $z_i = 0$, $i \in J_-$, and $z_i < 0$, $i \in J_0 \backslash J_-$, we deduce that (4.7) is a necessary condition for (CQA), i.e., (CQA) $\Rightarrow$ (4.7). We end the proof by showing that (4.7)$\Rightarrow$(4.8)$\Rightarrow$(CQA).

Assume that (CQA) does not hold. Then the convex cone

$$E := \left\{DG(\bar{u})h - z;\ h \in \mathcal{R}_{L^2_+(\Omega)}(\bar{u}) \cap \bar{q}^\perp;\ \ z_i \leq 0, i \in J_0;\ z_i = 0, i \in J_-\right\}$$

is not equal to $\mathbb{R}^m$. Since the latter is a finite dimensional space, the closure of E is not equal to $\mathbb{R}^m$. By the Hahn–Banach theorem, since E is a cone, $\exists\, \tilde\lambda \in \mathbb{R}^m$, $\tilde\lambda \neq 0$, such that $\langle \tilde\lambda, y\rangle \geq 0\ \forall y \in E$. It follows that (4.8(i)) holds, while $\tilde q$ defined by (4.8(iii)) is such that

$$-\langle \tilde\lambda, DG(\bar u)h\rangle = \int_\Omega \tilde q(x)h(x)dx \leq 0 \quad \forall h \in (L^2_+(\Omega) - \bar u) \cap \bar q^\perp. \tag{4.10}$$

Since the polar of the intersection of two closed convex cones is the closure of the sum of their polar cones, we have that

$$\tilde q \in \left(\left(L^2_+(\Omega) - \bar u\right) \cap \bar q^\perp\right)^- = \overline{(L^2_+(\Omega) - \bar u)^- + \mathbb{R}\bar q}.$$

Relation (4.8(ii)) follows.

Finally, suppose that (4.7) holds, but (4.8) does not hold. Let $(\tilde q, \tilde\lambda)$ satisfy (4.8). Then (4.10) holds. It follows that for each $h \in \mathcal{R}_{L^2_+(\Omega)}(\bar u) \cap \bar q^\perp$,

$$0 \leq -\int_\Omega \tilde q(x)h(x)dx = \sum_{1\leq i\leq m} \tilde\lambda_i DG_i(\bar u)h.$$

This and (4.7(ii)) imply $\tilde\lambda_i = 0\ \forall i \in J_0 \backslash J_-$. Then, since $L^2_+(\Omega)$ is polyhedric, with (4.7(i)) we obtain $\tilde\lambda_i = 0\ \forall i \in J_-$, in contradiction with the fact that $\tilde\lambda \neq 0$. □

Denote by $L^2(\Omega_+(\bar u))$ the Hilbert space of functions of $L^2(\Omega)$ that are a.e. null outside $\Omega_+(\bar u)$. From the above lemma, we deduce the following corollary, similar to [15, Thm. 5.2].

Corollary 4.4. *A sufficient condition for* (CQA) *is that the restriction of $DG(\bar u)$ to $L^2(\Omega_+(\bar u))$, with image space $\mathbb{R}^m$, is onto.*

We now discuss second-order optimality conditions. Since $Q_0(\cdot)$ is a Legendre form, we have the following result which is an immediate consequence of Theorem 2.10. Note that the assumption, that the Hessians $D^2G(\bar u)$ are weakly continuous, is typically satisfied if G represents state constraints, as will be the case in the examples to be seen later. The expression of the Lagrangian for problem $(\mathcal{P})$ is

$$\mathcal{L}(u, \lambda) := F(u) + \sum_{i=1}^m \lambda_i G_i(u).$$

Theorem 4.5. *Let $\bar u$ be an optimal solution of $(\mathcal{P})$, with associated Lagrange multiplier $(\bar q, \bar\lambda)$, satisfying condition* (CQA). *Assume that the Hessians $D^2G_i(\bar u)$ (for $i = 1, \ldots, m$) are weakly continuous. Then $\bar u$ satisfies the quadratic growth condition iff*

$$D^2_{uu}\mathcal{L}(\bar u, \bar\lambda)(h, h) > 0\ \ \forall h \in C(\bar u), \quad h \neq 0.$$

4.2. Problems with finitely many punctual state constraints. We consider in this subsection the case when the functions G_j, $1 \leq j \leq m$, are defined by

$$G_j(u) = y_u(x_j) - b_j.$$

Here $b \in \mathbb{R}^m$ and $x_j \in \Omega$, $1 \leq j \leq m$, are given. We denote $\bar y := y_{\bar u}$. A simple consequence of Lemma 4.3 follows.

LEMMA 4.6. *Assume that* $\Omega_+(\bar{u})$ *has a nonempty interior. Then the strict semi-linearized qualification condition* (CQA) *is satisfied.*

Proof. If the conclusion does not hold, then by Lemma 4.3, $\exists\ (\tilde{q}, \tilde{\lambda}) \in L^2(\Omega) \times \mathbb{R}^m$ with $\tilde{\lambda} \neq 0$, satisfying (4.8). It is a classical result (see, e.g., [12]) that $\tilde{q} \in L^2(\Omega) \cap W^{1,s}(\Omega)$, $\forall\ s < n/(n-1)$ and is the unique solution in $W^{1,1}(\Omega)$ of

$$-\Delta\tilde{q} + \phi'_y(x, \bar{y})\tilde{q} = -\textstyle\sum_{1\leq i\leq m} \lambda_i \delta(x_i) \text{ in } \Omega, \quad \tilde{q} = 0 \text{ on } \Gamma. \tag{4.11}$$

Here $\delta(x_i)$ stands for the Dirac measure at point x_i. Since $\tilde{q} = 0$ on the interior of Ω_+, and the latter is nonempty, we have by the unique extension theorem [35] that $\tilde{q} = 0$ over Ω except perhaps at the points x_j. But this implies $\tilde{\lambda} = 0$, in contradiction to the hypothesis. □

We now state the characterization of quadratic growth. By z_h we denote the solution of the linearized equation (4.2) with $y_u = \bar{y}$ and right-hand side h. As a consequence of Theorem 4.5, we have the following.

THEOREM 4.7. *Let* $\bar{u}$ *be a feasible point of* $(\mathcal{P})$, *with associated Lagrange multiplier* $(\bar{q}, \bar{\lambda})$, *and assume that the interior of* Ω_+ *is nonempty. Then* $\bar{u}$ *satisfies the quadratic growth condition iff* $\exists\ \bar{p} \in W^{1,s}(\Omega)\ \forall\ s < n/(n-1)$, *such that*

$$\bar{\lambda}_j \geq 0,\ 1 \leq j \leq m,\ \textit{and } \bar{\lambda}_j = 0 \qquad \textit{if } G_j(\bar{u}) < 0, \tag{4.12}$$

$$-\Delta\bar{p} + \phi'_y(x, \bar{y})\bar{p} = \bar{y} - y_d + \sum_{j=1}^{m} \bar{\lambda}_j \delta(x_j) \ \ \textit{in } \Omega, \quad \bar{p} = 0 \ \ \textit{on } \Gamma, \tag{4.13}$$

$$\int_\Omega (N\bar{u}(x) + \bar{p}(x))(u(x) - \bar{u}(x))\,dx \geq 0 \ \ \forall u \in L^2_+(\Omega), \tag{4.14}$$

and such that, $\forall\ h \in C(\bar{u})$, $h \neq 0$, *and* z_h *solution of* (4.2) *(in which* $\bar{y} = y_{\bar{u}}$*),*

$$\int_\Omega (Nh(x)^2 + z_h^2(x)^2 - \bar{p}(x)\phi''_y(x, \bar{y})z_h(x)^2)\,dx > 0. \tag{4.15}$$

We now discuss sensitivity of the solution of the optimal control problem with respect to the target y_d. Therefore we denote

$$F(u, y_d) := \frac{1}{2}\int_\Omega (y_u(x) - y_d(x))^2\,dx + \frac{N}{2}\int_\Omega u(x)^2\,dx.$$

Consider a target path, where $t \geq 0$,

$$y_d(t) = y_{d0} + ty_{d1} + \frac{t^2}{2}y_{d2} + o(t^2).$$

Note that

$$D_{y_d}F(u, y_{d0})y_{d1} = -\int_\Omega (\bar{y} - y_{d0})(x)y_{d1}(x)dx.$$

The subproblems to be considered here, corresponding to (LP) and (SP), are

$$\begin{array}{c}\text{Min}_{h\in L^2(\Omega)} \displaystyle\int_\Omega (N\bar{u} + \bar{p})(x)h(x)dx - \int_\Omega (\bar{y} - y_{d0})(x)y_{d1}(x)dx \\ \text{s.t. } h \geq 0 \text{ on } \Omega_0(\bar{u}); z_h(x_i) \leq 0,\ i \in J_0,\end{array} \tag{$\overline{\text{LP}}$}$$

and, denoting by z_h the solution of (4.2) (in which $y_u = \bar{y}$),

$$(\overline{\mathrm{SP}}) \qquad \mathrm{Min}_{h\in S(\overline{\mathrm{LP}})} D^2F(\bar{u}, y_{d0})((h, y_{d1}), (h, y_{d1})) - \int_\Omega (\bar{y} - y_{d0})(x) y_{d2}(x)dx.$$

(An expression of the Hessian of F in term of $\bar{p}$ and z_h is given in [6].)

Theorem 4.8. *Assume that $\bar{u}$ is the unique solution of* (AP) *and satisfies* (CQA) *as well as the second-order sufficient optimality condition* (4.15)*. Then*

(a) *The following expansion for the value function of* $(AP_{u(t)})$ *holds:*

$$\mathrm{val}\,(\mathrm{AP}_{u(t)}) = \mathrm{val}\,(\mathrm{AP}_{\bar{u}}) + t\mathrm{val}\,(\overline{\mathrm{LP}}) + \frac{t^2}{2}\mathrm{val}\,(\overline{\mathrm{SP}}) + o(t^2). \tag{4.16}$$

(b) *Let $u(t)$ be a path of $o(t^2)$-solutions. Then one has $u(t) = \bar{u} + O(t)$. Any weak limit-point of $t^{-1}(u(t) - \bar{u})$ is a strong limit-point and is a solution of* $(\overline{\mathrm{SP}})$*. In particular, if* $(\overline{\mathrm{SP}})$ *has a unique solution $\bar{h}$, then $u(t) = \bar{u} + t\bar{h} + o(t)$.*

Proof. It is easy to check that the solutions of the perturbed problem are uniformly bounded and that they strongly converge in $L^2(\Omega)$ to $\bar{u}$; see, e.g., [6]. In addition the Hessian of the Lagrangian, which is equal to the Hessian of the cost, is a Legendre form. Therefore the conclusion is a consequence of Theorem 3.2. □

4.3. Problems with integral state constraints. We consider in this subsection the case when the functions $G_j(u)$, $1 \le j \le m$, are defined by

$$G_j(u) = \int_\Omega g_j(y_u(x), x)\, dx.$$

The functions $g_j(u)$ are assumed to be twice continuously differentiable functions $\mathbb{R} \times \bar{\Omega} \to \mathbb{R}$. Then $G(\cdot)$ is itself a C^2 mapping. We know that the derivative of $u \to G_j(u)$, viewed as a function $L^2(\Omega) \to \mathbb{R}$, is $p_j(u) \in H^2(\Omega)$ solution of

$$-\Delta p_j + \phi'_y(x, y_u)p_j = D_y g_j(y_u(x), x) \ \text{ in } \Omega, \quad p_j = 0 \ \text{ on } \Gamma.$$

A simple consequence of Lemma 4.3 follows.

Lemma 4.9. *The strict qualification condition* (CQA) *is satisfied iff the following system has no solution $(\tilde{q}, \tilde{\lambda}) \in W^{1,s}(\Omega) \times \mathbb{R}^m$:*

$$\begin{array}{l} -\Delta\tilde{q} + \phi'_y(x, \bar{y})\tilde{q} = -\sum_{1\le i\le m} \lambda_i D_y g_j(\bar{y}(x), x) \quad in\ \Omega, \quad \tilde{q} = 0 \quad on\ \Gamma, \\ \tilde{q}(x) = 0 \qquad on\ \Omega_+(\bar{u}), \qquad \tilde{q}(x) \ge 0 \quad on\ \Omega_0(\bar{u})\backslash\Omega_-(\bar{u}), \\ \lambda \ne 0; \quad \lambda_i = 0, i \in J_+, \quad \lambda_i \ge 0, \qquad i \in J_0\backslash J_-. \end{array} \tag{4.17}$$

Let us give an example of such integral constraints for which condition (CQA) can be checked. Let $b \in \mathbb{R}^m$ and $a_j \in C(\overline{\Omega})$ for $1 \le j \le m$, with $a_j(x)$ of constant sign over its support $\Omega_j := \mathrm{supp}(a_j)$. Assume that these supports satisfy the following geometric relation:

$$\Omega_i \cap \Omega_j = \emptyset \text{ for } i \ne j; \quad \Omega \setminus (\cup_{1\le j\le m}\Omega_j) \ \text{ is connected.} \tag{4.18}$$

We consider the case when

$$g_j(u) := \int_\Omega a_j(x)y_u(x)dx - b_j.$$

We also assume the following:

$$\exists\ \Omega_*, \text{ open subset of } \Omega_+(u), \text{ s.t. } \Omega_* \cap \Omega_j = \emptyset, \qquad 1 \leq j \leq m. \tag{4.19}$$

LEMMA 4.10. *Under the above hypotheses, the strict qualification condition* (CQA) *is satisfied.*

Proof. If the conclusion does not hold, then $\exists\ (\tilde{q}, \tilde{\lambda})$ satisfying the condition of Lemma 4.9, and in particular

$$-\Delta\tilde{q} + \phi'_y(x,\bar{y})\tilde{q} = -\sum_{j \in J_0 \cup J_-} \tilde{\lambda}_j a_j \ \text{ in } \Omega, \quad \tilde{q} = 0 \ \text{ on } \Gamma, \tag{4.20}$$

as well as $\tilde{q} = 0$ on Ω_*. Set $A := \Omega \setminus (\cup_{1 \leq j \leq m} \Omega_j)$. Then A is a connected open set that contains Ω_*. Since Ω_* is open, and $\tilde{q} = 0$ on A, by the unique extension theorem [35] we obtain $\tilde{q} = 0$ on A, hence on $\partial\Omega_j$, $\forall\ 1 \leq j \leq m$. Let j be such that $\tilde{\lambda}_j \neq 0$. Let A_j be the interior of Ω_j, $1 \leq j \leq m$, and let

$$B_j := \{x \in \Omega;\ \text{dist}(x, A_j) \leq \varepsilon\}.$$

Take $\varepsilon > 0$ so small that $B_j \backslash A_j$ does not intersect Ω_i for $i \neq j$. Then $\tilde{q}$ satisfies

$$-\Delta\tilde{q} + \phi'_y(x, y_u)\tilde{q} = -\tilde{\lambda}_j a_j \ \text{ in } B_j, \quad \tilde{q} = 0 \text{ on } \partial B_j.$$

This equation has a unique solution in $H^1_0(B_j)$. Since a_j is of constant sign, $\tilde{q}$ is nonzero over the interior of B_j. But this is impossible, since the latter contains a nonempty open set included in A. □

Whenever (CQA) holds, we can state a characterization of quadratic growth. We omit the statement since it is similar to Theorem 4.2.

5. Conclusion and possible extensions. Our theoretical results extend those in [6], which discuss problems with polyhedric control constraints only. We were able to give an application of these results for control and state constrained optimal control problems when the number of state constraints is finite.

For technical reasons we discussed only the case when the space dimension n is less than or equal to 3. Extension of these results in the case $n > 3$ seems possible by combining the technique of this paper with the two norms approach [6, 25]. The latter would also allow our results to extend to the case of boundary control or to problems with a parabolic state equation.

It seems also possible to extend our results to the case when K_Y is not a polyhedron, taking advantage of the results in [9]. For instance, the set of semi definite positive matrices is a closed convex set that satisfies hypothesis (ii) of Theorem 2.7. On the other hand, the case of a punctual state constraint at every point of the domain Ω seems out of reach, since the strict qualification condition is probably not satisfied in that case.

Acknowledgments. The authors thank the two referees for their useful remarks.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces.* Academic Press, New York, 1975.

[2] S. AGMON, A. DOUGLIS, AND L. NIRENBERG, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* I, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[3] J. J. Alibert and J. P. Raymond, *A Lagrange multiplier theorem for control problems with state constraints*, Numer. Funct. Anal. Optim., 19 (1998), pp. 697–704.

[4] L. Barbet, *Analyse de sensibilité différentielle des solutions optimales d'inéquations variationnelles en dimension infinie*, C. R. Acad. Sci. Paris Sér. I Math., 315 (1992), pp. 1179–1182.

[5] H. Brézis, *Problèmes unilatéraux*, J. Math. Pures Appl., 51 (1972), pp. 1–168.

[6] J. F. Bonnans, *Second order analysis for control constrained optimal control problems of semilinear elliptic systems*, J. Appl. Math. Optim., 38 (1998), pp. 303–325.

[7] F. Bonnans and E. Casas, *An extension of Pontryagin's principle for state-constrained optimal control of semilinear elliptic equations and variational inequalities*, SIAM J. Control Optim., 33 (1995), pp. 274–298.

[8] J. F. Bonnans and R. Cominetti, *Perturbed optimization in Banach spaces* I*: A general theory based on a weak directional constraint qualification*, SIAM J. Control Optim., 34 (1996), pp. 1151–1171.

[9] J. F. Bonnans, R. Cominetti, and A. Shapiro, *Second-order optimality conditions based on parabolic second-order tangent sets*, SIAM J. Optim., 9 (1999), pp. 466–492.

[10] J. F. Bonnans and A. Shapiro, *Optimization problems with perturbations: A guided tour*, SIAM Rev., 40 (1998), pp. 228–264.

[11] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, manuscript.

[12] E. Casas, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309–1318.

[13] E. Casas, J. P. Raymond, and H. Zidani, *Optimal control problem governed by semilinear elliptic equations with integral control constraints and pointwise state constraints*, in International Conference on Control and Estimations of Distributed Parameter Systems, Internat. Ser. Numer. Math., W. Desch and F. Kappel, eds., Birkhaüser-Verlag, Basel, Switzerland, 1997.

[14] E. Casas, F. Tröltzsch, and A. Unger, *second-order sufficient optimality conditions for a nonlinear elliptic boundary control problem*, Z. Anal. Anwendungen, 15 (1996), pp. 687–707.

[15] E. Casas and F. Tröltzsch, *Second-order necessary optimality conditions for some state-constrained control problems of semilinear elliptic equations*, Appl. Math. Optim., 39 (1999), pp. 211–227.

[16] R. Cominetti, *Metric regularity, tangent sets and second-order optimality conditions*, Appl. Math. Optim., 21 (1990), pp. 265–287.

[17] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.

[18] A. Haraux, *How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities*, J. Math. Soc. Japan, 29 (1977), pp. 615–631.

[19] A. D. Ioffe and V. M. Tihomirov, *Theory of Extremal Problems*, North–Holland, Amsterdam, 1979.

[20] A. D. Ioffe, *On some recent developments in the theory of second-order optimality conditions*, in Optimization, S. Dolecki ed., Lecture Notes in Math. 1405, Springer-Verlag, Berlin, 1989, pp. 55–68.

[21] A. D. Ioffe, *Variational analysis of a composite function: A formula for the lower second-order epi-derivative*, J. Math. Anal. Appl., 160 (1991), pp. 379–405.

[22] H. Kawasaki, *An envelope-like effect of infinitely many inequality constraints on second-order necessary conditions for minimization problems*, Math. Programming, 41 (1988), pp. 73–96.

[23] J. L. Lions, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.

[24] K. Malanowski, *Second order conditions and constraint qualifications in stability and sensitivity analysis of solutions to optimization problems in Hilbert spaces*, Appl. Math. Optim., 25 (1992), pp. 51–79.

[25] K. Malanowski, *Two-norm approach in stability and sensitivity analysis of optimization and optimal control problems*, Adv. Math. Sci. Appl., 2 (1993), pp. 397–443.

[26] K. Malanowski, *Stability and sensitivity of solutions to nonlinear optimal control problems*, Appl. Math. Optim., 32 (1995), pp. 111–141.

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

[28] F. Mignot, *Contrôle dans les inéquations variationelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 25–39.

[29] Z. Páles and V. Zeidan, *First-and second-order necessary conditions for control problems*

with constraints, Trans. Amer. Math. Soc., 346 (1994), pp. 421–453.

[30] Z. Páles and V. Zeidan, *Nonsmooth optimum problems with constraints*, SIAM J. Control Optim., 32 (1994), pp. 1476–1502.

[31] J. P. Penot, *Optimality conditions in mathematical programming and composite optimization*, Math. Programming, 67 (1994), pp. 225–245.

[32] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.

[33] R. T. Rockafellar, *Conjugate Duality and Optimization*, CBMS–NSF Regional Conf. Ser. in Appl. Math., SIAM, Philadelphia, 1974.

[34] A. Shapiro, *Perturbation analysis of optimization problems in Banach spaces*, Numer. Funct. Anal. Optim., 13 (1992), pp. 97–116.

[35] J. C. Saut and B. Scheurer, *Sur l'unicité du problème de Cauchy et le prolongement unique pour des équations elliptiques à coefficients non localement bornés*, J. Differential Equations, 43 (1982), pp. 28–43.

[36] J. Sokolowski, *Sensitivity analysis of control constrained optimal control problems for distributed parameter systems*, SIAM J. Control Optim., 25 (1987), pp. 1542–1556.

[37] J. Zowe and S. Kurcyusz, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Optim., 5 (1979), pp. 49–62.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 6, pp. 1742–1750

AVERAGE $\mathcal{H}_2$ PERFORMANCE AND MAXIMAL PARAMETER PERTURBATION RADIUS FOR UNCERTAIN SYSTEMS*

KE-YOU ZHAO[†], MICHAEL J. GRIMBLE[‡], AND JAKOB STOUSTRUP[§]

Abstract. In this paper, methods are presented for calculating the maximal parameter perturbation bounds under $\mathcal{H}_2$ performance constraints for a family of uncertain systems and for calculating the average $\mathcal{H}_2$ performance under such parameter variations. The uncertain systems are described by state space models with nonlinear (polynomial) dependencies on real uncertain parameters. All results obtained are based on necessary and sufficient conditions. As a special virtue of the approach, the proposed algorithms for stability analysis and for performance analysis turn out to have exactly the same algebraic structure. An example illustrates the results and the algorithms.

Key words. $\mathcal{H}_2$ performance, stability, robustness, nonlinear perturbation, state space methods

AMS subject classifications. 93C73, 93D09

PII. S0363012996301725

1. Introduction.

Robust performance analysis for uncertain control systems, which is now receiving a great deal of attention (see [4, 9] and references therein), is a relatively new area in comparison with robust stability analysis. For linear time-invariant systems, the $\mathcal{H}_2$ performance metric arises naturally in a number of different physically meaningful situations; see [4, 6, 3]. The $\mathcal{H}_2$ performance of a linear time-invariant system is measured via the $\mathcal{H}_2$ norm of its transfer matrix. As long as this $\mathcal{H}_2$ norm is less than a given upper bound, the design can stop, and there is usually no need to seek the minimal norm and/or this might not be advisable due to robustness considerations.

Suppose now that the $\mathcal{H}_2$ norm of a nominal (stable) system is less than a given upper bound. Then the question is whether the norm is still less than this upper bound after suffering a parameter perturbation, or alternatively, how to find the maximal domain for perturbation parameters under stability and $\mathcal{H}_2$ norm constraints.

This paper will consider the latter problem and calculate the maximal perturbation interval or radius in perturbation parameter space. The results obtained are not only sufficient but also necessary. The paper is different from most previously published papers which deal with a fixed parameter domain and affine perturbations. One of our motivations comes from [4], which computed the supremum of the $\mathcal{H}_2$ norm in the case of an affine perturbation with perturbation parameter $q \in [0, 1]$. Also in similarity with that paper we shall compute not only the maximal perturbation radii subject to stability and performance constraints but also the average performance over a fixed perturbation set.

*Received by the editors April 7, 1996; accepted for publication (in revised form) February 26, 1999; published electronically September 30, 1999.

http://www.siam.org/journals/sicon/37-6/30172.html

[†] Department of Electrical Engineering, Qingdao University, Qingdao, 266071, People's Republic of China. The research of this author was supported by the National Education Commission and by the Shandong Provincial Science Foundation.

[‡] Industrial Control Centre, Department of Electronic and Electrical Engineering, University of Strathclyde, Graham Hills Building, 50 George Street, Glasgow G1 1QE, UK (m.grimble@icu.strath.ac.uk).

[§] Department of Control Engineering, Aalborg University, Fr. Bajersvej 7C, 9220 Aalborg, Denmark (jakob@control.auc.dk). The research of this author was supported by Danish Technical Research Council grant 95-00765.

The notation used throughout the paper is as follows. Denote the real number set by $\mathbf{R}$ and the complex plane (the complex open left half plane) by $\mathbf{C}$ ($\mathbf{C}^-$). Let **cs**: $\mathbf{R^{m\times n}} \to \mathbf{R^{mn}}$ be the column stacking operator on a matrix and $\oplus$: $\mathbf{R^{n\times n}} \times \mathbf{R^{m\times m}} \to \mathbf{R^{mn\times mn}}$ be the standard matrix Kronecker sum defined in [2]. Finally, let $\lambda_k(\cdot)$ be the kth eigenvalue of a square matrix.

2. Problem formulation. Consider a linear time-invariant system

$$\begin{aligned} \dot{x}(t) &= A(q)x(t) + B(q)w(t), \\ z(t) &= C(q)x(t), \end{aligned} \tag{2.1}$$

where $x \in \mathbf{R^n}$, $w \in \mathbf{R^m}$, and $z \in \mathbf{R^p}$ are state, disturbance, and performance vectors, respectively; $A(q)$, $B(q)$, and $C(q)$ are (of compatible dimension) continuous matrix functions of the perturbation parameter vector $q = [q_1, q_2, \ldots, q_l]^T \in \mathbf{R^l}$. The transfer function matrix from w to z can be expressed as $T(s,q) = C(q)(sI - A(q))^{-1}B(q)$. A square constant matrix is called stable if all of its eigenvalues lie in $\mathbf{C}^-$. The corresponding transfer function $T(s,q)$ is said to be stable for a given q if $A(q)$ is stable and its $\mathcal{H}_2$ norm is defined by

$$\|T(s,q)\|_2 \doteq \left\{ \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathbf{trace}\,[T(j\omega,q)T^*(j\omega,q)]\, d\omega \right\}^{1/2}, \tag{2.2}$$

where $T^*(s,q) \doteq T'(-s,q)$ and $(\cdot)'$ denotes transpose.

We shall make the following standing assumptions on the nominal system—given by $(A(0), B(0), C(0))$—and on the parameter dependence:

AS1. $A(0)$ is stable.

AS2. $\|T(s,0)\|_2^2 < \gamma$.

AS3. The system matrices may be parameterized as

$$\begin{aligned} A(q) &\doteq A_0 + qA_1 + \cdots + q^{m_1}A_{m_1}, \\ B(q) &\doteq B_0 + qB_1 + \cdots + q^{m_2}B_{m_2}, \\ C(q) &\doteq C_0 + qC_1 + \cdots + q^{m_3}C_{m_3}, \end{aligned}$$

where all of A_k, B_k, and C_k are given constant matrices.

Here, γ is a given positive constant which reflects the tolerance of the system as measured by the $\mathcal{H}_2$ performance (for instance, an acceptable output variance of (2.1) to a white noise signal w). The goal is to find the "maximal domain" in $\mathbf{R^l}$ so that $\|T(s,q)\|_2^2 < \gamma$ for every q in the domain. A prerequisite for doing this is that $A(q)$ must be stable for all q in this domain. This means that the robust stability analysis must be completed first (see relevant results in [1, 5, 7, 8]).

The relevant problems will in this paper only be considered for the single parameter case, i.e., $l = 1$. The two parameter case, $l = 2$, can at least in principle be handled by the approach described below applying a line search. However, for medium or large scale problems, computational issues will limit the practical use of this.

To formulate the problem of determining the maximal perturbation radius, first define

$$r_s^- \doteq \inf\{r < 0 : A(q) \text{ is stable } \forall q \in (r,0)\}, \tag{2.3}$$

$$r_s^+ \doteq \sup\{r > 0 : A(q) \text{ is stable } \forall q \in (0,r)\}, \tag{2.4}$$

$$r_2^- \doteq \inf\{r < 0 : A(q) \text{ is stable and } \|T(s,q)\|_2^2 < \gamma\ \forall q \in (r,0)\}, \tag{2.5}$$

$$r_2^+ \doteq \sup\{r > 0 : A(q) \text{ is stable and } \|T(s,q)\|_2^2 < \gamma\ \forall q \in (0,r)\}. \tag{2.6}$$

Then (r_s^-, r_s^+) is the maximal perturbation interval of q while keeping the stability of $A(q)$, and (r_2^-, r_2^+) is the maximal perturbation interval of q while keeping $\|T(s,q)\|_2^2 < \gamma$.

PROBLEM 2.1. *Suppose that system* (2.1) *satisfies* AS1, AS2, *and* AS3.

(a) *Find* r_s^- *and* r_s^+.

(b) *Find* r_2^- *and* r_2^+.

Remark 2.2. Obviously, $(r_2^-, r_2^+) \subset (r_s^-, r_s^+)$.

The notion of average performance can be defined in terms of the following problem.

PROBLEM 2.3. *Suppose that system* (2.1) *satisfies* AS1, AS2, *and* AS3 *and that two numbers* $\underline{q}$ *and* $\overline{q}$ *are given, where* $r_s^- < \underline{q} < \overline{q} < r_s^+$.

Find $\frac{1}{\overline{q} - \underline{q}} \int_{\underline{q}}^{\overline{q}} \|T(s,q)\|_2^2 \, dq.$

This definition follows the convention in [4]. It can be argued that an alternative problem formulation similar to Problem 2.3 but without the square would be interesting as well. That problem also admits a solution but not one which is as easy to interpret in terms of the problem parameters as the one given for Problem 2.3 below.

Remark 2.4. The integral boundaries of Problem 2.3 have been chosen to be strictly inside the stability interval (not on the closure). This is because, usually, the integral would become unbounded on the stability boundary.

3. Preliminaries.

The main idea in this paper is to transform functions that are rational in the independent variable (the uncertain parameter) into a matrix version of the companion form, utilizing the fact that the "denominator" is based on a matrix valued polynomial map. In what follows, we shall provide a matrix result which will prove useful in this respect.

Let $M(r) = M_0 + rM_1 + \cdots + r^m M_m$, where all of the M_k's are $n \times n$ constant matrices, and $|M_0| \neq 0$ ($|\cdot|$ denotes the determinant). Let

$$r^- \doteq \sup\{r < 0 : |M(r)| = 0\}, \tag{3.1}$$

$$r^+ \doteq \inf\{r > 0 : |M(r)| = 0\} \tag{3.2}$$

be the maximal perturbation bounds for nonsingularity of matrices. By simple operations on the matrix and its determinant (see [8]), it can be shown that

$$r^- = \frac{1}{\lambda_{min}^-(\mathbf{M})}, \tag{3.3}$$

$$r^+ = \frac{1}{\lambda_{max}^+(\mathbf{M})}, \tag{3.4}$$

where $\mathbf{M}$ is an $mn \times mn$ matrix given by

$$\mathbf{M} \doteq - \begin{pmatrix} \mathbf{O} & -\mathbf{I} & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & -\mathbf{I} & \cdots & \mathbf{O} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \cdots & -\mathbf{I} \\ M_0^{-1}M_m & M_0^{-1}M_{m-1} & M_0^{-1}M_{m-2} & \cdots & M_0^{-1}M_1 \end{pmatrix}, \tag{3.5}$$

$\lambda_{min}^-(\cdot)$ stands for the minimal value of the negative real eigenvalues (let $\lambda_{min}^-(\cdot) = 0^-$ if there exist no negative real eigenvalues) and $\lambda_{max}^+(\cdot)$ stands for the maximal value of the positive real eigenvalues (let $\lambda^+_{max}(\cdot) = 0^+$ if no positive real eigenvalues), respectively.

Formulae (3.3) and (3.4) suggest the following algorithm.

ALGORITHM 3.1 (the maximal perturbation bounds for nonsingularity of matrices).

Step 1. *Input M_k, $k = 0, 1, \ldots, m$, where $|M_0| \neq 0$;*

Step 2. *Define* $\mathbf{M}$ *as in* (3.5);

Step 3. *Calculate all the eigenvalues of* $\mathbf{M}$;

Step 4. *Find r^- and r^+ based on* (3.3) *and* (3.4)*, then output.*

Algorithm 3.1 will be one of the cornerstones below in solving Problems 2.1 and 2.3. Algorithm 3.1 is conceptually clear and easy to implement, although, admittedly, the numerical aspects can be quite involved for large scale problems, since the relevant matrices will be of very high order. Hence, the main applications for the results below will be in terms of small or medium scale problems.

The following lemma helps us to transform Problem 2.1(a) into that of the maximal perturbation bounds for the nonsingularity of matrices.

LEMMA 3.2. *Suppose that*

(1) *Q is a singly connected domain in $\mathbf{R}^1$, and $0 \in Q$,*

(2) *$A(0)$ is stable.*

Then $A(q)$ are stable $\forall q \in Q$ if and only if $|A(q) \oplus A(q)| \neq 0 \, \forall q \in Q$.

Proof. Recall the continuity of $A(q)$, $B(q)$, $C(q)$ in q and that

$$\lambda_k(A(q) \oplus A(q)) = \lambda_i(A(q)) + \lambda_j(A(q)), \qquad k = 1, \ldots, n^2;\ i, j = 1, \ldots, n.$$

From this observation the lemma becomes obvious. □

By using Lemma 3.2 it follows that

$$r_s^- = \sup\{q < 0 : |A(q) \oplus A(q)| = 0\} \qquad \text{(scalar case)}, \tag{3.6}$$

$$r_s^+ = \inf\{q > 0 : |A(q) \oplus A(q)| = 0\} \qquad \text{(scalar case)}, \tag{3.7}$$

$$r_s = \inf\{r : |A(q) \oplus A(q)| = 0 \quad \text{for some } q,\ \|q\| \leq r\} \qquad \text{(multiparameter case)}. \tag{3.8}$$

Instead of (2.2) in the frequency domain, use the state space approach to compute

$$\|T(s, q)\|_2^2 = \mathbf{trace}\{C'(q)C(q)Q(q)\},$$

where $Q(q) = Q(q)'$ satisfies

$$A(q)Q(q) + Q(q)A(q)' + B(q)B(q)' = 0.$$

It is easy to show the following compact formula (or see [4]):

$$\|T(s, q)\|_2^2 = -\mathbf{cs}[C'(q)C(q)]' \cdot [A(q) \oplus A(q)]^{-1} \cdot \mathbf{cs}[B(q)B'(q)]. \tag{3.9}$$

Going one step from (3.9), the following result is obtained, which helps transform Problem 2.1(b) into that of the maximal perturbation bounds for nonsingularity of matrices.

LEMMA 3.3. *Suppose that*

(1) *Q is a singly connected domain in $\mathbf{R}^1$, and $0 \in Q$,*

(2) *$A(q)$ are stable $\forall q \in Q$,*

(3) *$\|T(s, 0)\|_2^2 < \gamma$.*

Then $\|T(s,q)\|_2^2 < \gamma \ \forall\ q \in Q$ *if and only if* $|\mathbf{M}_\gamma(q)| \neq 0 \, \forall q \in Q$, *where*

$$\mathbf{M}_\gamma(q) \doteq A(q) \oplus A(q) + \frac{1}{\gamma}\mathbf{cs}[B(q)B'(q)] \cdot \mathbf{cs}[(C'(q)C(q)]'. \tag{3.10}$$

Proof. $\|T(s,q)\|_2^2 < \gamma \ \forall q \in Q$
$\Leftrightarrow \gamma + \mathbf{cs}[C'(q)C(q)]' \cdot [A(q) \oplus A(q)]^{-1} \cdot \mathbf{cs}[B(q)B'(q)] > 0 \, \forall q \in Q$ (from (3.9));
$\Leftrightarrow |\gamma I + [A(q) \oplus A(q)]^{-1} \cdot \mathbf{cs}[B(q)B'(q)] \cdot \mathbf{cs}[C'(q)C(q)]'| > 0 \, \forall q \in Q$ (use equality $|\gamma I + XY| = |\gamma I + YX|$);
$\Leftrightarrow |\gamma[A(q) \oplus A(q)]^{-1}| \cdot |\mathbf{M}_\gamma(q)| > 0 \, \forall q \in Q$ (from (3.10));
$\Leftrightarrow |\mathbf{M}_\gamma(q)| \neq 0 \, \forall q \in Q$ (due to the continuity of $A(q), B(q), C(q)$ to q, and Lemma 3.2).

The rest of the proof is trivial and thus omitted. □

By using Lemma 3.3 we obtain the following formulae:

$$r_2^- = \sup\{q \in (r_s^-, 0) : |\mathbf{M}_\gamma(q)| = 0\} \qquad \text{(scalar case)}, \tag{3.11}$$

$$r_2^+ = \inf\{q \in (0, r_s^+) : |\mathbf{M}_\gamma(q)| = 0\} \qquad \text{(scalar case)}, \tag{3.12}$$

$$r_2 = \inf\{r : r \leq r_s \text{ and } |\mathbf{M}_\gamma(q)| = 0 \quad \text{for some } q, \|q\| \leq r\} \qquad \text{(multiparameter case)}. \tag{3.13}$$

In section 2 we presented two types of problems. One is the maximal perturbation bounds for system stability; the other is the maximal perturbation bounds for system performance. Lemmas 3.2 and 3.3 help us to transform these two into the maximal perturbation bounds for nonsingularity of matrices. This means that the resulting algorithms will be similar in spirit.

4. Maximal stability and performance radii.

This section will describe the main formulae and algorithms.

By using matrix multiplication and the expressions of $A(q)$, $B(q)$, $C(q)$ in Problem 2.1, it can be seen that

$$A(q) \oplus A(q) = \mathbf{A}_0 + q\mathbf{A}_1 + \cdots + q^{m_1}\mathbf{A}_{m_1}, \tag{4.1}$$

$$\mathbf{cs}[B(q)B'(q)] = \mathbf{b}_0 + q\mathbf{b}_1 + \cdots + q^{2m_2}\mathbf{b}_{2m_2}, \tag{4.2}$$

$$\mathbf{cs}[C'(q)C(q)] = \mathbf{c}_0 + q\mathbf{c}_1 + \cdots + q^{2m_3}\mathbf{c}_{2m_3}, \tag{4.3}$$

where

$$\mathbf{A}_k = A_k \oplus A_k, \ k = 0, 1, \ldots, m_1,$$
$$\mathbf{b}_0 = \mathbf{cs}\left[B_0 B_0'\right], \ \ldots, \mathbf{b}_k = \mathbf{cs}\left[\sum_{i+j=k} B_i B_j'\right], \ldots, \mathbf{b}_{2m_2} = \mathbf{cs}\left[B_{m_2} B_{m_2}'\right],$$
$$\mathbf{c}_0 = \mathbf{cs}\left[C_0' C_0\right], \ \ldots, \mathbf{c}_k = \mathbf{cs}\left[\sum_{i+j=k} C_i' C_j\right], \ldots, \mathbf{c}_{2m_3} = \mathbf{cs}\left[C_{m_3}' C_{m_3}\right].$$

Substituting the above expressions for $A(q)$, $B(q)$, and $C(q)$ in (3.10), it can be written then as

$$\mathbf{M}_\gamma(q) = M_{0\gamma} + qM_{1\gamma} + \cdots + q^m M_{m\gamma}, \tag{4.4}$$

where $m = \max\{m_1, 2(m_2 + m_3)\}$ and

$$M_{0\gamma} = (A_0 \oplus A_0) + \frac{1}{\gamma}\mathbf{cs}\left[B_0 B_0'\right] \cdot \mathbf{cs}\left[C_0' C_0\right]', \tag{4.5}$$

and all of other $M_{k\gamma}$ (the detailed expressions are omitted) depend on $\mathbf{A}_i$, $\mathbf{b}_j$, and $\mathbf{c}_k$ in a similar fashion.

By recalling Algorithm 3.1 and using (3.6), (3.7), and (4.1), the following result is obtained.

THEOREM 4.1. *Assume that the system* (2.1) *satisfies* AS1, AS2, *and* AS3. *Then the following two statements are equivalent:*

(1) *system* (2.1) *is stable* $\forall\ |q| < \delta$*,*

(2) $\min\{-r_s^-, r_s^+\} > \delta$*.*

To compute the maximal perturbation stability bounds, we can devise the following algorithm from the above results.

ALGORITHM 4.2 (the maximal perturbation bounds for Problem 2.1(a)).

Step 1. *Input* A_k*,* $k = 0, 1, \dots, m$*, where* A_0 *must be stable*;

Step 2. *Calculate* $\mathbf{A}_k$*,* $k = 0, 1, \dots, m_1$;

Step 3. *Let* $M_k = \mathbf{A}_k$*, recall Algorithm* 3.1*, then compute* r^- *and* r^+;

Step 4. *Let* $r_s^- = r^-$ *and* $r_s^+ = r^+$*, and output.*

From AS2, Lemma 3.3, and (4.5), it can be shown that $|M_{0\gamma}| \neq 0$. By recalling Algorithm 3.1 and using (3.11), (3.12), and (4.4), the following result is obtained.

THEOREM 4.3. *Assume that the system* (2.1) *satisfies* AS1, AS2, *and* AS3. *Then the following two statements are equivalent:*

(1) $\|T(s,q)\|_2 < \gamma\ \forall\ |q| < \delta$*,*

(2) $\min\{-r_s^-, r_s^+\} > \delta$*.*

Similarly, to compute the maximal perturbation performance bounds, we can devise the following algorithm from the above results.

ALGORITHM 4.4 (the maximal perturbation bounds for Problem 2.1(b)).

Step 1. *Input* A_i*,* B_j*, and* C_k*, where we must have AS*1 *and AS*2;

Step 2. *Calculate* $\mathbf{A}_i$*,* $\mathbf{b}_j$*, and* $\mathbf{C}_k$*, and also* m;

Step 3. *Calculate* $M_{k\gamma}$;

Step 4. *Let* $M_k = M_{k\gamma}$*, and recall Algorithm* 3.1 *to get* r^- *and* r^+;

Step 5. *Output* $r_2^- = \max\{r_s^-, r^-\}$*,* $r_2^+ = \min\{r_s^+, r^+\}$*.*

Remark 4.5. Algorithms 4.2 and 4.4 do not need any iteration.

Reference [5] gave the maximal perturbation bounds for Problem 2.1(a) in the simplest case (affinely linear perturbation of a single parameter).

5. Average $\mathcal{H}_2$ performance.

To compute the average performance we follow the line of approach of [4]. In similarity with that approach we shall further assume that $B(q)$ and $C(q)$ are fixed matrices, i.e., we have the following uncertainty structure.

AS4. The system matrices may be parameterized as

$$\begin{aligned}
&A(q) \doteq A_0 + qA_1 + \cdots + q^{m_1} A_{m_1},\\
&B(q) \doteq B_0,\\
&C(q) \doteq C_0.
\end{aligned}$$

This assumption can be lifted at the cost of more complicated expressions. These, though, can be obtained easily for a specific application, for instance, by the use of a symbolic algebra package, and the more general result is straightforward, following the idea below.

We define the following matrix:

$$\mathbf{A} \doteq \begin{pmatrix} \mathbf{O} & -\mathbf{I} & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & -\mathbf{I} & \cdots & \mathbf{O} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \cdots & -\mathbf{I} \\ \mathbf{A}_0^{-1}\mathbf{A}_{m_1} & \mathbf{A}_0^{-1}\mathbf{A}_{m_1-1} & \mathbf{A}_0^{-1}\mathbf{A}_{m_1-2} & \cdots & \mathbf{A}_0^{-1}\mathbf{A}_1 \end{pmatrix},$$

where, as above, $\mathbf{A}_k = A_k \oplus A_k,\ k = 0, 1, \ldots, m_1$. Note that $\mathbf{A}_0$ is invertible due to assumption AS1. Also define

$$\mathbf{B} \doteq \begin{pmatrix} 0 \\ \vdots \\ \mathbf{A}_0^{-1}\mathbf{cs}(B_0 B_0') \end{pmatrix}.$$

Finally, we need

$$\mathbf{C} \doteq \begin{pmatrix} 0 & \ldots & 0 & \mathbf{cs}\,(C_0' C_0)' \end{pmatrix}.$$

With these definitions, we can obtain the following result for the *average* $\mathcal{H}_2$ performance of the parameter dependent system.

THEOREM 5.1. *Assume $A(q)$, $B(q)$, and $C(q)$ are as described in* AS4 *with $A(0)$ stable. Let $\underline{q}$ and $\overline{q}$ be two real numbers satisfying $r_s^- < \underline{q} < \overline{q} < r_s^+$, where r_s^- and r_s^+ are as defined in* (2.3) *and* (2.4). *Then*

$$\frac{1}{\overline{q} - \underline{q}} \int_{\underline{q}}^{\overline{q}} \|T(s,q)\|_2^2 \, dq = -\frac{1}{\overline{q} - \underline{q}} \mathbf{C}\mathbf{A}^{-1} \left(\log(I + \overline{q}\mathbf{A}) - \log(I + \underline{q}\mathbf{A}) \right) \mathbf{B},$$

where $\log(\cdot)$ *denotes the matrix logarithm, i.e., the inverse of the matrix exponential.*

Proof. It is straightforward using (3.9) to show that

$$\|T(s,q)\|_2^2 = -\mathbf{C}(I + q\mathbf{A})^{-1}\mathbf{B}.$$

Hence,

$$\int \|T(s,q)\|_2^2 \, dq = -\mathbf{C} \left(\int (I + q\mathbf{A})^{-1} \, dq \right) \mathbf{B} = -\mathbf{C} \left(\mathbf{A}^{-1} \log(I + q\mathbf{A}) \right) \mathbf{B}. \tag{5.1}$$

The last equality holds whenever the argument of the logarithm is a nonsingular matrix. This condition, however, is fulfilled in any open subset of (r_s^-, r_s^+) due to (3.6) and (3.7). □

In certain nongeneric cases (where controllability or observability is lost), it might make sense to extend the calculation of average performance to the boundaries of stability. In that case, the integral in (5.1) becomes more involved. Indeed, let T be a nonsingular matrix, such that

$$T^{-1}\mathbf{A}T = \begin{bmatrix} \tilde{\mathbf{A}} & 0 \\ 0 & \mathbf{A}_0 \end{bmatrix},$$

where $\tilde{\mathbf{A}}$ is nonsingular and $\mathbf{A}_0$ is nilpotent of order k. (One possibility is to compose $\tilde{\mathbf{A}}$ and $\mathbf{A}_0$ by Jordan blocks and to choose the columns of T as the corresponding generalized eigenvectors.)

Then it is easy to show that

$$\int \|T(s,q)\|_2^2 \, dq = -\mathbf{C}T \begin{bmatrix} \tilde{\mathbf{A}}^{-1} \log(I + q\tilde{\mathbf{A}}) & 0 \\ 0 & qI + \sum_{i=1}^{k-1} (-1)^i \frac{q^{i+1}}{i+1} \mathbf{A}_0^i \end{bmatrix} T^{-1}\mathbf{B}.$$

6. Example. An example with a single perturbation parameter is cited below. Let

$$A(q) = \begin{bmatrix} -2 & 1 \\ 0 & -1.5 \end{bmatrix} + q \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + q^2 \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + q^3 \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix},$$

$$B(q) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + q \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}, \; C(q) = \begin{bmatrix} 1 & 1 \end{bmatrix}.$$

It is easy to show that

$$A_0 = \begin{bmatrix} -2 & 1 \\ 0 & -1.5 \end{bmatrix}$$

is stable, that

$$T(s,0) = \begin{bmatrix} \frac{1}{s+2} & \frac{s+3}{(s+2)(s+1.5)} \end{bmatrix},$$

and that $\|T(s,0)\|_2^2 \approx 0.8214 < 1 = \gamma$. In this example it may be shown that

$$A(q) \oplus A(q) = \begin{bmatrix} -4 & 1 & 1 & 0 \\ 0 & -3.5 & 0 & 1 \\ 0 & 0 & -3.5 & 1 \\ 0 & 0 & 0 & -3 \end{bmatrix} + q \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$
$$+ q^2 \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} + q^3 \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix},$$

$$\mathbf{cs}[B(q)B'(q)] = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} + q \begin{bmatrix} 2 \\ 1 \\ 1 \\ 4 \end{bmatrix} + q^2 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 5 \end{bmatrix},$$

and

$$\mathbf{cs}[C'(q)C(q)] = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Furthermore,

$$\mathbf{M}_\gamma(q) = \begin{bmatrix} -3 & 2 & 2 & 1 \\ 0 & -3.5 & 0 & 1 \\ 0 & 0 & -3.5 & 1 \\ 1 & 1 & 1 & -2 \end{bmatrix} + q \begin{bmatrix} 2 & 3 & 3 & 2 \\ 1 & 1 & 1 & 2 \\ 1 & 1 & 1 & 2 \\ 4 & 4 & 4 & 4 \end{bmatrix}$$
$$+ q^2 \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & 1 & 1 \\ 2 & 1 & 1 & 1 \\ 5 & 6 & 6 & 5 \end{bmatrix} + q^3 \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

Finally, $(r_s^-, r_s^+) = (-1.6710, 0.7683)$ can be calculated, which shows that the family $A(q)$ is stable $\forall\ q \in (-1.6710, 0.7683)$, and $(r_2^-, r_2^+) = (-1.5670, 0.0442)$, meaning that $\|T(s,q)\|_2^2 < 1\ \forall\ q \in (-1.5670, 0.0442)$. These two intervals are furthermore the largest intervals with these properties.

If now, in compliance with Assumption AS4, we fix the input matrix

$$B(q) \equiv \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

we obtain a larger performance interval: $\left(r_2^-, r_2^+\right) = (-1.6668, 0.3182)$, where the $\mathcal{H}_2$ norm is bounded by 1. Moreover, in that interval, the average performance can be expressed in terms of

$$\begin{aligned} \sqrt{\int_{r_2^-}^{r_2^+} \|T(s,q)\|_2^2\, dq} &= \sqrt{-\frac{1}{r_2^+ - r_2^-} \mathbf{C}\mathbf{A}^{-1} \left(\log(I + r_2^+\mathbf{A}) - \log(I + r_2^-\mathbf{A})\right)\mathbf{B}}, \\ &\approx 0.7428 \end{aligned}$$

which for this case is in fact better than the nominal performance, $\|T(s,0)\|_2 \approx 0.9063$!

7. Conclusions. Methods for calculating the maximal parameter-perturbation bounds under $\mathcal{H}_2$ performance constraints for a family of systems described by state space models, with nonlinear dependence on real uncertain parameters, have been presented. The results are not conservative as the information of the system structure is used completely. The algorithms as presented here, for robust performance radii and for stability radii, are algebraically similar in nature. Finally, an explicit expression for average $\mathcal{H}_2$ performance for an uncertainty interval also has been presented.

REFERENCES

[1] J. Ackermann and B. Barmish, *Robust schur stability of a polytope of polynomials*, IEEE Trans. Circuits Systems, J. Fund. Theory Appl., 33 (1988), pp. 984–986.

[2] J. Brewer, *Kronecker products and matrix calculus in system theory*, IEEE Trans. Circuits Systems, J. Fund. Theory Appl., 25 (1978), pp. 772–781.

[3] J. Doyle, B. Francis, and A. Tannenbaum, *Feedback Control Theory*, Macmillan, New York, 1991.

[4] J. Friedman, P. Kabamba, and P. Khargonekar, *Worst-case and average $\mathcal{H}_2$ performance analysis against real constant parametric uncertainty*, Automatica, 31 (1995), pp. 649–657.

[5] M. Fu and B. Barmish, *Maximal unidirectional perturbation bounds for stability of polynomials and matrices*, Systems Control Lett., 11 (1988), pp. 173–179.

[6] D. Mustafa, *Non-iterative solution of a $\mathcal{H}_2$-optimal disturbance attenuation problem using static output feedback*, in Proceedings of the European Control Conference, Rome, Italy, 1995, pp. 555–560.

[7] S. Rern, P. Kabamba, and D. Bernstein, *Guardian map approach to robust stability of linear systems with constant real parameter uncertainty*, IEEE Trans. Automat. Control, 39 (1994), pp. 162–164.

[8] K.-Y. Zhao, *Maximal nonlinear perturbations bounds for matrices and polynomials with shifted left-sector stability*, Acta Automat. Sinica, 20 (1994), pp. 227–231.

[9] K. Zhou, P. Khargonekar, J. Stoustrup, and H. Niemann, *Robust performance of systems with structured uncertainties in state space*, Automatica J. IFAC, 31 (1995), pp. 249–255.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 6, pp. 1751–1766

NONCONVEX VARIATIONAL PROBLEMS RELATED TO A HYPERBOLIC EQUATION*

FABIÁN FLORES-BAZÁN† AND STEFANIA PERROTTA‡

Abstract. We first prove a new Lyapunov-type theorem which will yield existence of solutions to nonconvex minimum problems involving some hyperbolic equations on rectangular domains with Darboux boundary conditions. Some problems with obstacle and bang-bang results are also considered.

Key words. Lyapunov theorem, nonconvex minimum problems, relaxed problem, sequentially weakly lower semicontinuous function, biconjugate function, Darboux boundary conditions

AMS subject classifications. 49J20, 49J30, 49J45

PII. S0363012998332299

1. Introduction.

The classical Lyapunov theorem on the range of vector-valued measures, as established in [Ce, Chap. 16], has proved, in the absence of any general theory, to be a very important tool in the study of nonconvex minimization problems within the framework of optimal control [Ne], [Ce], [S1], [R1], [R2], [R3], and [R4] and calculus of variations [Ar], [Ce], [C-Co], [Cr], [Cr-M], [Mar], [C-F], [Ma], [Am-C], [Am-Ma], and [F1]. The setting on which the Lyapunov theorem works is as follows. We start by considering an auxiliary minimization problem (P**) associated to the original one (P). Problem (P**) is such that, if J and J^{**} are the functionals associated to problems (P) and (P**), respectively, then

$$J^{**}(u) \le J(u) \quad \forall \text{ u} \qquad \text{and} \qquad \min_u \; J^{**}(u) = J^{**}(\tilde{u}) \in \mathbb{R}.$$

Then, an application of the Lyapunov theorem shows the existence of a function $\bar{u}$ satisfying $J(\bar{u}) = J^{**}(\tilde{u}) = \min_u \; J^{**}(u)$, showing that $\bar{u}$ is a solution to (P). As a consequence, we could also have min (P**) = min (P). This equality suggests we consider J^{**} as the relaxed functional, defined as the greatest sequentially weakly lower semicontinuous functional majorized by J (e.g., [DM]). However, we actually call (P**) the convexified problem since the relaxation procedure requires some additional assumptions on the integrands which are not necessarily imposed in this paper.

The integral functionals treated under this setting have special structure, namely, they can be split into the sum of two integrals: one depending on the derivative of higher order (e.g., gradient, Laplacian, etc.), and the other depending on the state function itself. The former is the nonconvex part.

A multidimensional version of the Lyapunov theorem has been proven in [Br], although a similar result was already published in [A-L]. This version permits us

*Received by the editors January 7, 1998; accepted for publication (in revised form) November 24, 1998; published electronically September 30, 1999. The research of the first author was partially supported by Conicyt-Chile through FONDECYT 197-0931 and FONDAP-Matemáticas Aplicadas within the subprogram "Optimización." This work started during a visit of the second author to the Departamento de Ingeniería Matemática from the Universidad de Concepción, Concepción, Chile, and was continued while the first author was visiting I.C.T.P., Trieste.

http://www.siam.org/journals/sicon/37-6/33229.html

†Universidad de Concepción, Departamento de Ingeniería Matemática, Casilla 160-C, Concepción, Chile (fflores@ing-mat.udec.cl).

‡Dipartimento di Matematica Pura ed Applicata, Università di Modena, via Campi 213/B, I-41100, Modena, Italy (perrotta@mail.unimo.it).

to deal with nonconvex minimum problems involving some hyperbolic equations on rectangular domains with Darboux conditions on the entire boundary, as shown in [Br-F].

This paper is organized as follows. In section 2, a new Lyapunov-type theorem suitable for our purpose is established. This result involves a "unilateral" condition (see (ii) in Lemma 2.3) and is not valid if the set-valued map considered is not constant, as shown in Remark 2.6. In section 3, we present new classes of nonconvex integrals having minima. In such integrals we consider a monotonicity condition with respect to the state variable which replaces the concavity condition imposed in [Br-F]. Moreover, since our functionals involve a hyperbolic differential operator with Darboux boundary conditions on three sides of the rectangle, we also consider problems with obstacles (Corollary 3.4), unlike [Br-F], where the conditions are given on the entire boundary. The integrands we propose here give rise to integrals that are nondifferentiable and lower semicontinuous only along some special minimizing sequences. In section 4, we also apply the previously mentioned Lyapunov-type result to treat a bang-bang problem for the controlled equation $z_{xy} = u(x, y)$ (Theorems 4.4 and 4.6).

2. A Lyapunov-type result.

Before establishing our Lyapunov-type result, we recall some basic definitions from set-valued analysis. We refer to the book [Au-Fr, Chapter VIII]. In particular, a set-valued map $T : Y \subset \mathbb{R}^m \to \mathbb{R}^n$ is a map from Y to the subsets of $\mathbb{R}^n$. In what follows measurability is with respect to a Lebesgue measure.

Definition 2.1. *Given a measurable set $Y \subset \mathbb{R}^m$, the set-valued map $T : Y \to \mathbb{R}^n$ is called* measurable *if the inverse image of each open set is a measurable set. In other words, if for every open set $V \subset \mathbb{R}^n$, we have that*

$$T^{-1}(V) \doteq \Big\{ y \in Y : \ T(y) \cap V \neq \emptyset \ \Big\}$$

is measurable.

The next proposition will be useful in the proof of our main lemma.

Proposition 2.2. *Let $Y \subset \mathbb{R}^m$ be a closed set, $F : Y \to \mathbb{R}^n$ be a measurable set-valued map with closed and nonempty values, and $\psi : \mathbb{R}^n \times Y \to \mathbb{R}$ be a function such that the map $x \mapsto \psi(x, y)$ is continuous $\forall y \in Y$, $y \mapsto \psi(x, y)$ is measurable on Y $\forall x \in \mathbb{R}^n$. Then, the set-valued map $T : Y \to \mathbb{R}^n$,*

$$T(y) = \Big\{ x \in F(y) : \ \psi(x, y) = 0 \ \Big\},$$

which is supposed to take nonempty values, is measurable with closed values.

Proof. That $T(y)$ is closed follows from the continuity of $\psi(\cdot, y)$ and the closedness of $F(y)$. By virtue of Theorem 8.1.4 in [Au-Fr], we need only verify that the graph of T is measurable. It is defined by Graph $T \doteq \{(x, y) : \ x \in T(y)\}$. The conclusion is then implied by the equality Graph T = Graph $F \cap \psi^{-1}(0)$, which is trivially satisfied. □

In the following, the pair (x, y) with $y = (x_2, x_3, \ldots, x_m)$ denotes a vector in $\mathbb{R}^m$, Q denotes the closed cube in $\mathbb{R}^m$ defined by $Q \doteq \prod_{i=1}^m [a_i, b_i]$, and Q' denotes the closed cube in $\mathbb{R}^{m-1}$ corresponding to the last $m - 1$ components of Q, i.e., $Q' \doteq \prod_{i=2}^m [a_i, b_i] \subset \mathbb{R}^{m-1}$. Given a convex set $K \subset \mathbb{R}^n$, we denote by extr K the set of extreme points of K, that is, the set of such $x \in K$ such that $K \backslash \{x\}$ is convex as well, and by $\langle \cdot, \cdot \rangle$ we denote the scalar product in $\mathbb{R}^n$.

The next result is a variant of Lemma 2.2 in [Am-Ma], suitable for our purpose. We refer to section 4 for a further extension of this result. Applications of it to variational problems arising in economics will appear elsewhere; see [F2] and [F-R].

LEMMA 2.3. *Let $S \subset \mathbb{R}^n$ be a k-dimensional relative open simplex with vertices $c_0, c_1, \dots, c_k$, $E \subset Q \subset \mathbb{R}^m$ be a measurable set, and $v : E \to S$ be a measurable function. Then, there exists a measurable function $w : E \to \operatorname{extr} S$ such that*

(i) *$\int_{a_1}^{b_1} w(x,y)\chi_E(x,y)dx = \int_{a_1}^{b_1} v(x,y)\chi_E(x,y)dx$ for almost all $y \in Q'$;*

(ii) *for every $x \in [a_1, b_1]$, every $\nu \in C_+$, we have*

$$\int_{a_1}^{x} \langle w(r,y), \nu\rangle \chi_E(r,y)dr \geq \int_{a_1}^{x} \langle v(r,y), \nu\rangle \chi_E(r,y)dr \qquad \textit{for almost all } y \in Q'.$$

Here, $C_+ \doteq \{\nu : \langle c_i - c_{i+1}, \nu\rangle \geq 0 \ \ \forall\, i = 0, 1, \dots, k-1\}$.

Proof. A measurable selection theorem allows us to write $v(x,y) = \sum_0^k p_i(x,y)c_i$ for suitable measurable functions $p_i : E \to [0,1]$ satisfying $\sum_0^k p_i \equiv 1$. For any $\nu \in C_+$, we set $\alpha_i = \langle c_i, \nu\rangle$ for $i = 0, \cdots, k$. Thus, $\alpha_0 \geq \alpha_1 \geq \cdots \geq \alpha_k$.

We claim that there exist measurable functions $\delta_i : Q' \to [a_1, b_1]$, $i = 1, \dots, k$, such that $\delta_i(y) \leq \delta_{i+1}(y)$ and, by putting $\delta_0 \equiv a_1$, $\delta_{k+1} \equiv b_1$, one has

$$(2.1) \qquad \int_{\delta_i(y)}^{\delta_{i+1}(y)} \chi_E(x,y)dx = \int_{a_1}^{b_1} p_i(x,y)\chi_E(x,y)dx \qquad \text{for almost all } y \in Q'.$$

To prove (2.1), we proceed recursively as follows. Assuming δ_i is known for $i = 0, \dots, j$, we will define δ_{j+1}. Let us consider the function

$$\psi(x,y) = \int_{a_1}^{x} \chi_E(r,y)dr - \int_{a_1}^{b_1} \sum_{i=0}^{j} p_i(r,y)\chi_E(r,y)dr.$$

This function is such that $x \mapsto \psi(x,y)$ is continuous for almost every (a.e.) y; $y \mapsto \psi(x,y)$ is measurable for every x; $\psi(b_1,y) \geq 0$ since $\sum_0^k p_i \equiv 1$; and

$$\begin{aligned}\psi(\delta_j(y),y) &= \int_{a_1}^{\delta_j(y)} \chi_E(r,y)dr - \sum_{i=0}^{j} \int_{a_1}^{b_1} p_i(r,y)\chi_E(r,y)dr \\ &= -\int_{a_1}^{b_1} p_j(r,y)\chi_E(r,y)dr \leq 0.\end{aligned}$$

Thus, by the previous proposition, the set-valued map $T(y) = \{x \in [\delta_j(y), b_1] : \psi(x,y) = 0\}$, $y \in Q'$, is measurable and then admits at least a measurable selection $\delta_{j+1} : Q' \to [a_1, b_1]$ (see Theorem 8.1.3 in [Au-Fr], for instance). In particular, we have $\delta_{j+1}(y) \geq \delta_j(y)$ and

$$\int_{\delta_j(y)}^{\delta_{j+1}(y)} \chi_E(x,y)dx = \int_{a_1}^{b_1} p_j(x,y)\chi_E(x,y)dx \qquad \text{for almost all } y \in Q'.$$

This proves claim (2.1). A desired function satisfying the requirements of the lemma is given by

$$w(x,y) = \sum_{i=0}^{k} c_i \chi_{E_i(y)\cap E}(x,y) = \sum_{i=0}^{k} c_i \chi_{E_i \cap E}(x,y),$$

where $E_i(y) = [\delta_i(y), \delta_{i+1}(y)[\times Q'$ for $i = 0, \dots, k-1$ and $E_k(y) = [\delta_k(y), b_1] \times Q'$. On the other hand, $E_i = \text{hyp } \delta_{i+1} \setminus \text{hyp } \delta_i$ for $i = 0, \dots, k$. Here hyp δ_i means the hypograph of the function δ_i defined by hyp $\delta_i \doteq \{(y,t) \in Q' \times [a_1, b_1] \ : \ \delta_i(y) \geq t\}$.

Let us now prove part (i).

$$\begin{aligned}\int_{a_1}^{b_1} w(x,y)\chi_E(x,y)dx &= \int_{a_1}^{b_1} \sum_{i=0}^{k} c_i \chi_{E_i(y)\cap E}(x,y)dx = \sum_{i=0}^{k} c_i \int_{\delta_i(y)}^{\delta_{i+1}(y)} \chi_E(x,y)dx \\ &= \sum_{i=0}^{k} c_i \int_{a_1}^{b_1} p_i(x,y)\chi_E(x,y)dx = \int_{a_1}^{b_1} v(x,y)\chi_E(x,y)dx.\end{aligned}$$

It remains only to prove part (ii). Fix any $y \in Q'$ and x such that $\delta_j(y) \leq x \leq \delta_{j+1}(y)$ for $j = 0, \dots, k$. Then

$$\begin{aligned}&\int_{a_1}^{x} \langle w(r,y), \nu\rangle \chi_E(r,y)dr = \int_{a_1}^{x} \sum_{i=0}^{k} \alpha_i \chi_{E_i(y)\cap E}(r,y)dr \\ &= \sum_{i=0}^{j-1} \alpha_i \int_{\delta_i(y)}^{\delta_{i+1}(y)} \chi_E(r,y)dr + \int_{\delta_j(y)}^{x} \alpha_j \chi_E(r,y)dr \\ &= \sum_{i=0}^{j-1} \alpha_i \int_{a_1}^{b_1} p_i(r,y)\chi_E(r,y)dr + \int_{\delta_j(y)}^{x} \alpha_j \chi_E(r,y)dr \\ &= \sum_{i=0}^{j-1} \alpha_i \int_{a_1}^{x} p_i(r,y)\chi_E(r,y)dr + \sum_{i=0}^{j-1} \alpha_i \int_{x}^{b_1} p_i(r,y)\chi_E(r,y)dr \\ &\qquad + \int_{\delta_j(y)}^{x} \alpha_j \chi_E(r,y)dr \\ &\geq \sum_{i=0}^{j-1} \alpha_i \int_{a_1}^{x} p_i(r,y)\chi_E(r,y)dr + \alpha_j \int_{x}^{b_1} \left[1 - \sum_{i=j}^{k} p_i(r,y)\right] \chi_E(r,y)dr \\ &\qquad + \int_{\delta_j(y)}^{x} \alpha_j \chi_E(r,y)dr \\ &= \sum_{i=0}^{j-1} \alpha_i \int_{a_1}^{x} p_i(r,y)\chi_E(r,y)dr + \alpha_j \int_{\delta_j(y)}^{b_1} \chi_E(r,y)dr \\ &\qquad - \alpha_j \int_{x}^{b_1} \sum_{i=j}^{k} p_i(r,y)\chi_E(r,y)dr \\ &= \sum_{i=0}^{j-1} \alpha_i \int_{a_1}^{x} p_i(r,y)\chi_E(r,y)dr + \sum_{i=j}^{k} \alpha_j \int_{a_1}^{x} p_i(r,y)\chi_E(r,y)dr \\ &\geq \int_{a_1}^{x} \sum_{i=0}^{j-1} \alpha_i p_i(r,y)\chi_E(r,y)dr + \int_{a_1}^{x} \sum_{i=j}^{k} \alpha_i p_i(r,y)\chi_E(r,y)dr \\ &= \int_{a_1}^{x} \sum_{i=0}^{k} \alpha_i p_i(r,y)\chi_E(r,y)dr = \int_{a_1}^{x} \langle v(r,y), \nu\rangle \chi_E(r,y)dr.\end{aligned}$$

The latter proves (ii) and the proof of the lemma is concluded. □

Remark 2.4. Under the same assumptions of Lemma 2.3, one can also obtain the existence of another function w satisfying (i) and

For every $x \in [a_1, b_1]$, every $\nu \in C_-$, we have

$$\int_{a_1}^{x} \langle w(r,y), \nu\rangle \chi_E(r,y)dr \leq \int_{a_1}^{x} \langle v(r,y), \nu\rangle \chi_E(r,y)dr \qquad \forall y \in Q'.$$

Here, $C_- \doteq \{\nu : \langle c_i - c_{i+1}, \nu\rangle \leq 0 \ \ \forall\, i = 0, 1, \ldots, k-1\}$.

Remark 2.5. One cannot expect that, in addition to (i) and (ii) in Lemma 2.3, also holds an equality like (i) but integrated with respect to another variable. In other words, assume for simplicity that $n = 1$, $m = 2$, $a_1 = a_2 = 0$, $b_1 = b_2 = 1$, $E = Q \subset \mathbb{R}^2$, and $\nu = 1$. Then one cannot expect that besides satisfying (i) and (ii) in Lemma 2.3, there is also

(iii) $\int_0^1 w(x,y)dy = \int_0^1 v(x,y)dy$ for almost all $x \in \ [0,1]$.

In fact, by integrating (iii) and using the Tonelli–Fubini theorem, one obtains

$$\int_0^1 \Big[\int_0^x w(r,y)dr - \int_0^x v(r,y)dr\Big]dy = 0 \qquad \forall x \in [0,1].$$

Setting $u(x,y) = \int_0^x (w(r,y) - v(r,y))dr$, the above equality becomes $\int_0^1 u(x,y)dy = 0$ $\forall x \in [0,1]$ and, because of (ii), $u(x,y) \geq 0$ $\forall x \in [0,1]$, for a.e. $y \in [0,1]$. Thus, for every $x \in [0,1]$ there exists a null set $N(x) \subset [0,1]$ such that $u(x,y) = 0\ \forall\, y \in [0,1] \setminus N(x)$. It turns out that $u(x,y) = 0\ \forall\, x \in Q_1$, $\forall\, y \in [0,1] \setminus N$ with $N = \bigcup_{x \in Q_1} N(x)$ (independent of x), being a null set, where Q_1 is the set of rational numbers in $[0,1]$. Since $u(\cdot, y)$ is continuous, the latter implies that $u(x,y) = 0$ $\forall x \in [0,1]$, a.e. $y \in [0,1]$. It follows that

$$\int_B [w(r,s) - v(r,s)]dsdr = 0 \qquad \forall \text{ Borel set } B \subset Q.$$

Hence, $w(r,s) = v(r,s)$ for almost all (r,s) in Q. The latter gives a contradiction since v takes values in the open simplex S and w takes values only on its extreme points. In other words, we have shown that one cannot obtain, in Lemma 2.3, an equality like (i) where the integral is respect to one variable and an inequality like (ii) integrated with respect to another variable.

Remark 2.6. The result in Lemma 2.3 is not valid if one considers a set-valued map depending on x. In fact, simply take $n = 2$, $m = 2$, $E = Q = [0,1]^2$, $\nu = (1,0)$, the set-valued map $\Phi(x,y) = \{\lambda(1,x) : \ \lambda \in \ [0,1]\}$ (instead of $\Phi(x,y) \equiv S$), and $v(x,y) = \frac{1}{2}(1,x)$. Then, it is not difficult to show that there exists no function w taking values in $\{(0,0),(1,x)\}$ such that (i) and (ii) continues to be valid (compare with the example in [Am-Ma]).

3. Nonconvex minimum problems involving a hyperbolic equation.

We shall consider, on the rectangle $Q = [a,b] \times [c,d] \subset \mathbb{R}^2$, Darboux-type boundary conditions for the equation

$$z_{xy}(x,y) = u(x,y), \tag{3.1}$$

$$z(x,y) = \psi(x,y), \qquad (x,y) \in \partial_+ Q, \tag{3.2}$$

where $\partial_+ Q = \{a\} \times [c,d] \cup [a,b] \times \{c\} \cup \{b\} \times [c,d]$ and ψ is a continuous function on $\partial_+ Q$ such that the restriction of ψ to each one of the three sides of $\partial_+ Q$ is absolutely continuous with derivative in L^p for $p \in \ ]1, +\infty[$.

We denote by $W^{*,p}(Q,\mathbb{R}^n)$ the space of all functions $z : Q \to \mathbb{R}^n$ in L^p whose distributional derivatives z_x, z_y, z_{xy} are in L^p. This space becomes a Banach space with the norm $\|z\|_* = \|z\|_p + \|z_x\|_p + \|z_y\|_p + \|z_{xy}\|_p$ as shown in [S2]. It is easy to see that the following integral representation for $z \in W^{*,p}$ is valid:

$$z(x,y) = z(x,c) + z(a,y) - z(a,c) + \int_a^x \int_c^y z_{xy}(r,s)dsdr.$$

In particular, given any $u \in L^p$, every solution to (3.1) and (3.2) is in $W^{*,p}(Q,\mathbb{R}^n)$ and satisfies

$$z(x,y) = \psi(x,c) + \psi(a,y) - \psi(a,c) + \int_a^x \int_c^y u(r,s)dsdr. \tag{3.3}$$

In the following $\mathcal{L} \otimes \mathcal{B}_k$ will denote the product of the Lebesgue σ-algebra on Q with the Borel σ-algebra on $\mathbb{R}^k$. We recall that a function $f : Q \times \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ is called $\mathcal{L} \otimes \mathcal{B}_k$-measurable or simply measurable if the inverse image under f of every closed subset of $\overline{\mathbb{R}}$ is in $\mathcal{L} \otimes \mathcal{B}_k$. By h^{**}, we denote the biconjugate (bipolar) function defined as the greatest convex lower semicontinuous function not greater than h. Whenever $\xi \in \mathbb{R}^n$, $|\xi|$ stands, unless otherwise specified, for the Euclidean norm in $\mathbb{R}^n$.

For fixed $\nu \in \mathbb{R}^n$, on the functional $J_1 : W^{*,p}(Q,\mathbb{R}^n) \to \mathbb{R} \cup \{+\infty\}$

$$J_1(z) = \int_a^b \int_c^d h(z_{xy})dydx + \int_a^b \int_c^d g(x,y,\langle z,\nu\rangle)dydx + \int_a^b \int_c^d f(x,y,\langle z_y,\nu\rangle)dydx,$$

to be minimized, we assume the following hypothesis.

Hypothesis H1. We assume $1 < p < +\infty$. The function $f : Q \times \mathbb{R} \to \mathbb{R}$ is such that

(f_1) f is $\mathcal{L} \otimes \mathcal{B}_1$-measurable;

(f_2) the map $\rho \mapsto f(x,y,\rho)$ is convex and nonincreasing for a.e. $(x,y) \in Q$;

(f_3) there exists a constant $\alpha_1 \geq 0$ and a function $\beta_1 \in L^1$ such that

$$f(x,y,\rho) \geq -\alpha_1|\rho|^p - \beta_1(x,y).$$

The function $g : Q \times \mathbb{R} \to \mathbb{R}$ is such that

(g_1) g is $\mathcal{L} \otimes \mathcal{B}_1$-measurable;

(g_2) the map $\eta \mapsto g(x,y,\eta)$ is lower semicontinuous and nonincreasing for a.e. $(x,y) \in Q$;

(g_3) there exists a constant $\alpha_2 \geq 0$ and a function $\beta_2 \in L^1$ such that

$$g(x,y,\eta) \geq -\alpha_2|\eta|^p - \beta_2(x,y).$$

The function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is such that

(h_1) h is lower semicontinuous;

(h_2) there exist some constants $\alpha_3 > 0$, $\beta_3 \in \mathbb{R}$ such that

$$h(\xi) \geq \alpha_3|\xi|^p - \beta_3;$$

(h_3) setting $K = \{\xi \in \mathbb{R}^n : \ h^{**}(\xi) < h(\xi)\}$, we require that $K \subset \bigcup_{i\in I} S_i$ where I is a countable set and each S_i is a relative open simplex in $\mathbb{R}^n$ such that $h^{**} = h$ on extr S_i and h^{**} is affine on every S_i. Here the sets S_i are supposed to be disjoint. There is no assumption if $n = 1$.

In addition, for fixed $\nu \in \mathbb{R}^n$, we assume

$$\alpha_3 - (\alpha_2(d-c)^p 2^{p-1} + \alpha_1)2^{p-1}(b-a)^p|\nu|^p > 0. \tag{3.4}$$

We are interested in the following minimization problem:

$$\min_{z \in Z} J_1(z), \tag{P$_1$}$$

where

$$Z = \Big\{ z \in W^{*,p}(Q, \mathbb{R}^n) : z \text{ satisfies (3.2)} \Big\}.$$

We are now in a position to state the first main result.

THEOREM 3.1. *Let f, g, h satisfy Hypothesis* H1 *above. Assume the functional J_1 is finite at some $z \in Z$. Then problem* (P$_1$) *admits at least a solution.*

Proof. (a) We first consider the convexified problem associated with (P$_1$):

$$\min_{z \in Z} J_1^{**}(z), \tag{P_1^{**}}$$

where J_1^{**} is obtained from J by replacing h by h^{**}. Because of the assumptions on the integrands, $\lambda \doteq \inf J_1^{**}(z)$ is finite. Take any minimizing sequence (z^k) in Z. Then, by the condition on (3.4), we obtain $\|z^k_{xy}\|_p \leq B$ for some positive constant B. Therefore, there exists a subsequence, still indexed by k, such that $z^k_{xy} \rightharpoonup \xi$ in $L^p(Q, \mathbb{R}^n)$. This implies that $\int_a^x \int_c^y z^k_{xy}(r,s)dsdr \to \int_a^x \int_c^y \xi(r,s)dsdr$ for $(x,y) \in Q$. Therefore, $z^k(x,y) \to \tilde{z}(x,y)$ for $(x,y) \in Q$, where

$$\tilde{z}(x,y) = \psi(x,c) + \psi(a,y) - \psi(a,c) + \int_a^x \int_c^y \xi(r,s)dsdr.$$

This implies that $\tilde{z} \in W^{*,p}(Q, \mathbb{R}^n)$. Since Q is a compact set and z^k and $\tilde{z}$ are continuous, $z^k \to \tilde{z}$ uniformly. Thus $\tilde{z} \in Z$. On the other hand, it is not difficult to prove that $z^k_y \rightharpoonup \tilde{z}_y$ in L^p and $z^k_x \rightharpoonup \tilde{z}_x$ in L^p. Hence $\tilde{z}$ is a minimizer for J_1^{**} since h^{**} and $f(x,y,\cdot)$ are convex functions.

(b) Let $\tilde{z} \in Z$ be any solution to problem (P$_1^{**}$). We set for every $i \in I$, $E_i = \{(x,y) \in Q : \tilde{z}_{xy}(x,y) \in S_i\}$, and apply Lemma 2.3 to obtain a measurable function w_i taking values in extr S_i on E_i, such that, for every $i \in I$,

(i) $\int_a^b w_i(x,y)\chi_{E_i}(x,y)dx = \int_a^b \tilde{z}_{xy}(x,y)\chi_{E_i}(x,y)dx$ for a.e. $y \in [c,d]$, and

(ii) for every $x \in [a,b]$, we have

$$\int_a^x \langle w_i(r,y), \nu\rangle \chi_{E_i}(r,y)dr \geq \int_a^x \langle \tilde{z}_{xy}(r,y), \nu\rangle \chi_{E_i}(r,y)dr \qquad \text{for a.e. } y \in [c,d].$$

Put $E_0 = Q \setminus \bigcup_i E_i$, and define $w : Q \to \mathbb{R}^n$ by

$$w(x,y) = \tilde{z}_{xy}(x,y)\chi_{E_0}(x,y) + \sum_{i \in I} w_i(x,y)\chi_{E_i}(x,y).$$

Clearly, this function is measurable and, because of the growth condition (h_2) and (h_3)—h^{**} is affine in S_i—w is in L^p since the integral $\int\int h^{**}(\tilde{z}_{xy}(x,y))dydx$ is finite. Moreover, by Vitali's convergence theorem, it follows that

$$\int_a^b w(x,y)dx = \int_a^b \tilde{z}_{xy}(x,y)dx \qquad \text{for a.e. } y \in [c,d] \tag{3.5}$$

and for every $x \in [a,b]$

$$\int_a^x \langle w(r,y),\nu\rangle dr \geq \int_a^x \langle \tilde{z}_{xy}(r,y),\nu\rangle dr \qquad \text{for a.e. } y \in [c,d]. \tag{3.6}$$

Thus, the function $z : Q \to \mathbb{R}^n$ given by

$$z(x,y) = \psi(x,c) + \psi(a,y) - \psi(a,c) + \int_a^x \int_c^y w(r,s)dsdr \tag{3.7}$$

is in $W^{*,p}(Q,\mathbb{R}^n)$ and satisfies the boundary condition (3.2). The latter follows from (3.5) and the integral representation for z and $\tilde{z}$.

On the other hand, since h^{**} is affine on each S_i, (3.5) implies

$$\begin{aligned}\int_a^b \int_c^d h(z_{xy}(x,y))dydx &= \int_a^b \int_c^d h^{**}(z_{xy}(x,y))dydx \\ &= \int_a^b \int_c^d h^{**}(\tilde{z}_{xy}(x,y))dydx.\end{aligned} \tag{3.8}$$

It follows from (3.6) that for every $x \in [a,b]$, we have

$$\int_a^x \langle z_{xy}(r,y),\nu\rangle dr \geq \int_a^x \langle \tilde{z}_{xy}(r,y),\nu\rangle dr \qquad \text{for a.e. } y \in [c,d].$$

As a consequence,

$$\langle z_y(x,y),\nu\rangle \geq \langle \tilde{z}_y(x,y),\nu\rangle \quad \text{and} \quad \langle z(x,y),\nu\rangle \geq \langle \tilde{z}(x,y),\nu\rangle \tag{3.9}$$

for a.e. $(x,y) \in Q$, which obviously implies

$$\int_a^b \int_c^d f(x,y,\langle z_y(x,y),\nu\rangle)dydx \leq \int_a^b \int_c^d f(x,y,\langle \tilde{z}_y(x,y),\nu\rangle)dydx, \tag{3.10}$$

$$\int_a^b \int_c^d g(x,y,\langle z(x,y),\nu\rangle)dydx \leq \int_a^b \int_c^d g(x,y,\langle \tilde{z}(x,y),\nu\rangle)dydx. \tag{3.11}$$

Hence, (3.8), together with (3.10) and (3.11), implies that z is a solution to problem (P_1) since $\tilde{z}$ is a solution to (P_1^{**}) and $h^{**} \leq h$. □

Remark 3.2. The assumption $K \subset \bigcup_{i\in I} S_i$, (h_3) in Hypothesis H1, is imposed (instead of an equality) in order to include some cases in which, for instance, the integrand h has rotational symmetry; that is, h is of the form $h(\xi) = \bar{h}(|\xi|)$. This fact is exhibited in the following example. Let us fix $0 < t_1 < t_2$ and consider an even function $\bar{h}$ such that, besides satisfying the superlinear growth condition (h_2), we also assume, for $i = 1,2$, that $\bar{h}(t) = \bar{h}(t_i) = \bar{h}^{**}(t_i)$ if $0 \leq t \leq t_1$, $\bar{h}(t_1) < \bar{h}(t)$ if $t_1 < t < t_2$, and $\bar{h}(t) = \bar{h}^{**}(t)$ $\forall t > t_2$ with $\bar{h}_+^{**\prime}(t_2) > 0$. Setting $h(\xi) = \bar{h}(|\xi|)$, we have $h^{**}(\xi) = \bar{h}^{**}(|\xi|)$ and

$$K \doteq \Big\{ \xi \in \mathbb{R}^n : \bar{h}^{**}(|\xi|) < \bar{h}(|\xi|) \Big\} = \Big\{ \xi \in \mathbb{R}^n : \ t_1 < |\xi| < t_2 \Big\}.$$

Thus, K is not a countable union of simplices but satisfies the inclusion assumption (see Lemma 4.3 later) as required in (h_3) and then the theorem can be applied.

Motivated by the fact that (h_3) is a consequence of (h_2) if $h : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, we will establish another existence result whose proof follows the reasoning of the preceding theorem componentwise. Before proceeding, we give some notation. By $\tilde{z} \preceq z$ we mean $\tilde{z}^i \leq z^i$ $\forall i = 1, \dots, n$ and a function $g : \mathbb{R}^n \to \mathbb{R}$ is said to be nondecreasing (nonincreasing) if $\tilde{z} \preceq z$ implies $g(\tilde{z}) \leq g(z)$ $(\geq)$. For fixed $p > 1$, $|\xi| = [\sum_{j=1}^n |\xi^j|^p]^{\frac{1}{p}}$ whenever $\xi = (\xi^1, \dots, \xi^n)$. Thus, if $z \in L^p(Q, \mathbb{R}^n)$, then $\|z\|_p = [\sum_{j=1}^n \int_Q |z^j(x,y)|^p dxdy]^{\frac{1}{p}}$.

Hypothesis H2. We assume $1 < p < +\infty$. The function $f : Q \times \mathbb{R}^n \to \mathbb{R}$ is such that

(f_1) f is $\mathcal{L} \otimes \mathcal{B}_n$-measurable;

(f_2) the map $\rho \mapsto f(x,y,\rho)$ is convex and nonincreasing for a.e. $(x,y) \in Q$;

(f_3) there exists a constant $\alpha_1 \geq 0$ and a function $\beta_1 \in L^1$ such that

$$f(x,y,\rho) \geq -\alpha_1 |\rho|^p - \beta_1(x,y).$$

The function $g : Q \times \mathbb{R}^n \to \mathbb{R}$ is such that

(g_1) g is $\mathcal{L} \otimes \mathcal{B}_n$-measurable;

(g_2) the map $\eta \mapsto g(x,y,\eta)$ is lower semicontinuous and nonincreasing for a.e. $(x,y) \in Q$; (g_3) there exists a constant $\alpha_2 \geq 0$ and a function $\beta_2 \in L^1$ such that

$$g(x,y,\eta) \geq -\alpha_2 |\eta|^p - \beta_2(x,y).$$

The function $h_j : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, $j = 1, \dots, n$, is such that

(h_1) h_j is lower semicontinuous;

(h_2) there exist some constants $\alpha_3 > 0$, $\beta_3 \in \mathbb{R}$ such that

$$h_j(\xi) \geq \alpha_3 |\xi|^p - \beta_3.$$

In addition, we assume

$$\alpha_3 - (\alpha_2 (d-c)^p 2^{p-1} + \alpha_1) 2^{p-1} (b-a)^p > 0. \tag{3.12}$$

Now, our purpose is to consider the following minimization problem:

$$\min_{z \in Z} J_2(z), \tag{P$_2$}$$

where $J_2 : Z \to \mathbb{R} \cup \{+\infty\}$ is given by

$$J_2(z) = \int_a^b \int_c^d \sum_{j=1}^n h_j(z_{xy}^j) dydx + \int_a^b \int_c^d g(x,y,z) dydx + \int_a^b \int_c^d f(x,y,z_y) dydx$$

and

$$Z = \Big\{ z \in W^{*,p}(Q, \mathbb{R}^n) : z \text{ satisfies } (3.2) \Big\}.$$

We have the second main theorem.

THEOREM 3.3. *Let f, g, h_j satisfy Hypothesis* H2 *above. Assume the functional J_2 is finite at some $z \in Z$. Then problem* (P$_2$) *admits at least a solution.*

Proof. We start by considering the convexified problem

$$\min_{z \in Z} J_2^{**}(z), \tag{P$_2^{**}$}$$

where J_2^{**} is obtained from J_2 by replacing each h_j with h_j^{**}. We proceed as in the proof (part (a)) of the previous theorem to conclude that problem (P_2^{**}) has at least a solution, say $\tilde{z} = (\tilde{z}^1, \dots, \tilde{z}^n)$. On the other hand, it is well known that under assumption (h_2) in Hypothesis H2, one has

$$\left\{\xi \in \mathbb{R} : h_j^{**}(\xi) < h_j(\xi)\right\} = \bigcup_{i \in I_j} \,]a_i^j, b_i^j[,$$

where the intervals $]a_i^j, b_i^j[$ are disjoint and finite with I_j being a countable set. In this situation each $]a_i^j, b_i^j[$ plays the role of the simplex S_i in Assumption H1. We now reason as in part (b) in the proof of Theorem 3.3. That is, setting

$$E_i^j = \left\{(x, y) \in Q : \ \tilde{z}_{xy}^j(x, y) \in \ ]a_i^j, b_i^j[\ \right\}$$

for every $j = 1, \dots, n$ and every $i \in I_j$, we apply Lemma 2.3 (with $]a_i^j, b_i^j[$ instead of S, E_i^j instead of E and $\nu = 1$) to obtain a measurable function w_i^j taking values in $\{a_i^j, b_i^j\}$ on E_i^j, such that

(i) $\int_a^b w_i^j(x, y)\chi_{E_i^j}(x, y)dx = \int_a^b \tilde{z}_{i,xy}^j(x, y)\chi_{E_i^j}(x, y)dx$ for a.e. $y \in [c, d]$ and

(ii) for every $x \in [a, b]$, we have

$$\int_a^x w_i^j(r, y)\chi_{E_i^j}(r, y)dr \geq \int_a^x \tilde{z}_{i,xy}^j(r, y)\chi_{E_i^j}(r, y)dr \qquad \text{for a.e. } y \in [c, d].$$

Put $E_0^j = Q \setminus \bigcup_{i \in I_j} E_i^j$, and define $z^j : Q \to \mathbb{R}$ by

$$z_{xy}^j(x, y) = \tilde{z}_{xy}^j(x, y)\chi_{E_0^j}(x, y) + \sum_{i \in I_j} w_i^j(x, y)\chi_{E_i^j}(x, y),$$
$$z^j(a, y) = \psi^j(a, y), \ z^j(x, c) = \psi^j(x, c).$$

Exactly as in the proof of the previous theorem, it can be proven that $\tilde{z}^j \in W^{*,p}(Q, \mathbb{R})$ satisfies the boundary condition (3.2),

$$\int_a^b\int_c^d h_j(z_{xy}^j(x, y))dydx = \int_a^b\int_c^d h_j^{**}(z_{xy}^j(x, y))dydx = \int_a^b\int_c^d h_j^{**}(\tilde{z}_{xy}^j(x, y))dydx,$$

and $z_y^j(x, y) \geq \tilde{z}_y^j(x, y)$, $z^j(x, y) \geq \tilde{z}^j(x, y)$ for a.e. (x, y) in Q. Consequently, by setting $z = (z^1, \dots, z^n)$, we obtain

$$\int_a^b\int_c^d \sum_{j=1}^n h_j(z_{xy}^j(x, y))dydx = \int_a^b\int_c^d \sum_{j=1}^n h_j^{**}(\tilde{z}_{xy}^j(x, y))dydx,$$

and by the monotonicity assumption on f and g,

$$\int_a^b\int_c^d f(x, y, z_y(x, y))dydx \leq \int_a^b\int_c^d f(x, y, \tilde{z}_y(x, y))dydx$$

and

$$\int_a^b\int_c^d g(x, y, z(x, y))dydx \leq \int_a^b\int_c^d g(x, y, \tilde{z}(x, y))dydx.$$

These facts show that z is a solution to problem (P_2), since $\tilde{z}$ is a solution to problem (P_2^{**}) and $h_j^{**} \leq h_j$. $\square$

The application of Lemma 2.3 in the proof of Theorems 3.1 and 3.3 allows us to construct a function z from $\tilde{z}$ satisfying

$$\langle z(x,y), \nu \rangle \geq \langle \tilde{z}(x,y), \nu \rangle \quad \text{and} \quad \tilde{z}(x,y) \preceq z(x,y), \tag{3.13}$$

respectively. Therefore, setting

$$Z_1 = \Big\{ z \in W^{*,p}(Q, \mathbb{R}^n) : z \text{ satisfies } (3.2), \langle z, \nu \rangle \geq \varphi_1 \text{ in } Q \Big\},$$

$$Z_2 = \Big\{ z \in W^{*,p}(Q, \mathbb{R}^n) : z \text{ satisfies } (3.2),\ \varphi_2 \preceq z \text{ in } Q \Big\},$$

for given measurable functions $\varphi_1 : Q \to \mathbb{R}$ and $\varphi_2 : Q \to \mathbb{R}^n$, we also obtain the following corollary.

Corollary 3.4. *Assume, in addition to Hypothesis* H1 *or* H2*, that some measurable constraints, as above, are given. Then the problems*

$$\min_{z \in Z_1} J_1(z) \quad \textit{and} \quad \min_{z \in Z_2} J_2(z)$$

admit at least a solution.

Remark 3.5. (i) By Remark 2.4, Theorems 3.1 and 3.3 continue to be valid in case $f(x,y,\cdot)$ and $g(x,y,\cdot)$ are nondecreasing functions. Moreover, under this assumption, the previous corollary also holds with the reverse inequality appearing in the definition of Z_1 and Z_2. In fact, Remark 2.4 asserts the existence (similar to (3.13)) of a function z satisfying

$$\langle z(x,y), \nu \rangle \leq \langle \tilde{z}(x,y), \nu \rangle \quad \text{or} \quad z(x,y) \preceq \tilde{z}(x,y)$$

according to whether we are dealing with problem Z_1 or Z_2, both with the reverse inequalities. Thus, if $\langle \tilde{z}(x,y), \nu \rangle \leq \varphi_1(x,y)$ (respectively, $\tilde{z} \preceq \varphi_2$), then $\langle z(x,y), \nu \rangle \leq \varphi_1(x,y)$ (respectively, $z \preceq \varphi_2$), and hence the conclusion follows since $f(x,y,\cdot)$ and $g(x,y,\cdot)$ are nondecreasing functions.

(ii) As a consequence of the monotonicity of $f(x,y,\cdot)$ and $g(x,y,\cdot)$, together with the fact that z is also a solution to (P_1^{**}) (respectively, (P_2^{**})), one can conclude that for almost all (x,y) in Q,

$$\begin{cases} f(x,y,\langle z_y(x,y), \nu \rangle) = f(x,y,\langle \tilde{z}_y(x,y), \nu \rangle), \\ g(x,y,\langle z(x,y), \nu \rangle) = g(x,y,\langle \tilde{z}(x,y), \nu \rangle) \end{cases}$$

(respectively,

$$\begin{cases} f(x,y,z_y(x,y)) = f(x,y,\tilde{z}_y(x,y)), \\ g(x,y,z(x,y)) = g(x,y,\tilde{z}(x,y))). \end{cases}$$

The last part of the preceding remark yields the following corollary.

Corollary 3.6. *Assume, in addition to the hypothesis of Theorems* 3.1 *or* 3.3*, that $n = 1$ and for almost all $(x,y) \in Q$ one of the functions $f(x,y,\cdot)$ or $g(x,y,\cdot)$ is strictly decreasing in $\mathbb{R}$ and $\nu = 1$. Then every solution to the convexified problem* $(\mathrm{P}_\mathrm{i}^{**})$ $(i = 1,2)$ *is a solution to the original (nonconvex) problem* (P_i) $(i = 1,2)$.

Notice that the previous corollary admits the following variant. Instead of assuming that $f(x,y,\cdot), g(x,y,\cdot)$ are strictly decreasing, one can impose that both functions satisfy $f(x,y,\rho) = f_1(x,y)f_2(\rho)$, $g(x,y,\eta) = g_1(x,y)g_2(\eta)$ with suitable assumptions on f_i, g_i. Here, f_2, g_2 might be piecewise monotone.

Example 3.7. In what follows I_K will denote the indicator function of the set K, i.e., $I_K(\xi) = 0$ if $\xi \in K$ and $I_K(\xi) = +\infty$ otherwise. For fixed $p \geq 1$, let us consider the problem of minimizing

$$J_1(z) = \int_0^1 \int_0^1 I_{\{-1,+1\}}(z_{xy}(x,y))dydx$$

among those functions z in $W^{*,p}([0,1] \times [0,1], \mathbb{R})$ satisfying $z(0,y) = z(x,0) = z(1,y) = 0$ for $x \in [0,1]$, $y \in [0,1]$. We denote this space by Z. Clearly, this problem admits infinitely many solutions. In fact, for every $k \in \mathbb{N}$, it is not difficult to check that the function z^k defined by

$$z^k(x,y) = \begin{cases} -\Big(x - \dfrac{2i-2}{2^k}\Big)y & \text{if } \dfrac{2i-2}{2^k} \leq x \leq \dfrac{2i-1}{2^k}, \\ \Big[-\dfrac{1}{2^k} + \Big(x - \dfrac{2i-1}{2^k}\Big)\Big]y & \text{if } \dfrac{2i-1}{2^k} \leq x \leq \dfrac{2i}{2^k} \end{cases}$$

($i = 1, 2, \dots, 2^{k-1}$) is a minimizer for J_1 in Z. Notice that $z^k \to 0$ uniformly. We now consider the functional J_2,

$$J_2(z) = \int_0^1 \int_0^1 I_{\{-1,+1\}}(z_{xy}(x,y))dydx + \int_0^1 \int_0^1 |z(x,y)|^p dydx,$$

defined on the same space as J_1. One can easily prove that the sequence defined above is minimizing for J_2, i.e., $\lim_k J_2(z^k) = \inf J_2(z) = 0$. However, since every function in $W^{*,p}$ is continuous, the minimum value of J_2 is not attained. Notice that the monotonicity assumption (g_2) in Hypothesis H1 or H2 is not satisfied. The same conclusion also is reached if we consider the functional $J_3 : Z \to \mathbb{R}$,

$$J_3(z) = J_1(z) + \int_0^1 \int_0^1 |z_y(x,y)|^p dydx,$$

since the sequence, as above, continues to be minimizing for J_3 and $z_y^k \to 0$ uniformly. Observe that $z_x^k \rightharpoonup 0$ in L^p.

Remark 3.8. Certainly, because of the analogy of the results in Lemma 2.3, we also obtain existence results for the functional J_1 (or J_2) depending on z_x in place of z_y, but in this case the boundary data have to be given on $\partial^- Q = [a,b] \times \{c\} \cup \{a\} \times [c,d] \cup [a,b] \times \{d\}$. On the other hand, due to the last part of Remark 2.5, our approach cannot be applied to a functional depending on z_x and z_y simultaneously. However, we believe that such a result, in general, is not true even if we were unable to find a concrete example.

4. Some bang-bang theorems for the simplest hyperbolic equation. This section is devoted to prove a *weak* bang-bang theorem for the hyperbolic inclusion

$$z_{xy}(x,y) \in C \qquad \text{for a.e. } (x,y) \in Q \tag{4.1}$$

where C is a compact convex subset of $\mathbb{R}^n$, $z \in W^{*,\infty}(Q, \mathbb{R}^n)$, and all the notations are those introduced in section 3. Some of the topological concepts to be used here have been introduced in section 2. Recall that, given a set E, co E and $\mathrm{aff}(E)$ stand for the convex hull of E and for the smallest affine space containing E, respectively ([Roc]). Moreover, if C is a convex subset of $\mathbb{R}^n$, a face F of C is said to be *exposed* if it is the intersection of C and a supporting hyperplane of C. The normal cone to C at a point $x \in \mathbb{R}^n$ is defined by

$$N_C(x) = \{\xi \in \mathbb{R}^n \,:\, \langle \xi, x-y\rangle \geq 0 \;\forall y \in C\}$$

and the tangent cone

$$T_C(x) = \{\zeta \in \mathbb{R}^n \,:\, \langle \zeta, \xi\rangle \leq 0 \;\forall \xi \in N_C(x)\}\,.$$

Definition 4.1. *Let C be a convex set and e be a point such that $e \notin C$. We say that the supporting hyperplane H to C strictly separates C and e if C is contained in a closed half-space associated with H and e belongs to the opposite open half-space.*

First we prove the following technical proposition.

Proposition 4.2. *Let $P \subset \mathbb{R}^n$ be an n-dimensional convex polytope (see* [Roc]*) and e and x, $e \in \mathbb{R}^n \backslash P$ and $x \in \partial P$, be such that $]x,e] \cap P$ is empty. Then there exists an $(n-1)$-dimensional face F of P such that $x \in F$ and* $\mathrm{aff}(F)$ *strictly separates P and e.*

Proof. Let F_i, $i = 1,\dots,k$, be the $(n-1)$-dimensional faces of P and F_i, $i = 1,\dots,k_x$, be those which contain x. For every $i = 1,\dots,k$ denote by ν_i the exterior normal vector at the face F_i and by π_i the supporting half-space to P at every point of F_i. Thus, for every point $y \in \partial\pi_i$, the tangent and the normal cones to π_i at y are given by $T_{\pi_i}(y) = \pi_i - y$ and $N_{\pi_i}(y) = \mathbb{R}_0^+\nu_i$, respectively. Recalling that $P = \cap_{i=1}^k \pi_i$, by Corollary 23.8.1 in [Roc] (see also [Au-Fr, p. 141]) the normal cone to P at x is given by $N_P(x) = \sum_{i=1}^k N_{\pi_i}(x) = \sum_{i=1}^{k_x} N_{\pi_i}(x)$; hence

$$N_P(x) = \left\{\sum_{i=1}^{k_x} \lambda_i\nu_i \,:\, \lambda_i \geq 0,\; i = 1,\dots k_x\right\}.$$

As a consequence, the tangent cone to P at x is given by

$$T_P(x) = \bigcap_{i=1}^{k_x} T_{\pi_i}(x) = \bigcap_{i=1}^{k_x}(\pi_i - x).$$

Taking into account that every face of P is exposed, from $]x,e] \cap P = \emptyset$ it follows that $(e-x) \notin T_P(x)$. Therefore there exists $i \in \{1,\dots,k_x\}$ such that $e \notin \pi_i$, and the conclusion follows. □

The following result is a sharper version of Lemma 2.3 in [Am-Ma]. We prove that every n-dimensional compact convex set C can be written, up to the boundary, as a union of a countable family of n-dimensional simplices (with pairwise disjoint interior) having their vertices in the set of extreme points of C.

Lemma 4.3. *Let C be an n-dimensional compact convex subset of $\mathbb{R}^n$. Then there exists a countable family $\mathcal{S}$ of n-dimensional simplices such that*

(i) int $C \subset \cup\{S \,:\, S \in \mathcal{S}\}$*;*

(ii) int $S \cap$ int $S' = \emptyset$ *if $S, S' \in \mathcal{S}$ and $S \neq S'$;*

(iii) extr $S \subseteq$ extr C *for every $S \in \mathcal{S}$.*

Proof. If C has only a finite number of extreme points, then it is a polytope and the proof is trivial. Otherwise, let E be a countable dense subset of extr C and set

$E = \{e_i : i \in \mathbb{N}\}$ such that $e_0, e_1, \dots, e_n$ are affinely independent. We are going to construct the family $\mathcal{S}$ by a recursive argument.

At the first step set $S_1^1 = \text{co}\{e_0, \dots, e_n\}$, $C_1 = S_1^1$, and $n_1 = 1$. Obviously, S_1^1 is an n-dimensional simplex. At the second step, consider the point $e_{n+1} \in E$ and denote by F_k^1, $k = 1, \dots, n_2$, the $(n-1)$-dimensional faces of C_1 that satisfy the following property: for every $k = 1, \dots, n_2$ the hyperplane $\text{aff}(F_k^1)$ generated by F_k^1 strictly separates C_1 and e_{n+1}. By Proposition 4.2 the set $\{F_k^1 : k = 1, \dots, n_2\}$ is not empty. For every $k = 1, \dots, n_2$ denote by S_k^2 the n-dimensional simplex given by $S_k^2 = co(F_k^1 \cup \{e_{n+1}\})$. The finite family $\{S_1^1, S_k^2 : k = 1, \dots, n_2\}$ fulfills properties (ii) and (iii) in the statement. Finally, set $C_2 = C_1 \cup \{S_k^2 : k = 1, \dots, n_2\}$. We claim that $C_2 = \text{co}\{e_0, \dots, e_{n+1}\}$.

Obviously $C_2 \subseteq \text{co}\{e_0, \dots, e_{n+1}\}$. In order to prove the opposite inclusion consider $x = \sum_{i=0}^{n+1} \lambda_i e_i$, $0 \le \lambda_i \le 1$, $\sum_{i=0}^{n+1} \lambda_i = 1$. If $\lambda_{n+1} = 0$, then $x \in S_1^1 \subset C_2$. If $\lambda_{n+1} = 1$, then $x = e_{n+1} \in C_2$. Otherwise, if $0 < \lambda_{n+1} < 1$, set $\bar{x} = \sum_{i=0}^{n} \lambda_i/(1 - \lambda_{n+1})e_i$. Then $\bar{x} \in S_1^1$ and $x = (1 - \lambda_{n+1})\bar{x} + \lambda_{n+1}e_{n+1}$. Let $\tilde{x}$ be the *last* point in $[\bar{x}, e_{n+1}] \cap \partial C_1$. By Proposition 4.2 there exists an $(n-1)$-dimensional face F of C_1 such that $\tilde{x} \in F$ and $\text{aff}(F)$ strictly separates C_1 and e_{n+1}. The definition of the simplices S_k^2 implies that the segment whose extreme points are $\tilde{x}$ and e_{n+1} is included in C_2. Hence, the segment $[\bar{x}, e_{n+1}]$ is also included in C_2 and the claim is proven.

Suppose now we have defined $m - 1$ finite families

$$\left\{S_k^h : k = 1, \dots, n_h\right\}, \qquad h = 1, \dots, m-1,$$

of simplices whose interior points are mutually disjoint and whose extreme points are in $\{e_0, \dots, e_{n+m-2}\}$, and $m - 1$ convex sets

$$C_h = C_{h-1} \cup \left(\bigcup_{k=1}^{n_h} S_k^h\right) = \text{co}\{e_0, \dots, e_{n+h-1}\}, \qquad h = 1, \dots, m-1 \quad (C_0 = \emptyset).$$

In particular $C_{m-1} = \cup_{h=1}^{m-1} \cup_{k=1}^{n_h} S_k^h = \text{co}\{e_0, \dots, e_{n+m-2}\}$. Following the same ideas of the second step, denote by $\{F_k^{m-1} : k = 1, \dots n_m\}$ the family of all $(n-1)$-dimensional faces of C_{m-1} such that $\text{aff}(F_k^{m-1})$ strictly separates C_{m-1} and e_{n+m-1}. By Proposition 4.2 this family is not empty. For every $k \in \{1, \dots, n_m\}$, define the simplex $S_k^m = \text{co}(F_k^{m-1} \cup \{e_{n+m-1}\})$. Clearly, the family $\{S_k^h : h = 1, \dots, m, k = 1, \dots, n_h\}$ satisfies items (ii) and (iii) of the statement. Moreover, setting $C_m = C_{m-1} \cup (\cup_{k=1}^{n_m} S_k^m)$, one can prove, as in the previous step, that $C_m = \text{co}\{e_0, \dots, e_{n+m-1}\}$.

Recursively, one obtains the countable family of simplices $\mathcal{S} = \{S_k^h : h \in \mathbb{N}^+, k = 1, \dots, n_h\}$ satisfying (ii) and (iii) and such that, for every $m \ge 1$, $\cup_{h=1}^{m} \cup_{k=1}^{n_h} S_k^h = \text{co}\{e_0, \dots, e_{n+m-1}\} = C_m$. Hence also (i) is fulfilled,

$$\text{int}\, C \subset \text{co}\, E = \bigcup_{m \ge 1} C_m = \bigcup \left\{S : S \in \mathcal{S}\right\},$$

and the lemma is proven. □

We are now in a position to establish our first bang-bang theorem for the inclusion (4.1).

THEOREM 4.4. *Let C be a convex compact subset in $\mathbb{R}^n$, let $\tilde{z}$ be a solution of the inclusion* (4.1). *Then, for every vector $\nu \in \mathbb{R}^n$ there exists a function $z \in W^{*,\infty}(Q, \mathbb{R}^n)$ such that*

(i) $z_{xy}(x,y) \in \partial_r C$ *(the relative boundary of* C*) for a.e.* $(x,y) \in Q$*;*

(ii) $z(x,y) = \tilde{z}(x,y)$ *for every* $(x,y) \in \partial_+ Q$*;*

(iii)$\langle z(x,y), \nu\rangle \geq \langle \tilde{z}(x,y), \nu\rangle$ *for every* $(x,y) \in Q$*.*

Proof. It is not restrictive to assume that C has dimension n. By Lemma 4.3, the interior of C is the union of countably many relative open pairwise disjoint simplices (the faces of the n-dimensional simplices in $\mathcal{S}$). Denote this family by $\{S_k : k \in \mathbb{N}\}$. Then Q can be partitioned as a union of countably many measurable sets, $Q = F \cup (\cup_{k\in\mathbb{N}} E_k)$, where

$$F = \Big\{(x,y) \in Q \, : \, \tilde{z}_{xy}(x,y) \in \partial C\Big\},$$

$$E_k = \Big\{(x,y) \in Q \, : \, \tilde{z}_{xy}(x,y) \in S_k\Big\}, \qquad k \in \mathbb{N}.$$

For every k in $\mathbb{N}$, let w_k be the function obtained by applying Lemma 2.3 for $E = E_k$, $S = S_k$, and $v = \tilde{z}_{xy}$. Let w be the function defined by $w = \tilde{z}_{xy}\chi_F + \sum_{k\in\mathbb{N}} \chi_{E_k}$. Arguing as in the proof of Theorem 3.1, one can prove that the function $z : Q \to \mathbb{R}^n$ given by

$$z(x,y) = \tilde{z}(x,c) + \tilde{z}(a,y) - \tilde{z}(a,c) + \int_a^x \int_c^y w(r,s)dsdr$$

is in $W^{*,\infty}(Q)$ and satisfies (i), (ii), and (iii). □

Remark 4.5. Notice that, if in the previous theorem $\tilde{z}$ is such that $\tilde{z}_{xy} \in \mathrm{ri}\ C$ (the relative interior of C) a.e. in Q, we obtain that $z_{xy} \in \mathrm{extr}\ C$ a.e. in Q. Thus, the previous result allows us to generalize Lemma 2.3 to the case when S is a convex relative open bounded subset of $\mathbb{R}^n$, and also Theorem 3.1 by substituting S_i, in Hypothesis H1, with a convex relative open bounded subset of $\mathbb{R}^n$. Others assumptions on S_i remain valid.

In case $n = 2$ we obtain the following bang-bang theorem in the *strong* form.

THEOREM 4.6. *Let* $\tilde{z}$ *be a solution of the inclusion* (4.1) *where* C *is a compact and convex subset of* $\mathbb{R}^2$*. Then, for every vector* $\nu \in \mathbb{R}^2$ *there exists a function* $z \in W^{*,\infty}(Q, \mathbb{R}^2)$ *such that*

(i) $z_{xy}(x,y) \in \mathrm{extr}\ C$ *for a.e.* $(x,y) \in Q$*;*

(ii) $z(x,y) = \tilde{z}(x,y)$ *for every* $(x,y) \in \partial_+ Q$*;*

(iii) $\langle z(x,y), \nu\rangle \geq \langle \tilde{z}(x,y), \nu\rangle$ *for every* $(x,y) \in Q$*.*

Proof. For $n = 2$ the boundary of C is the union of its extreme points and of its one-dimensional faces. Notice that, since the one-dimensional Hausdorff measure of ∂C is finite, the one-dimensional faces of C are at most countable. Hence, the proof is similar to that of the previous theorem, where we add to the family $\{S_k : k \in \mathbb{N}\}$ the relative interior of the one-dimensional faces of C and we call F the set

$$F = \Big\{(x,y) \in Q \, : \, \tilde{z}_{xy}(x,y) \in \mathrm{extr}\ C\Big\}. \qquad \square$$

Acknowledgments. Both authors gratefully acknowledge the financial support received from CONICYT (Chile) through the Proyecto FONDAP-Matemáticas Aplicadas and the warm hospitality of the members of their institutions.

REFERENCES

[Am-C] M. Amar and A. Cellina, *On passing to the limit for non convex variational problems*, Asymptot. Anal., 9 (1994), pp. 135–148.

[Am-Ma] M. Amar and C. Mariconda, *A nonconvex variational problem with constraints*, SIAM J. Control Optim., 33 (1995), pp. 299–307.

[A-L] V. I. Arkin and V. L. Levin, *Convexity of values of vectors integrals, theorems on measurable choice and variational problems*, Russian Math. Surveys, 27 (1972), pp. 21–85.

[Ar] Z. Artstein, *On a variational problem*, J. Math. Anal. Appl., 45 (1974), pp. 404–415.

[Au-Fr] J. P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.

[Br] A. Bressan, *A multidimensional Lyapunov type theorem*, Studia Math., 106 (1993), pp. 121–128.

[Br-F] A. Bressan and F. Flores-Bazán, *Multivariable Aumann integral and controlled wave equations*, J. Math. Anal. Appl., 189 (1995), pp. 315–334

[C-Co] A. Cellina and G. Colombo, *On a classical problem of the calculus of variations without convexity assumptions*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 97–106.

[C-F] A. Cellina and F. Flores-Bazán, *Radially symmetric solutions of a class of problems of the calculus of variations without convexity assumptions*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 9 (1992), pp. 465–478.

[Ce] L. Cesari, *Optimization-Theory and Applications*, Springer-Verlag, New York, 1983.

[Cr] G. Crasta, *An existence result for noncoercive nonconvex problems in the calculus of variations*, Nonlinear Anal., 26 (1996), pp. 1527–1533.

[Cr-M] G. Crasta and A. Malusa, *Existence results for noncoercive variational problems*, SIAM J. Control Optim., 34 (1996), pp. 2064–2076.

[DM] G. Dal Maso, *An Introduction to $\Gamma-$Convergence*, Birkhäuser, Boston, 1993.

[F1] F. Flores-Bazán, *On radial solutions to nonconvex variational problems*, Houston J. Math., 22 (1996), pp. 161–181.

[F2] F. Flores-Bazán, *Optimal Solutions in an Allocation process for a Continuum of Traders*, Technical Report 98–08, D.I.M., Universidad de Concepción, Concepción, Chile, 1998; J. Global Optim., to appear.

[F-R] F. Flores-Bazán and J. P. Raymond, *A Variational Problem Related to a Continuous-Time Allocation Process for a Continuum of Traders*, in preparation.

[Mar] P. Marcellini, *Nonconvex integrals of the calculus of variations*, in Methods of Nonconvex Analysis, Lectures Notes in Math. 1446, Springer, New York, 1990, pp. 16–57.

[Ma] C. Mariconda, *On a parametric problem of the calculus of variations without convexity assumptions*, J. Math. Anal. Appl., 170 (1992), pp. 291–297.

[Ne] L. W. Neustadt, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.

[O] C. Olech, *The Lyapunov theorem: its extensions and applications*, in Methods of Nonconvex Analysis, Lectures Notes in Math. 1446, Springer, NY, 1990, pp. 84–103.

[Pul] G. Pulvirenti, *Existence theorems for an optimal control problem relative to a linear, hyperbolic partial differential equation*, J. Optim. Theory Appl., 7 (1971), pp. 109–117.

[R1] J. P. Raymond, *Existence theorems in optimal control problems without convexity assumptions*, J. Optim. Theory Appl., 67 (1990), pp. 109–132.

[R2] J. P. Raymond, *Existence theorems without convexity assumptions for optimal control problems governed by parabolic and elliptic systems*, Appl. Math. Optim., 26 (1992), pp. 39–62.

[R3] J. P. Raymond, *Problèmes de Calcul des Variations et de Contrôle Optimal: Existence et regularité des solutions*, Thèse d'habilitation, Université Paul Sabatier, Toulouse, 1990.

[R4] J. P. Raymond, *Existence and Bang-Bang Theorems for Control Problems Governed by Hyperbolic Equations*, in Calculus of Variations, Homogenization and Continuum Mechanics, Ser. Adv. Math. Appl. Sci. 18, World Scientific, River Edge, N.J., 1993.

[Roc] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton NJ, 1972.

[S1] M. B. Suryanarayana, *Existence theorems for optimization problems concerning linear, hyperbolic partial differential equations without convexity conditions*, J. Optim. Theory Appl., 19 (1976), pp. 47–61.

[S2] M. B. Suryanarayana, *A Sobolev space and a Darboux problem*, Pacific J. Math., 69 (1977), pp. 535–550.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 6, pp. 1767–1779

STABILIZATION OF HIGH EIGENFREQUENCIES OF A BEAM EQUATION WITH GENERALIZED VISCOUS DAMPING[*]

SHENGJIA LI[†], JINGYUAN YU[‡], ZHANDONG LIANG[†], AND GUANGTIAN ZHU[§]

Abstract. In this paper, using the one-dimensional vibrating beam equation with generalized viscous damping as a model of vibration of flexible robot arms, it is shown that for such a system the high eigenmodes decay at a uniform rate. The proof is obtained by perturbation theory of linear operators [Y. H. Luo, *Acta Math. Sinica*, 32 (1991), pp. 556–563] and asymptotic estimates of eigenvectors based on an earlier work of G. B. Birkhoff and M. H. Stone. Feedback control for this class of system is investigated, and a finite-dimensional controller is presented for an exponentially stable closed-loop system. Our method may be used to study nondissipative systems.

Key words. generalized viscous damping, eigenfrequencies, feedback control, exponential decay

AMS subject classifications. 35B35, 35B37, 93B50

PII. S0363012997322546

1. Introduction.

In recent years there has been much interest in the topic of control and stabilization of vibrating equations. For example, see [6], [13], [14], and [16] for serially connected beams, [12], [20], [21], [22], [23] for vibrating strings or second order wave equations, and [1], [5], [18] for beam equations with locally distributed damping.

In this article, we consider the beam equation

$$(1.1)\qquad \begin{cases} \dfrac{\partial^2}{\partial t^2}u(x,t) + \dfrac{\partial^4}{\partial x^4}u(x,t) + b(x)\dfrac{\partial}{\partial t}u(x,t) = 0, & 0 < x < \ell, \quad t > 0, \\ u(0,t) = \dfrac{\partial}{\partial x}u(0,t) = \dfrac{\partial^2}{\partial x^2}u(\ell.t) = \dfrac{\partial^3}{\partial x^3}u(\ell,t) = 0, & t \geq 0, \\ u(x,0) = u_0(x), \qquad \frac{\partial}{\partial t}u(x,0) = u_2(x), \quad 0 \leq x \leq \ell, \end{cases}$$

where the viscous damping $b(x)$ is a continuous real function satisfying the hypothesis. Such a perturbing term takes place due to medium impurities, distributed and/or boundary friction, small viscous effects, etc.

It is well known that the analysis of damping is important in the understanding of vibrating system behavior. In [1] and [5], Chen et al. have considered the asymptotic average decay rate for the wave equation with variable coefficient viscous damping based on an earlier work of Birkhoff and Langer. For second order wave equations, a rigorous treatment of the asymptotics of the eigenvalues is given in [21], one of the conjectures in [1] is proven to be wrong in general, and the results in [20] are extended to allow sign-change damping terms, and a sufficient condition for stability in that case is given in [23]. Unfortunately, for the beam equation (1.1) the Birkhoff–Langer technique cannot be applied; information about the asymptotic average decay rate of this equation is still unknown. Whether it is valid or not is an open question [1],

[*] Received by the editors June 9, 1997; accepted for publication (in revised form) June 30, 1998; published electronically October 4, 1999. This research was supported by NSFC and the Science Foundation of Shanxi under project 69674011.

http://www.siam.org/journals/sicon/37-6/32254.html

[†] Department of Mathematics, Shanxi University, Taiyuan 030006, People's Republic of China (sjli@mail.sxu.edu.cn).

[‡] Institute of Information and Control, Beijing 100037, People's Republic of China.

[§] Institute of System Sciences, Academia Sinica, Beijing 100080, People's Republic of China.

[5]. In this paper, we prove that all high eigenmodes decay at a uniform rate. This result answers the conjecture [1] in the generalized case: there exists an $n_1 > 0$, such that $\text{Re}\lambda_{nj} < -\frac{\delta_0}{2} + \varepsilon$ for $n > n_1$, $j = 1, 2$, where $\varepsilon : 0 < \varepsilon < \frac{\delta_0}{2}$ and $\lambda_{nj} \in \sigma_0(A)$ (the spectrum of A). Finally, in order to stabilize the system exponentially, a finite-dimensional controller is presented. The main difference between the work of this paper and that of [1], [5], and others is that our condition on $b(x)$ is not necessarily dissipative but only "more positive than negative [5]." Our method may be used to study nondissipative systems.

In order to study (1.1), we define the differential operator A_0 as follows:

$$\begin{cases} A_0\phi(x) = \frac{d^4}{dx^4}\phi(x), & \phi(x) \in D(A_0), \\ D(A_0) = \{\phi; \phi \in H^4(0,\ell), & \text{and} \quad \phi(0) = \phi'(0) = \phi''(\ell) = \phi'''(\ell) = 0\}. \end{cases}$$

Obviously A_0 is positive in $L^2(0,\ell)$. By the equivalent transformation $y_1(t) = u(t)$, $y_2(t) = \frac{d}{dt}u(t)$, (1.1) becomes

$$\begin{cases} \dfrac{d}{dt}\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = A \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix}, \quad t > 0, \\ \begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \end{cases} \tag{1.2}$$

where $A = \left(\begin{smallmatrix} 0 & I \\ -A_0 & -b(x) \end{smallmatrix}\right)$ and $D(A) = D(A_0) \times D(A_0^{\frac{1}{2}})$. The energy inner product is defined on $D(A)$ by $((\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}), (\begin{smallmatrix} v_1 \\ v_2 \end{smallmatrix})) = (A_0^{\frac{1}{2}}u_1, A_0^{\frac{1}{2}}v_1) + (u_2, v_2)$ for $u_1, v_1 \in D(A_0^{\frac{1}{2}})$, $u_2, v_2 \in D(A_0)$, and the Hilbert space $\mathbf{H}$ is defined as the closure of D(A) in this energy inner product.

Hypothesis. We assume that $b(x) \in C[0,\ell]$ and satisfies

$$\int_0^l b(x)|\phi_{nj}(x)|^2 dx \geq \delta_0 > 0, \qquad j = 1, 2, \quad n > n_0, \tag{1.3}$$

where $\delta_0 > 0$ is a constant, n_0 is a positive integer, and ϕ_{nj}, $j = 1, 2$, $n = 1, 2, \ldots,$ are the orthonormalized eigenfunctions of the eigenequation

$$\begin{cases} \dfrac{\partial^2}{\partial t^2}u(x,t) + \dfrac{\partial^4}{\partial x^4}u(x,t) = 0, & 0 < x < \ell, \quad t > 0, \\ u(0,t) = \dfrac{\partial}{\partial x}u(0,t) = \dfrac{\partial^2}{\partial x^2}u(\ell,t) = \dfrac{\partial^3}{\partial x^3}u(\ell,t) = 0, & t \geq 0. \end{cases} \tag{1.1$'$}$$

The assumptions on the damping coefficient $b(x)$ are very weak; the value of $b(x)$ can even be negative on some small parts of the interval $[0,\ell]$.

The organization of this paper is as follows. In section 2 we obtain the asymptotic formula for the eigenvectors of A. In section 3 we get the stronger result that the operator $A = \begin{pmatrix} 0 & I \\ -A_0 & -b(x) \end{pmatrix}$ generates a C_0-semigroup and all generalized eigenvectors of A form an unconditional basis of H. In section 4 we prove that, for a given $\varepsilon : 0 < \varepsilon < \frac{\delta_0}{2}$, there exists an $n_1 > n_0$ such that $\text{Re}\lambda_{nj} < -\frac{\delta_0}{2} + \varepsilon$ for $n > n_1$, where $\lambda_{nj} \in \sigma(A)$. In section 5, a finite-dimensional controller is presented for exponentially stable closed-loop system.

2. Asymptotic estimations of the eigenvectors of $\boldsymbol{A}$. In this section, we will estimate the eigenvectors of A. For convenience, we suppose that $\ell = 1$. The following is obvious.

Lemma 2.1. *The eigenequation of* (1.1) *is equivalent to the eigenequation of* (1.2). *That is, a complex λ satisfies the eigenequation*

$$\begin{cases} \lambda^2\phi(x) + \lambda b(x)\phi(x) + \phi^{(4)}(x) = 0, & x \in (0,1), \\ \phi(0) = \phi'(0) = \phi''(1) = \phi'''(1) = 0, \end{cases} \tag{2.1}$$

if and only if λ satisfies the eigenequation

$$\lambda \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} = A \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}.$$

We only need to estimate those of the quadratic eigenfunction (2.1). *For convenience, we divide the complex plane into*

$$S_n = \left\{ \rho;\ \frac{n\pi}{8} \le \arg\rho \le \frac{n+1}{8}\pi \right\}, \quad n = 0,1,2,3,4,5,6,7.$$

Then, for every sector S_n, we can arrange ω_1, ω_2, ω_3, ω_4(the roots of $\omega^4 = -1$), such that

$$\operatorname{Re}(\rho\omega_1) \le \operatorname{Re}(\rho\omega_2) \le \operatorname{Re}(\rho\omega_3) \le \operatorname{Re}(\rho\omega_4) \quad \textit{for } \rho \in S_n. \tag{2.2}$$

In fact, for S_0, we choose $\omega_1 = e^{\frac{3\pi}{4}i}, \omega_2 = e^{\frac{5\pi}{4}i}, \omega_3 = e^{\frac{\pi}{4}i}, \omega_4 = e^{-\frac{\pi}{4}i}$. We can verify that (2.2) *is true for $\rho \in S_0$. Similarly, we can see that* (2.2) *holds for S_n, $1 \le n \le 7$. Let $\lambda = \rho^2$. For* (2.1), *we have a general result.*

Lemma 2.2. *The equation*

$$\phi^{(4)}(x) + \rho^2 b(x)\phi(x) + \rho^4\phi(x) = 0, \quad x \in [0,1],\ \rho \in S_n,\ |\rho| > b_0,\ 0 \le n \le 7, \tag{2.3}$$

has four linearly independent solutions, ϕ_k, $k = 1,2,3,4$, which satisfy

$$\frac{d^j}{dx^j}\phi_k(x,\rho) = \rho^j e^{\rho\omega_k x}\left[\omega_k^j + O\left(\frac{1}{\rho}\right)\right], \quad j = 0,1,2,3, \tag{2.4}$$

where $b_0 = \max|b(x)|$ and ω_k, $k = 1,2,3,4$, are the four roots of the equation $\omega^4 = -1$ satisfying (2.2). *If $b(x) = 0$, then $O(\frac{1}{\rho}) = 0$.*

Proof. Because $e^{\rho\omega_j x}$, $j = 1,2,3,4$, are the basic solutions of the equation $\phi^{(4)} + \lambda^2\phi = 0$, by the method of variation of constants, the solution of (2.3) can be represented as

$$y(x,\rho) = \sum_{j=1}^{4} c_j e^{\rho\omega_j x} + \frac{1}{4\rho}\int_0^x \left(\sum_{j=1}^{4} \omega_j e^{\rho\omega_j(x-\xi)}\right) b(\xi)y(\xi)d\xi, \tag{2.5}$$

where c_j, $j = 1,2,3,4$, are arbitrary constants. For a fixed k, we set

$$\begin{aligned} &c_j' = c_j, && 1 \le j \le k, \\ &c_j' = c_j + \tfrac{1}{4\rho}\textstyle\int_0^1 \omega_j e^{-\rho\omega_j\xi} b(\xi)y(\xi)d\xi, && k+1 \le j \le 4. \end{aligned} \tag{2.6}$$

Then (2.5) can be rewritten as

$$\begin{aligned} y(x,\rho) = \sum_{j=1}^{4} c'_j e^{\rho\omega_j x} + \tfrac{1}{4\rho}\int_0^x \left(\sum_{j=1}^{k}\omega_j e^{\rho\omega_j(x-\xi)}\right) b(\xi)y(\xi)d\xi \\ - \frac{1}{4\rho}\int_x^1 \left(\sum_{j=k+1}^{4}\omega_j e^{\rho\omega_j(x-\xi)}\right) b(\xi)y(\xi)d\xi. \end{aligned} \tag{2.7}$$

Taking $c'_k = 1$, $c'_n = 0$, $n \neq k$, we get

$$\begin{aligned} y_k(x,\rho) = e^{\rho\omega_k x} + \frac{1}{4\rho}\int_0^x \left(\sum_{j=1}^{k}\omega_j e^{\rho\omega_j(x-\xi)}\right) b(\xi)y_k(\xi)d\xi \\ - \frac{1}{4\rho}\int_x^1 \left(\sum_{j=k+1}^{4}\omega_j e^{\rho\omega_j(x-\xi)}\right) b(\xi)y_k(\xi)d\xi\,, \quad k=1,2,3,4. \end{aligned} \tag{2.8}$$

If $y_k(x,\rho), 1 \le k \le 4$, are solutions of (2.2), differentiating y_k with respect to x, we get

$$\begin{aligned} \frac{d^j}{dx^j}y_k(x,\rho) = (\rho\omega_k)^j e^{\rho\omega_k x} + \frac{1}{4\rho}\int_0^x \left(\sum_{\nu=1}^{k}\omega_\nu^{j+1}\rho^j e^{\rho\omega_\nu(x-\xi)}\right) b(\xi)y_k(\xi)d\xi \\ - \frac{1}{4\rho}\int_x^1 \left(\sum_{\nu=k+1}^{4}\omega_\nu^{j+1}\rho^j e^{\rho\omega_\nu(x-\xi)}\right) b(\xi)y_k(\xi)d\xi, \\ k=1,2,3,4,\ j=0,1,2,3. \end{aligned} \tag{2.9}$$

Let

$$\frac{d^j}{dx^j}y_k(x,\rho) = \rho^j e^{\rho\omega_k x} z_{kj}(x,\rho)\,, \quad k=1,2,3,4,\ j=0,1,2,3; \tag{2.10}$$

we get

$$\begin{aligned} z_{kj}(x,\rho) = \omega_k^j + \frac{1}{4\rho}\int_0^x \left(\sum_{\nu=1}^{k}\omega_\nu^{j+1} e^{\rho(\omega_\nu-\omega_k)(x-\xi)}\right) b(\xi)z_{k0}(\xi,\rho)d\xi, \\ - \frac{1}{4\rho}\int_x^1 \left(\sum_{\nu=k+1}^{4}\omega_\nu^{j+1} e^{\rho(\omega_k-\omega_\nu)(\xi-x)}\right) b(\xi)z_{k0}(\xi,\rho)d\xi, \\ k=1,2,3,4,\ j=0,1,2,3. \end{aligned} \tag{2.11}$$

Let

$$K_{kj0}(x,\xi,\rho) = \begin{cases} \dfrac{1}{4}\left(\displaystyle\sum_{\nu=1}^{k}\omega_\nu^{j+1} e^{\rho(\omega_\nu-\omega_k)(x-\xi)}\right) b(\xi)\,, & \xi < x, \\ -\dfrac{1}{4}\left(\displaystyle\sum_{\nu=k+1}^{4}\omega_\nu^{j+1} e^{\rho(\omega_k-\omega_\nu)(\xi-x)}\right) b(\xi)\,, & \xi > x, \end{cases} \tag{2.12}$$

$$k=1,2,3,4,\ j=0,1,2,3.$$

From (2.12), (2.11) can be rewritten as

$$z_{kj}(x,\rho)=\omega_k^j+\frac{1}{\rho}\int_0^1 K_{kj0}(x,\xi,\rho)z_{k0}(\xi,\rho)d\xi\,,\quad k=1,2,3,4,\ j=0,1,2,3. \tag{2.13}$$

From inequality (2.2), we have

$$|e^{-\rho(\omega_k-\omega_\nu)(x-\xi)}|=e^{(x-\xi)^{\mathrm{Re}}(\rho\omega_\nu-\rho\omega_k)}\le 1\quad \text{for}\ \ \nu\le k,\ \xi<x,\ \rho\in S_n,$$

and

$$|e^{-\rho(\omega_k-\omega_\nu)(x-\xi)}|=e^{(x-\xi)^{\mathrm{Re}}(\rho\omega_\nu-\rho\omega_k)}\le 1\quad \text{for}\ \ \nu> k,\ \xi>x,\ \rho\in S_n.$$

So

$$|K_{kj0}(x,\xi,\rho)|\le \max\nolimits_{0\le\xi\le 1}|b(\xi)|=b_0. \tag{2.14}$$

By Picard's iteration, we see that (2.12) has a unique solution

$$z_{kj}(x,\rho)\quad \text{for}\ \ \rho\in S_n \text{ and } |\rho|>b_0,$$

satisfying

$$z_{kj}(x,\xi)=\omega_k^j+O\left(\frac{1}{\rho}\right)\quad \text{as}\ \ \rho\longrightarrow\infty. \tag{2.15}$$

From (2.15), we get (2.4).

Finally, we prove that (2.8) is the solution of (2.3) for $|\rho|$ large enough. For this, in the following, we prove (2.7) is a solution of (2.3) for every (c_1',c_2',c_3',c_4') and $|\rho|>b_0$. Because (2.6) is a linear mapping from (c_1,c_2,c_3,c_4) to (c_1',c_2',c_3',c_4') , we prove only that mapping (2.6) is invertible for $|\rho|>b_0$. Supposing (2.6) had a solution $c_j\neq 0$ for $c_j'=0$, the equation

$$\begin{aligned} y(x,\rho)=&\frac{1}{4\rho}\int_0^x\left(\sum_{\nu=1}^k\omega_\nu e^{\rho\omega_\nu(x-\xi)}\right)b(\xi)y(\xi,\rho)d\xi\\ &-\frac{1}{4\rho}\int_0^x\left(\sum_{\nu=k+1}^4\omega_\nu e^{\rho\omega_\nu(x-\xi)}\right)b(\xi)y(\xi,\rho)d\xi \end{aligned} \tag{2.16}$$

has a nonzero solution. Differentiating (2.16), we get

$$\frac{d^j}{dx^j}y(x,\rho)=\rho^j e^{\rho\omega_k x}z_j(x,\rho)\,,\quad j=0,1,2,3, \tag{2.17}$$

where z_j satisfy

$$\begin{aligned} z_j(x,\rho)=&\frac{1}{4\rho}\int_0^x\left(\sum_{\nu=1}^k\omega_\nu e^{-\rho(\omega_k-\omega_\nu)(x-\xi)}\right)b(\xi)z_0(\xi,\rho)d\xi\\ &-\frac{1}{4\rho}\int_0^x\left(\sum_{\nu=k+1}^4\omega_\nu e^{\rho(\omega_\nu-\omega_k)(x-\xi)}\right)b(\xi)z_0(\xi,\rho)d\xi,\ j=0,1,2,3. \end{aligned} \tag{2.18}$$

Let $M(\rho)=\max_{0\le x\le 1}|z_j(x,\rho)|\,,\, j=0,1,2,3$. We get

$$|z_j(x,\rho)|\le\left[\frac{b_0k}{4|\rho|}+\frac{b_0(4-k)}{4|\rho|}\right]M(\rho)\,,\, j=0,1,2,3. \tag{2.19}$$

That is,

$$M(\rho)\le\frac{b_0}{|\rho|}M(\rho). \tag{2.20}$$

From (2.20), we see that $M(\rho)=0$, and $z_j(x,\rho)=0$ $(j=0,1,2,3)$ for $|\rho|>b_0$. Then $y(x,\rho)=e^{\rho\omega_k x}z_0(x,\rho)=0$ for $|\rho|>b_0$. This is a contradiction with $y(x,\rho)\neq 0$. Therefore mapping (2.7) is the solution of (2.3). The proof is complete. □

In order to get an asymptotic formula for eigenfunctions of (2.1), we first estimate the eigenvalues of (2.1).

Lemma 2.3. *The eigenvalues of* (2.1) *are the following sequences:*

$$\begin{aligned}\lambda_{n1}&=\left(-n+\frac{1}{2}\right)^2\pi^2\left[i+O\left(\frac{1}{n^2}\right)\right], \quad n=1,2,\dots,\\ \lambda_{n2}&=\left(n+\frac{1}{2}\right)^2\pi^2\left[i+O\left(\frac{1}{n^2}\right)\right], \quad n=1,2,\dots.\end{aligned} \tag{2.21}$$

Proof. Let $\rho^4=\lambda^2$ and $S_0=\{\rho\ ;\ \ -\frac{\pi}{4}\le\arg\rho\le 0\}$. Taking $\omega_1=e^{\frac{5\pi i}{4}}$, $\omega_2=e^{\frac{3\pi i}{4}}$, $\omega_3=e^{\frac{7\pi i}{4}}$, $\omega_4=e^{\frac{\pi i}{4}}$, then $\mathrm{Re}\omega_1\rho\le\mathrm{Re}\omega_2\rho\le\mathrm{Re}\omega_3\rho\le\mathrm{Re}\omega_4\rho$ for $\rho\in S_0$, and $e^{\omega_1\rho}\to 0, e^{\omega_4\rho}\to\infty$ $(\rho\to\infty,\rho\in S_0)$ exponentially. By Lemma 2.2 for $\rho\in S_0$ (2.2) has four independent solutions ϕ_1, ϕ_2, ϕ_3, ϕ_4, which satisfy (2.4). By (2.4), we get

$$\begin{aligned}&\phi_k(0)=1+O\left(\frac{1}{\rho}\right), && \phi_k'(0)=\rho\omega_k\left[1+O\left(\frac{1}{\rho}\right)\right],\\ &\phi_k''(1)=\rho^2\omega_k^2e^{\rho\omega_k}\left[1+O\left(\frac{1}{\rho}\right)\right], && \phi_k'''(1)=\rho^3\omega_k^3e^{\rho\omega_k}\left[1+O\left(\frac{1}{\rho}\right)\right],\\ &&& k=1,2,3,4.\end{aligned} \tag{2.22}$$

From this estimation, the eigendeterminant $\Delta(\lambda)=0$ can be represented as

$$\begin{vmatrix}\omega_1[1] & \omega_2[1] & \omega_3[1] & \omega_4[1]\\ [1] & [1] & [1] & [1]\\ (\omega_1)^3e^{\omega_1\rho}[1] & (\omega_2)^3e^{\omega_2\rho}[1] & (\omega_3)^3e^{\omega_3\rho}[1] & (\omega_4)^3e^{\omega_4\rho}[1]\\ (\omega_1)^2e^{\omega_1\rho}[1] & (\omega_2)^2e^{\omega_2\rho}[1] & (\omega_1)^2e^{\omega_3\rho}[1] & (\omega_4)^2e^{\omega_4\rho}[1]\end{vmatrix}, \tag{2.23}$$

where $[1]=1+O(\frac{1}{\rho})$. By computing, we get

$$\Delta(\lambda)=-2(e^{\rho\omega_3}+e^{\rho\omega_2})+e^{\rho\omega_3}O\left(\frac{1}{\rho}\right). \tag{2.24}$$

Noting that $\omega_1=-\omega_4,\omega_2=-\omega_3$, (2.24) becomes

$$e^{\rho\omega_2}\Delta(\lambda)=-2(e^{2\rho\omega_2}+1)+O\left(\frac{1}{\rho}\right). \tag{2.25}$$

Because equations $e^{2\rho\omega_2} + 1 = 0$ have roots

$$\rho_{n1} = \omega_2^{-1}\left(n + \frac{1}{2}\right)\pi i, \quad n = 0, 1, 2, 3, \ldots, \tag{2.26}$$

by Rouche's theorem, we can see that the equation $\Delta(\lambda) = 0$ has a sequence of simple roots represented as

$$\hat{\rho}_{n1} = \rho_{n1} + O\left(\frac{1}{n}\right), \quad n = 1, 2, 3, \ldots. \tag{2.27}$$

From (2.27), we get the second representation of (2.21).

Taking $S_1 = \{\rho\ ;\ 0 \leq \arg\rho \leq \frac{\pi}{4}\}$, $\omega_1 = e^{\frac{3\pi i}{4}}$, $\omega_2 = e^{\frac{5\pi i}{4}}$, $\omega_3 = e^{\frac{\pi i}{4}}$, $\omega_4 = e^{-\frac{\pi i}{4}}$ and using the above method, we can prove the other sequence in (2.21). The proof is complete. □

Corollary 2.4. *The eigenvalues of* (1.1′) *satisfy the following asymptotic formula*

$$\begin{aligned} \lambda_{n1} &= (-n + \tfrac{1}{2})^2\pi^2[i + O_1(\tfrac{1}{n^2})], \quad n = 1, 2, \ldots, \\ \lambda_{n2} &= (n + \tfrac{1}{2})^2\pi^2[i + O_2(\tfrac{1}{n^2})], \qquad n = 1, 2, \ldots. \end{aligned} \tag{2.22′}$$

By Lemmas 2.2 *and* 2.3, *we can get the asymptotic estimate of the eigenvectors of* (2.1).

Lemma 2.5. *There exist eigenvectors of* (2.1) *satisfying the following asymptotic formula:*

$$\begin{gathered} y_{n1}(x) = -\sqrt{2}e^{(-n+\frac{1}{2})\pi ix} - i\sqrt{2}e^{(n-\frac{1}{2})i\pi x} + O(\tfrac{1}{n}), \\ y_{n2}(x) = \sqrt{2}e^{-(n+\frac{1}{2})\pi ix} + i\sqrt{2}e^{(n+\frac{1}{2})\pi ix} + O(\tfrac{1}{n}), \\ n = 1, 2, 3, \ldots, \quad x \in [a, b] \subset (0, 1), \end{gathered} \tag{2.28}$$

where a and b are two arbitrary constants which satisfy $0 < a < b < 1$.

Furthermore, there exists a constant $M > 0$ such that $|y_{nj}(x)| \leq M$ for $x \in [0, 1]$ and $j = 1, 2,\ n = 1, 2, \ldots$.

Proof. Let $\rho^4 = \lambda^2$, $\omega_1 = e^{\frac{3\pi i}{4}}$, $\omega_2 = e^{\frac{5\pi i}{4}}$, $\omega_3 = e^{\frac{\pi i}{4}}$, $\omega_4 = e^{\frac{7\pi i}{4}}$. We consider Lemma 2.4 on the sector $S_0 = \{\rho\ ;\ 0 \leq \arg\rho \leq \frac{\pi}{4}\}$. Because $\phi_j(x, \lambda) = e^{\rho\omega_j x}[1 + O(\frac{1}{\rho})]$, $j = 1, 2, 3, 4$, are the basic solutions of the equation $\phi^{(4)}(x) + \lambda b(x)\phi(x) + \lambda^2\phi(x) = 0$, the eigenvectors corresponding to the eigenvalue $\lambda_{nk}, k = 1, 2, n = 1, 2, \ldots$, are $y_{nk}(x) = y(x, \lambda_{nk}) = \sum_{j=1}^4 c_j\phi_j(x, \lambda_{nk})$, and c_j $(j = 1, 2, 3, 4)$ satisfy the following equations:

$$\begin{cases} \phi_1'(0, \lambda_{nk})c_1 + \phi_2'(0, \lambda_{nk})c_2 + \phi_3'(0, \lambda_{nk})c_3 + \phi_4'(0, \lambda_{nk})c_4 = 0, \\ \phi_1(0, \lambda_{nk})c_1 + \phi_2(0, \lambda_{nk})c_2 + \phi_3(0, \lambda_{nk})c_3 + \phi_4(0, \lambda_{nk})c_4 = 0, \\ \phi_1'''(1, \lambda_{nk})c_1 + \phi_2'''(1, \lambda_{nk})c_2 + \phi_3'''(1, \lambda_{nk})c_3 + \phi_4'''(1, \lambda_{nk})c_4 = 0, \\ \phi_1''(1, \lambda_{nk})c_1 + \phi_2''(1, \lambda_{nk})c_2 + \phi_3''(1, \lambda_{nk})c_3 + \phi_4''(1, \lambda_{nk})c_4 = 0\,. \end{cases} \tag{2.29}$$

By the theory of linear algebra, we obtain

$$y_{nk}(x) = \begin{vmatrix} \phi_1(x, \lambda_{nk}) & \phi_2(x, \lambda_{nk}) & \phi_3(x, \lambda_{nk}) & \phi_4(x, \lambda_{nk}) \\ \phi_1(0, \lambda_{nk}) & \phi_2(0, \lambda_{nk}) & \phi_3(0, \lambda_{nk}) & \phi_4(0, \lambda_{nk}) \\ \phi_1'''(1, \lambda_{nk}) & \phi_2'''(1, \lambda_{nk}) & \phi_3'''(1, \lambda_{nk}) & \phi_4'''(1, \lambda_{nk}) \\ \phi_1''(1, \lambda_{nk}) & \phi_2''(1, \lambda_{nk}) & \phi_3''(1, \lambda_{nk}) & \phi_4''(1, \lambda_{nk}) \end{vmatrix}. \tag{2.30}$$

By the representation of $\phi_{nk}(x)$ for $x \in [a,b]$, $y_{nk}(x)$ can be rewritten as

$$y_{nk}(x) = \begin{vmatrix} e^{\omega_1\rho_{nk}x}[1] & e^{\omega_2\rho_{nk}x}[1] & e^{-\omega_2\rho_{nk}x}[1] & e^{\omega_4\rho_{nk}(x-1)}[1] \\ 1+O(\frac{1}{n}) & 1+O(\frac{1}{n}) & 1+O(\frac{1}{n}) & O(\frac{1}{n}) \\ (\omega_1)^3e^{\omega_1\rho_{nk}}[1] & (\omega_2)^3e^{\omega_2\rho_{nk}}[1] & (\omega_3)^3e^{\omega_3\rho_{nk}}[1] & (\omega_4)^3[1] \\ (\omega_1)^2e^{\omega_1\rho_{nk}}[1] & (\omega_2)^2e^{\omega_2\rho_{nk}}[1] & (\omega_1)^2e^{\omega_3\rho_{nk}}[1] & (\omega_4)^2[1] \end{vmatrix}$$

$$\begin{aligned} &= \begin{vmatrix} e^{\omega_1\rho_{nk}x} & e^{\omega_2\rho_{nk}x} & e^{-\omega_2\rho_{nk}x} & 0 \\ 1 & 1 & 1 & 0 \\ 0 & (\omega_2)^3e^{\omega_2\rho_{nk}} & (\omega_3)^3e^{\omega_3\rho_{nk}} & (\omega_4)^3 \\ 0 & (\omega_2)^2e^{\omega_2\rho_{nk}} & (\omega_3)^2e^{\omega_3\rho_{nk}} & (\omega_4)^2 \end{vmatrix} + O\left(\frac{1}{n}\right) \\ &= e^{\omega_1\rho_{nk}x}\left[\begin{vmatrix} \omega_3 & \omega_4 \\ 1 & 1 \end{vmatrix}\omega_3^2\omega_4^2e^{\omega_3\rho_{nk}} - \begin{vmatrix} \omega_2 & \omega_4 \\ 1 & 1 \end{vmatrix}\omega_2^2\omega_4^2e^{\omega_2\rho_{nk}}\right] \\ &\quad + e^{\omega_2\rho_{nk}}e^{-\omega_2\rho_{nk}x}\begin{vmatrix} \omega_2 & \omega_4 \\ 1 & 1 \end{vmatrix}\omega_2^2\omega_4^2 - e^{-\omega_2\rho_{nk}}e^{\omega_2\rho_{nk}x}\begin{vmatrix} \omega_3 & \omega_4 \\ 1 & 1 \end{vmatrix}\omega_3^2\omega_4^2 \\ &\quad + O\left(\frac{1}{n}\right). \end{aligned} \tag{2.31}$$

Taking $k=1,\ n=1,2,\dots,$ in (2.31) we get

$$y_{n1}(x) = (i+1)\sqrt{2}e^{\omega_1\omega_2^{-1}(-n+\frac{1}{2})\pi ix} - \sqrt{2}e^{(-n+\frac{1}{2})\pi ix} - i\sqrt{2}e^{(n-\frac{1}{2})\pi ix} + O\left(\frac{1}{n}\right). \tag{2.32}$$

Taking $\omega_1 = e^{\frac{5\pi i}{4}}$, $\omega_2 = e^{\frac{3\pi i}{4}}$, $\omega_3 = e^{\frac{7\pi i}{4}}$, $\omega_4 = e^{\frac{1\pi i}{4}}$, we consider Lemma 2.5 on the sector $S_0 = \{\rho\ ;\ -\frac{\pi}{4} \le \arg\rho \le 0\}$. By the above method, we can get

$$y_{n2}(x) = -(i+1)\sqrt{2}e^{\omega_1\omega_2^{-1}(n+\frac{1}{2})\pi ix} + \sqrt{2}e^{-(n+\frac{1}{2})\pi ix} + i\sqrt{2}e^{(n+\frac{1}{2})\pi ix} + O\left(\frac{1}{n}\right). \tag{2.33}$$

The proof is complete. □

By Lemma 2.2 and (2.30), we get the following corollary.

COROLLARY 2.6. *The eigenfunctions of* $(1.1')$ *satisfy the following asymptotic formula:*

$$\begin{gathered} y_{n1}(x) = -\sqrt{2}e^{(-n+\frac{1}{2})\pi ix} - i\sqrt{2}e^{(n-\frac{1}{2})i\pi x} + O_1(\tfrac{1}{n}), \\ y_{n2}(x) = \sqrt{2}e^{-(n+\frac{1}{2})\pi ix} + i\sqrt{2}e^{-(n+\frac{1}{2})\pi ix} + O_2(\tfrac{1}{n}), \\ n = 1,2,3,\dots, \quad x \in [a,b] \subset (0,1), \end{gathered} \tag{2.34}$$

where a and b are two arbitrary constants: $0 < a < b < 1$.

Furthermore, there exists a constant $M > 0$ such that $|y_{nj}(x)| \le M$ for $x \in [0,1]$ and $j = 1,2,\ n = 1,2,\dots.$

3. Properties of operator *A*. In this section, we discuss some properties of A. Let $\{\Psi_{nj}\}$, $j = 1,2$, $n = 1,2,\dots$, be the set of all generalized eigenvectors of A. We have the following result.

THEOREM 3.1. *A is a generator of a C_0-semigroup $T(t)$, and all of the eigenvectors $\{\Psi_{nj}\}$ $(j = 1,2, n = 1,2,\dots)$ of A form an unconditional basis of H.*

Proof. Obviously A generates a C_0-semigroup $T(t)$. By Corollary 2.4, $\mu_{nj} = (\frac{1}{2}+(-1)^j n)^2[i+O_1(\frac{1}{n})]$, $j = 1, 2$, $n = 1, 2, \dots$, are the eigenvalues of $A_1 = \begin{pmatrix} 0 & I \\ -A_0 & 0 \end{pmatrix}$, and the eigenvectors $\begin{pmatrix} \phi_n \\ \mu_{nj}\phi_n \end{pmatrix}$ of A_1 corresponding to μ_{nj}, $j = 1, 2$, $n = 1, 2, \dots$, form an orthogonal basis of H. Then A_1 is a (D)-operator [3]; that is, A_1 has a discrete spectrum and its eigenvectors $\begin{pmatrix} \phi_n \\ \mu_n\phi_n \end{pmatrix}$ form a basis of H. Because

$$\left\| \begin{pmatrix} 0 & 0 \\ 0 & -b(x) \end{pmatrix} \Phi \right\| \le b_0\|\phi_2\| = b_0\|A_1^0\|\|\Phi\|, \qquad \Phi \in D(A_1), \tag{3.1}$$

where $A_1^0 = I$, $\Phi = (\phi_1, \phi_2)^T$, $b_0 = \max_{x\in[0,1]} |b(x)|$, by Corollary 11 [2], A is also a (D)-operator. That is, A has also a discrete spectrum and its generalized eigenvectors $\{\Psi_{nj}\}$ ($j = 1, 2$, $n = 1, 2, \dots$) form a basis of H. □

4. The estimate for the real parts of the eigenvalues of A.

We now estimate the real parts of the eigenvalues of A.

LEMMA 4.1. *Let* $\sigma(A) = \{\mu_{nj};\quad j = 1, 2, n = 1, 2, 3, \dots\}$. *There exists a constant* n_0 *such that* $\mathrm{Re}\mu_{nj} < -\frac{\delta_0}{2} + \varepsilon$ *for* $n > n_0$, *where* $0 < \varepsilon < \frac{\delta_0}{2}$.

Proof. Let $\begin{pmatrix} \phi_{1nj} \\ \phi_{2nj} \end{pmatrix}$, $j = 1, 2, n = 1, 2, \dots$, be the eigenvector of A corresponding to the eigenvalue μ_{nj}. By Lemma 2.1, $\phi_{2nj} = \mu_{nj}\phi_{1nj}$ and ϕ_{1nj} satisfies the equation

$$\begin{cases} \phi_{1nj}^{(4)}(x) + \mu_{nj}b(x)\phi_{1nj}(x) + \mu_{nj}^2\phi_{1nj}(x) = 0, \\ \phi_{1nj}(0) = \phi'_{1nj}(0) = \phi''_{1nj}(1) = \phi'''_{1nj}(1) = 0\,. \end{cases} \tag{4.1}$$

Forming the inner product of (4.1) with ϕ_{1nj}, we obtain

$$\mu_{nj}^2\|\phi_{1nj}\|^2 + \mu_{nj}\int_0^1 b(x)|\phi_{1nj}(x)|^2dx + \|\phi''_{1nj}\|^2 = 0. \tag{4.2}$$

Equation (4.2) is equivalent to

$$2(\mathrm{Re}\mu_{nj})(\mathrm{Im}\mu_{nj})\|\phi_{1nj}\|^2 + (\mathrm{Im}\mu_{nj})\int_0^1 b(x)|\phi_{1nj}(x)|^2dx = 0 \tag{4.3}$$

and

$$((\mathrm{Re}\mu_{nj})^2 - (\mathrm{Im}\mu_{nj})^2)\|\phi_{1nj}\|^2 + (\mathrm{Re}\mu_{nj})\int_0^1 b(x)|\phi_{1nj}|^2dx + \|\phi''_{1nj}\|^2 = 0. \tag{4.4}$$

For convenience, we let $\|\phi_{1nj}\| = 1$. From (4.3) and (4.4) we get

$$\begin{aligned} \mathrm{Re}\mu_{nj} &= -\tfrac{1}{2}\textstyle\int_0^1 b(x)|\phi_{1nj}|^2dx \\ &\quad \pm \tfrac{1}{2}\sqrt{(\textstyle\int_0^1 b(x)|\phi_{1nj}(x)|^2dx)^2 - 4\|\phi''_{1nj}\|^2\|\phi_{1nj}\|^2}, && \text{if } \mathrm{Im}\mu_{nj} = 0, \\ \mathrm{Re}\mu_{nj} &= -\tfrac{1}{2}\textstyle\int_0^1 b(x)|\phi_{1nj}(x)|^2dx, && \text{if } \mathrm{Im}\mu_{nj} \neq 0. \end{aligned} \tag{4.5}$$

Because $|\mu_{nj}| \to \infty$, by (4.5), (2.25), the above corollaries, and the hypothesis, we see that there exists a constant n_0 such that $\mathrm{Re}\mu_{nj} < -\frac{\delta_0}{2} + \epsilon$, for $n \ge n_0$. Therefore the proof is complete. □

From Lemma 4.1, we can get the following decomposition result.

THEOREM 4.2. *The system* (1.2) *can be decomposed into*

$$\tfrac{d}{dt}Y_{n_1}(t) = A_{n_1}Y_{n_1}(t), \quad Y_{n_1} = P_{n_1}Y_0, \quad t > 0, \tag{4.6}$$

$$\frac{d}{dt}Y_r(t) = A_rY_r(t), \quad Y_r = P_rY_0 \quad , t > 0, \tag{4.7}$$

where P_{n_1} is a projection; $P_r = I - P_{n_1}$, $Y_{n_1} = P_{n_1}Y$, $Y_r = P_rY$, $A_{n_1} = P_{n_1}AP_{n_1}$, and $A_r = P_rAP_r$. In addition, A_r generates a semigroup $T_r(t) = P_rT(t)P_r$ with the growth property

$$T_r(t) \leq M_r(\varepsilon)e^{-(\delta_1-\varepsilon)t}, \quad t \geq 0, \quad 0 < \varepsilon < \delta_1 = \frac{\delta_0}{2}, \tag{4.8}$$

where $M_r(\varepsilon) \geq 1$ is a constant with respect to P_r, ε, and δ_0.

Proof. According to Lemma 4.1, there are only $2n_1$ eigenvalues, $\mu_{nj} \in \sigma(A)$, such that Re$\mu_{nj} > -\delta_1 + \epsilon$. We assume that they are $\mu_{11}, \mu_{12}, \ldots, \mu_{n_1 1}, \mu_{n_1 2}$. Let $P_{n_1} = (2i\pi)^{-1}\sum_{n=1}^{n_1}\sum_{j=1}^{2}\oint_{C_{nj}} R(\lambda; A)d\lambda$, where $C_{nj}, j = 1, 2, n = 1, 2, \ldots, n_0$, is a circle, centered at μ_{nj}, such that no other μ_{nj} belongs to its interior. Obviously, $P_r = \sum_{\mu_{nj}\in\sigma_0(A)}(2i\pi)^{-1}\oint_{V_{nj}} R(\lambda; A)d\lambda$, where $\sigma_0(A) = \sigma(A)\backslash\{\mu_{nj}, j = 1, 2; n = 1, 2, \ldots, n_1\}$, V_{nj} is a small circle defined similarly to C_{nj}. Let $Y_{n_1} = P_{n_1}Y$, $Y_r = P_rY$, $A_{n_1} = P_{n_1}AP_{n_1}$, $A_r = P_rAP_r$; we can obtain (4.6) and (4.7). We now prove (4.8). By the definition of P_{n_1} and P_r, we know that $H = (P_{n_1}H)\bigoplus(P_rH)$ and P_rH is an invariant subspace of the semigroup T(t). On P_rH, $T_r(t) = P_rT(t)P_r$ is a C_0-semigroup with generator $A_r = P_rAP_r$. Obviously, A_r is a (D)-operator in P_rH because A is a (D)-operator in H. Therefore, for any $Y_r \in P_rH$,

$$T_r(t)Y_r(t) = \sum_{n=n_1+1}^{N_0}\sum_{j=1}^{2}e^{\mu_{nj}t}\sum_{k=0}^{n(j)}\alpha_{knj}\phi_{knj} + \sum_{j=1}^{2}\sum_{n=N_0+1}^{\infty}\alpha_{nj}e^{\mu_{nj}t}\phi_{nj}, \tag{4.9}$$

where $\{\phi_{knj}\}$ are the generalized unit eigenvectors of A_r, $N_0 > 0$ is a constant, and

$$\begin{aligned}\|T_r(t)Y_r\| &\leq M_0\left(\sum_{j=1}^{2}\sum_{n=n_1+1}^{N_0} e^{2\mathrm{Re}\mu_{nj}t}\sum_{k=0}^{n(j)}|\alpha_{knj}|^2 + \sum_{j=1}^{2}\sum_{n=N_0+1}^{\infty}|\alpha_{nj}|^2e^{2Re\mu_{nj}t}\right)^{\frac{1}{2}}\\ &\leq M_0M_r(\varepsilon)e^{-(\delta_1-\varepsilon)t}\left(\sum_{n=n_1+1}^{N_0}\sum_{j=1}^{2}\sum_{k=0}^{n(j)}|\alpha_{knj}|^2 + \sum_{j=1}^{2}\sum_{k=n_1+1}^{\infty}|\alpha_{nj}|^2\right)^{\frac{1}{2}}\\ &\leq M_0M_r(\varepsilon)M_1e^{-\delta_1t}\|Y_r\|, \end{aligned} \tag{4.10}$$

where M_0 and M_1 are constants relative to the unconditional bases [3]. The proof of (4.8) is complete. □

5. Finite-dimensional state feedback controller. In this section, we design a finite-dimensional state feedback controller such that system (1.1) is stable.

THEOREM 5.1. *For a given ε : $0 < \varepsilon < \delta_1$, let there be no eigenvalue of A on the line $\{\lambda|\ \mathrm{Re}\lambda = -(\delta_1 - \varepsilon)\}$. Then, there exists a bounded linear operator $\hat{\mathcal{C}}_\epsilon$ such that $A + \hat{\mathcal{C}}_\varepsilon$ generates an exponentially stable C_0-semigroup $T_\varepsilon(t)$; that is, there exists a constant $M_\varepsilon \geq 1$ such that*

$$\|T_\epsilon(t)\| \leq M_\epsilon e^{-(\delta_1-\epsilon)t}, \quad t \geq 0. \tag{5.1}$$

Proof. From Theorem 4.2, we know that $A = A_{n_1} + A_r$, where A_{n_1} is invariant on $H_1 = P_{n_1}H$, and A_r is invariant on $H_2 = P_r H$. Because H_1 is a finite-dimensional subspace spanned by the generalized eigenfunctions of A_{n_1}, A_{n_1} is equivalent to a finite matrix A_1. By the theory of lumped parameter systems, we can find a matrix C_1 such that (A_1, C_1) is controllable. Then there exists a matrix K_1 such that all eigenvalues of $A_1 + C_1K_1$ are assigned to the half plane $\{\lambda \mid \mathrm{Re}\lambda < -(\delta_1 - \varepsilon)\}$ and $\sigma(A_1 + C_1K_1) \cap \sigma(A_r) = \emptyset$. Let C_ϵ be the operator on H_1 corresponding to C_1K_1. Let $\hat{\mathcal{C}}_\varepsilon = C_\varepsilon P_{n_0}$. By the finite state feedback $\hat{\mathcal{C}}_\varepsilon$, we get a new system:

$$\begin{cases} \dfrac{d}{dt}Y(t) = (A + \hat{\mathcal{C}}_\varepsilon)Y(t), \\ Y(0) = Y_0 \in H\,. \end{cases} \tag{5.2}$$

By the definition of $\hat{\mathcal{C}}_\epsilon$, we see that $A + \hat{\mathcal{C}}_\epsilon$ generates a C_0 semigroup $T_\epsilon(t)$, $t \geq 0$ on H, and there exists a constant $M_\epsilon \geq 1$ such that $\|T_\varepsilon(t)\| \leq M_\varepsilon e^{-(\delta_1-\varepsilon)t}$, $t \geq 0$. The proof is complete. □

Because system (5.2) is not equivalent to (1.2), the stability result of system (5.2) does not directly imply that of system (1.1). So we will find a control for system (1.1) such that the closed loop system of (1.1) is stable.

For convenience, we write $\hat{\mathcal{C}}_\varepsilon$ in the form of a matrix:

$$\hat{\mathcal{C}}_\varepsilon = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}. \tag{5.3}$$

Lemma 5.2. *Let u(x,t), $\frac{\partial u(x,t)}{\partial t}$, $\frac{\partial u(x,t)}{\partial x}$, and $\frac{\partial^2 u(x,t)}{\partial x^2}$ be continuous. Then the system*

$$\begin{cases} y_1 = u + \displaystyle\int_0^t (1,0)\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}\begin{pmatrix} y_1(s) \\ y_2(s) \end{pmatrix} ds, \\ y_2 = \dfrac{du}{dt} \end{cases} \tag{5.4}$$

has a unique solution.

By the theory of Volterra integral equations, the proof of Lemma 5.2 *is simple.*

For system (1.1)*, we take the following control:*

$$f(t) = (0,1)\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} - A_0 \int_0^t (1,0)\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}\begin{pmatrix} y_1(s) \\ y_2(s) \end{pmatrix} ds, \tag{5.5}$$

where y_1 and y_2 are defined by (5.4)*. Then system* (1.1) *controlled by $f(t)$ can be written as*

$$\begin{cases} \dfrac{d^2u(t)}{dt^2} + B\dfrac{d}{dt}u(t) + A_0u(t) = f(t)\,, \quad t > 0, \\ u(0) = u_0\,, \quad \dfrac{d}{dt}u(0) = u_1\,, \quad u_0, u_1 \in L^2[0,l], \end{cases} \tag{5.6}$$

where $B = b(x)$. By the transformation

$$\begin{cases} y_1 = u + \int_0^t (1,0)\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}\begin{pmatrix} y_1(s) \\ y_2(s) \end{pmatrix} ds, \\ y_2 = \dfrac{du}{dt}, \end{cases} \tag{5.7}$$

system (5.6) *becomes*

$$\left\{\begin{aligned} &\frac{d}{dt}Y(t) = (A + \mathcal{C}_\epsilon)Y(t)\,, \quad t > 0 \\ &Y(0) = \begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix} = \begin{pmatrix} u_0 \\ u_1 \end{pmatrix}, \end{aligned}\right. \tag{5.8}$$

where $Y(t) = (y_1(y), y_2(t))^T$. *By Theorem* 5.1, *system* (5.7) *generates an exponentially stable C_0-semigroup.*

THEOREM 5.3. *The solution $u(x,t)$ of closed-loop system* (5.6) *converges exponentially fast to*

$$-\int_0^\infty (1,0)\hat{\mathcal{C}}_\epsilon \begin{pmatrix} y_1(s) \\ y_2(s) \end{pmatrix} ds,$$

where $y = (y_1, y_2)^T$, the solution of (5.8).

Proof. Because $u(x,t_2) - u(x,t_1) = \int_{t_1}^{t_2} \frac{\partial u(x,t)}{\partial t} dt = \int_{t_1}^{t_2} y_2(t)dt$ and

$$\|u(x,t_2) - u(x,t_1)\| \le \int_{t_1}^{t_2} \|y_2(t)\|dt \le \int_{t_1}^{t_2} Me^{-\delta_0 t}dt,$$

we get

$$\lim_{t\to\infty} u(x,t) = u(x,\infty) \in H^2(0,l).$$

Because

$$y_1(t) = u(x,t) + \int_0^t (1,0)\hat{\mathcal{C}}_\epsilon \begin{pmatrix} y_1(s) \\ y_2(s) \end{pmatrix} ds$$

and

$$\|u(x,t) + \int_0^t (1,0)\hat{\mathcal{C}} \begin{pmatrix} y_1(s) \\ y_2(s) \end{pmatrix} ds\| = \|y_1(t)\| \le \|Y(0)\|Me^{-\delta_0 t} \to 0,$$

we have

$$\lim_{t\to\infty} \int_0^t (1,0)\hat{\mathcal{C}}_\epsilon \begin{pmatrix} y_1(s) \\ y_2(s) \end{pmatrix} ds = \int_0^\infty (1,0)\hat{\mathcal{C}}_\epsilon \begin{pmatrix} y_1(s) \\ y_2(s) \end{pmatrix} ds. \qquad \square$$

Remark. For system (1.1), it is still not clear how many eigenvalues have nonnegative real parts.

REFERENCES

[1] G. CHEN, S. A. FULLING, F. J. NARCOWCH, AND C. QI, *An asymptotic average decay rate for the wave equation with variable coefficient viscous damping*, SIAM. J. Appl. Math., 50 (1990), pp. 1341–1347.

[2] Y. H. LUO, *The root property on a certain discrete spectrum operators*, Acta. Math. Sinica, 32 (1991), pp. 556–563.

[3] I. SINGER, *Bases in Banach Spaces*, Springer-Verlag, Berlin, Heidelberg, New York, 1970.

[4] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1983.

[5] G. CHEN, S. A. FULLING, F. J. NARCOWCH, AND S. SUN, *Exponential decay energy of evolution equations with locally distributed dampings*, SIAM. J. Appl. Math., 51 (1991), pp. 266–301.

[6] G. Chen, M. C. Delfour, A. M. Krall, and G. Payre, *Modeling, stabilization and control of serially connected beams*, SIAM J. Control Optim., 25 (1987), pp. 526–546.

[7] S. J. Li, *The analytic property of the C_0-semigroup corresponding to the slender flying vehicle system with structural damp*, Act. Shanxi University. Scis., 4 (1987), pp. 6–11.

[8] S. J. Li, *Time dependent elastic system*, Acta Math. Sci. (Chinese), 15(supp) (1995), pp. 39–59.

[9] J. Y. Yu, S. J. Li, and G. T. Zhu, *The stability of the time variable elastic system*, Sci. China, Ser. E, 39 (1996), pp. 92–102.

[10] C. Z. Xu and G. Sallet, *On spectrum and Riesz basis assignment of infinite-dimensional linear systems by bounded linear feedbacks*, SIAM J. Control Optim., 34 (1996), pp. 521–541.

[11] M. H. Stone, *A comparison of the series of Fourier and Birkhoff*, Trans. Amer. Math. Soc., 28 (1926), pp. 695–767.

[12] M. J. Balas, *Feedback control of dissipative hyperbolic distributed parameter systems with dimensional controllers*, J. Math. Anal. Appl., 98 (1984), pp. 1–24.

[13] R. Rebarber, *Exponential stability of coupled beams with dissipative joints: A frequency domain approach*, SIAM J. Control Optim., 33 (1995), pp. 1–28.

[14] G. Chen, S. G. Krants, D. L. Russell, C. E. Wayne, H. H. West, and M. P. Coleman, *Analysis, designs and behavior of dissipative joints for coupled beams*, SIAM J. Appl. Math., 49 (1989), pp. 1665–1693.

[15] G. Chen, M. P. Coleman, and H. H. West, *Pointwise stabilization in the middle of the span for second order systems, nonuniform and uniform exponential decay of solutions*, SIAM J. Appl. Math., 47 (1987), pp. 751–780.

[16] F. Conrad, *Stabilization of beams by pointwise feedback control*, SIAM J. Control Optim., 28 (1990), pp. 423–437.

[17] R. Curtain and A. Prichard, *Infinite Dimensional System Theory*, New York, Springer-Verlag, 1978.

[18] M. C. Dekfour, J. Lagnese, and M. P. Polis, *Stabilization of hyperbolic system using concentrated sensors and actuators*, IEEE Trans Automat. Control, AC-31 (1986), pp. 1091–1096.

[19] D. X. Feng and G. T. Zhu, *Spectral properties of principal operators for a class of elastic vibration problem*, Kexue Tongbao, 26 (1981), pp. 1473–1475.

[20] S. Hanssen and E. Zuazua, *Exact controllability and stabilization of a vibrating string with an interior point mass*, SIAM J. Control. Optim., 33 (1995), pp. 1357–1391.

[21] S. Cox and E. Zuazua, *The rate at which energy decays in a damped string*, Comm. Partial Differential Equations, 19 (1994), pp. 213–243.

[22] P. Freitas, *On some eigenvalue problems related to the wave equation with indefinite damping*, J. Differential Equations, 127 (1996), pp. 320–335.

[23] P. Freitas and E. Zuazua, *Stability results for the wave equation with indefinite damping*, J. Differential Equations, 132 (1996), pp. 338–352.

SIAM J. Control Optim.
Vol. 37, No. 6, pp. 1780–1807

ADMISSION CONTROL FOR COMBINED GUARANTEED PERFORMANCE AND BEST EFFORT COMMUNICATIONS SYSTEMS UNDER HEAVY TRAFFIC*

EITAN ALTMAN† AND HAROLD J. KUSHNER‡

Abstract. Communications systems often have many types of users. Since the users share the same resource, there is a conflict in their needs. This conflict leads to the imposition of controls on admission or elsewhere. In this paper, there are two types of customers, GP (Guaranteed Performance) and BE (Best Effort). We consider an admission control of GP customer which has two roles. First, to guarantee the performance of the existing GP customers, and second, to regulate the congestion for the BE users. The optimal control problem for the actual physical system is difficult. A heavy traffic approximation is used, with optimal or nearly optimal controls. It is shown that the optimal values for the physical system converge to that for the limit system and that good controls for the limit system are also good for the physical system. This is done for both the discounted and average cost per unit time cost criteria. Additionally, asymptotically, the pathwise average (not mean) costs for the physical system are nearly minimal when good nearly optimal controls for the limit system are used. Numerical data show that the heavy traffic optimal control approach can lead to substantial reductions in waiting time for BE with only quite moderate rejections of GP, under heavy traffic. It also shows that the controls are often linear in the state variables. The approach has many advantages. It is robust, simplifies the analysis (both analytical and numerical), and allows a more convenient study of the parametric dependencies. Even if optimal control is not wanted, the approach is very convenient for a systematic exploration of the possible tradeoffs among the various cost components. This is done by numerically solving a series of problems with different weights on the costs. We can then get the best tradeoffs and the control policies which give them.

Key words. admission control, control of communications systems, control of queueing networks, heavy traffic limits, ergodic control, singular control, weak convergence

AMS subject classifications. 90B22, 60K25, 60K30, 60F17, 93E20, 93E25

PII. S0363012998333517

1. Introduction.

Broadband-ISDN (integrated services data network) high speed networks allow the possibility of integrating different services into one single telecommunications network. In particular, they handle applications that require guaranteed quality of service (QoS), such as bounded delays, bounded cell loss rates, and a guarantee on the throughput. Such a service is called Guaranteed Performance (GP). On the other hand, they also support more flexible applications, such as data transfer, that are less sensitive to instantaneous variations in available bandwidth, delays, jitter, etc., and which do not require guarantees on throughputs or delays. We call the latter "Best Effort" (BE) traffic. In the context of Asynchronous Transfer Mode (ATM), this corresponds to the Available Bit Rate (ABR) service category [1], which can adapt to the bandwidth unused by the GP service classes. In the context of the Internet, the BE traffic are the TCP/IP connections, which use a congestion avoidance mechanism [17] so as to adapt to the available bandwidth, in contrast with real-time applications that use UDP (UDP is a protocol which does not adapt its transmission

*Received by the editors February 3, 1998; accepted for publication (in revised form) January 10, 1999; published electronically October 4, 1999. The work was supported by contract DAAH04-96-1-0075 from the Army Research Office and NSF grant ECS9703895.

http://www.siam.org/journals/sicon/37-6/33351.html

†INRIA, 2004 Route des Lucioles, B.P. 93, 06902 Sophia-Antipolis Cedex, France (eitan.altman@sophia.inria.fr).

‡Division of Applied Mathematics, Brown University, Providence, RI 02912 (hjk@dam.brown.edu).

rate to the current congestion state of the system) such as the phone over the Internet [5]. In particular, we suppose that the BE users share the remaining bandwidth, as for example, in the standard Internet. (Note that, unlike ATM, the Internet does not provide performance guarantees for real-time applications. However, in practice, the TCP/IP connections do adapt their transmission rate to the available bandwidth left over after the UDP sessions; see [3].)

Since GP and BE share the same resources of the network, there is a conflict in needs, and the question of admission control arises. The purpose of this paper is to analyze the impact that Connection Admission Control (CAC) performed on GP traffic has on the BE traffic, in order to properly allocate resources in a dynamic manner as well as guarantee QoS.

In ATM networks, a large part of the architecture has been standardized [1]. Implementation of the CAC, however, is not standardized and is left to the network manager. This has motivated much research and a large literature has emerged. Several approaches have been used in designing CACs. The first approach is based on characterizing the input traffic of a source by some simple parameters, such as the effective bandwidth; basically it estimates the bandwidth that should be given to a source to guarantee that the cell loss rate does not exceed (in some asymptotic sense) some small number. Examples of this approach are [10, 15, 13, 18, 19, 31, 32, 37]. A diffusion-based statistical CAC was proposed in [12].

A recent approach to admission control estimates (on-line) the consequence of possible admission, using real-time measures of the network activity [14, 35, 36]. An optimal-control approach for call admission, using a Markov Decision Processes approach, has been used in a number of papers [11, 16, 34]. All the cited papers focus on the impact of admission control on GP traffic. To the best of our knowledge, our paper is the first to consider the optimal design of the CAC for the GP sessions in terms of the performance of both the BE and the GP sessions.

The CAC was not intended to be performed on BE traffic, since BE traffic is sufficiently flexible so as not to deteriorate QoS that is required by GP connections. In particular, the ATM forum has decided [1] that ABR sessions will not be subject to CAC (unless they require from the network a guarantee on minimum cell rate). We shall therefore assume that BE sessions are always admitted into the network, unless some very large limit is reached. This limit might represent the number of sessions that can be handled simultaneously by the switches. This limit will not depend on the available bandwidth.

Quite often, CACs are designed taking into account only the performance requirements of GP sessions. However, as already noted in [2, 4], under appropriate operating conditions in the network (in particular, the heavy-traffic conditions, which correspond to an efficient utilization of the resources) it is possible to improve substantially the performance (delays and throughputs) of the BE sessions by modifying *slightly* the CAC for the GP sessions.

References [2, 4] contain analysis of the performance of GP and BE sessions for given CACs (performed on GP sessions) that take into account also the performance of BE sessions. In this paper, we go one further step and provide a design based on an optimization approach. We consider the *optimal* admission control arising when the number of sessions and the available bandwidth are taken to be large and scaled in an appropriate way so as to perform in the desirable heavy traffic region.

The approach taken is that of optimal control in the heavy traffic regime. In this regime, the system has little spare bandwidth over what is needed to handle the "average load." It is shown that the physical system can be well approximated by a controlled (reflected) diffusion process, and that good controls for this diffusion process are good for the physical system under heavy traffic. The resulting controls are easily implementable (and are often simply linear in the state variables) and can give a significant improvement of the performance of the BE sessions by only a small rejection of GP sessions. The work is a continuation of the past work on the control of communications and queueing systems under heavy traffic conditions [29, 27, 26, 22, 30, 25, 28, 23].

The extensive numerical results in [25] illustrate the power of such an approach. Excellent controls, system performance, and very useful information which would not otherwise be available were obtained. The paper [30], concerned with trunk line problems, also shows the power of the heavy traffic approach for modeling, simplifying and controlling very complex systems.

The control problem for the actual physical model is quite hard, if not impossible to solve. The system is not always Markovian. Even in the Markovian case, the number of states can be extremely large. Additionally, even for an uncontrolled Markovian system, the computation of the steady state overflows (losses) and delays is difficult. The heavy traffic approach provides much analytical and computational simplification. The optimally controlled costs for the physical problem converge to the optimally controlled cost for the limit problem, and a nearly optimal control for the limit problem is also nearly optimal for the physical problem if the traffic is "heavy."

In applications, the traffic is not always heavy. But this regime is one of the crucial ones for design, analogous to the case of the trunk line problem for long distance teletraffic. The analysis illustrates efficient use of resources. In any large system, there are many alternative uses of the resources and continuous tradeoffs among them. Here we can see, for example, the effect of marginal reallocations of bandwidth within a control context. Of course, serious control problems do exist even outside of the heavy traffic regime; for example, even if the system is heavily overbuilt from the heavy traffic perspective, the structure of the burstiness in arrivals can lead to important control problems. In this regard, it is worth noting that, in the numerical study in [25] of controlled multiplexers in the heavy traffic regime, the buffer is empty or nearly empty most of the time, and it is the burstiness of the input which yields the losses and the control problem.

There are many other advantages to the combined heavy traffic and optimal control approach. Under broad conditions, it yields the appropriate dimensioning of the system and shows that good performance can often be achieved with only modestly extra bandwidth (over average requirements). The "limit" variables can be interpreted as "aggregated" states. The analysis takes advantage of the laws of large numbers and central limit theorems that come into play as the size of the system grows. The "limit" or aggregate equations can be used to compute nearly optimal controls for the physical system, and to get good estimates of various measures of performance. The relative simplicity of the form of the heavy traffic limit process facilitates understanding the parametric dependencies. As in [25], we can obtain both qualitative and quantitative information which is often very hard to get. To do so otherwise, one would need to study many particular cases of systems with different parameters and sizes whose relationships would remain obscure. The reference [33] contains examples of other types of applications.

The method allows convenient numerical approximations (see [24, 25]) whose complexity is much less than that of even the physical Markov systems. In fact, the basic Markov chain approximation method of [24] has been coded for problems of this type and is publicly available.[1] The format is robust: bursty data, priorities, time dependence, dependence of arrivals on the system state, and other useful extensions can be readily handled.

An important question concerns the tradeoff between gains in QoS for BE versus losses for GP. The heavy traffic formulation allows a systematic exploration of this. One solves the optimization problem with various weights for the associated losses and costs and computes the values of the individual components of the cost under the optimal controls. This yields the set of possible tradeoffs under good operating policies. One can then choose the control which gets the best performance for one component with a constraint on the others. Usually, optimization for its own sake is not the main interest. But the use of optimizing or optimal control methods for exploration of possibilities is of greater interest and provides a powerful design tool. The value of this approach was amply demonstrated in [25], and the large gains due to the use of feedback control were demonstrated. This is equally valid for the present problem. Indeed, numerical data shows that substantial reductions (say, 30% or even much more) in BE delays can be obtained with only very modest levels of rejection of GP customers. The percent rejected depends on the system parameters and the arrival rates, and it goes to zero as the system grows, for any fixed percentage gain in delay.

The "state space" collapse phenomenon illustrated in section 5, when there are many GP and BE subclasses, allows one to explore the effects on performance and control of complex customer requirements. One can do the numerics for aggregations of several classes as well as for the original problem to get insights into robustness of the model and performance under varying conditions. This would be very hard under any current alternative approach. Indeed, the limit allocations to the different classes are consistent with the (ad hoc) weighted fair-shares for different types of BE customers which were defined by the ATM forum for sharing bandwidth (see [1, section I.3]).

The model in section 4 can be used to study the effects of finite bandwidth constraints on individual users. The multicast problem of section 6 shows how seemingly complex forms of the problem can be put into the context of a general theory.

The structure of the paper is as follows. The basic model is presented in section 2, and the input-output equation is written in a way that is convenient for the heavy traffic analysis. In the basic model, the BE customers share the available bandwidth equally, whatever it is. The general methods used for the analysis of this case also apply with little modification to the subsequent cases. Section 3 deals with the weak convergence of the heavy traffic approximations and the discounted cost function as the traffic intensity approaches unity. Many of the ideas have been used in the references in various ways, despite the difference in the problems considered. Because of this and to save space, we provide detailed outlines with references where possible. There is a new problem with "tightness" of the set of the (singular) controls, but we show that they can be approximated such that they are tight.

Section 4 considers the case where there is an upper limit to the bandwidth that any one of the BE customers can use. This changes the dynamics and the scaling, but the previous analysis can be carried over. In order to illustrate the power and

[1] See the home page of the Lefschetz Center for Dynamical Systems, Brown University Dept. of Applied Mathematics home page http://www.dam.brown.edu/lcds.html. Select the software link to get the documentation and codes.

versatility of the approach, in section 5 we consider extensions where there are several classes of BE customers, which might be allocated bandwidth in class dependent ways. There is a surprising degeneracy, which can be exploited to simplify the analysis and subsequent numerical work.

Section 6 extends the basic model to a multicast case. Here there are two channels (any number could be used), each with its own class of GP customers and its own control. But the BE customers must be transmitted simultaneously on each of the channels. Many variations of the format are possible. The ergodic cost problem is dealt with in section 7. In section 7 we impose a constraint on the rate at which GP arrivals can be rejected, so that previous results on the ergodic cost problem can be exploited. The form of the heavy traffic limit equations implies that this constraint is not very restrictive and this is borne out by numerical data. But in section 8 we show that it is not a restriction.

The optimal mean cost per unit time for the physical system converges to the optimal "ergodic" value for the limit system as the size of the system and time go to infinity in any way at all. Furthermore, the limit of the pathwise (not mean) average costs per unit time cannot be better than the optimal ergodic value for the limit system. The pathwise average costs for the physical system can be made to be arbitrarily close to this ideal limit by using a nice nearly optimal (for the limit system) control on the physical system, under heavy traffic. The pathwise results are important, since in any single application, we have just one sample path, and the mean values are less important than the pathwise values. Furthermore, a "nice" nearly optimal control for the limit system provides nearly optimal values for the physical system, under heavy traffic, in both a mean and pathwise sense.

Some of the development is similar to that in recent works on control under heavy traffic (e.g., [29, 22, 25, 23]), although the exact form of the adaptation is not entirely obvious for the present problem. This is particularly true of the derivation of the basic setup in the first part of section 3. Because of this, we have referred to the literature whenever possible. There remain many new methodological features apart from the novelty of an important application in a broad context: The approximation of the singular controls, both for the discounted and the ergodic problem, the degeneration (or state space collapse) for the multi BE/GP class problem, the adaptation of the formulation to multicast, and other extensions, with state dependent dynamics.

2. Problem 1: The basic setup. In this section, we set up the notation and evolution equations for the *basic problem*. The development for the more complex problem classes is similar. There are two classes of users, which we refer to as class 0 and class 1. Class 0 refers to the BE traffic and class 1 the GP traffic. As usual in the analysis of systems under conditions of heavy traffic, the mean service capacity is slightly greater than the mean demand. The system is parameterized by a parameter N. Both the system capacity, excess capacity (over what is required for the average demands), and demand grow as $N \to \infty$, with the *relative* excess capacity going to zero. This will be formalized below. A consequence of the heavy traffic analysis is that these relationships represent good design in that very good performance can be obtained. A well designed heavy traffic operating regime implies that the system is not overbuilt and can handle the demands placed on it.

In this and in the next section, the bandwidth is normalized so that each member of class 1 requires one unit of bandwidth. The members of class 0 share whatever bandwidth is left. An applications paradigm is that each arrival is the work for a "session," whose work arrives essentially at once, and is buffered on arrival. This models well the more general situation in which the input rate of packets within a session is not less than the transmission rate, so that when we allocate a given bandwidth to a session then this bandwidth is indeed used by it. In section 4 we shall relax this assumption. The channel is time shared, with a guaranteed time going to the class 1 customers, and the time remaining to the class 0 customers. Thus, there is no limitation on the rate at which work can be done on the set of class 0 customers except for the available capacity.

Let $\alpha_l^{i,N}, l = 1, \dots,$ denote the interarrival times for the members of class i, and set $E\alpha_l^{i,N} = \bar{\alpha}^{i,N}$. For a given N, $\{\alpha_l^{i,N}\}_{l=1,\dots,i=0,1}$ are assumed to be mutually independent. To conform with standard usage we also define the normalized mean "rates" $\lambda^{i,N} = 1/N\bar{\alpha}^{i,N}$. The service times for the members of class 1 are exponentially distributed with constant rate $\mu^{1,N}$. Class 1 can be controlled in that any requested admission can be denied by the controller. The system is a "loss" system in that any customer denied admission disappears from the system. For analytic convenience we suppose further that there is a constant $B_+^0 > 0$ such that no class 0 customers are admitted to the system if the current number of class 0 customers in the system is greater than $\sqrt{N}B_+^0$. It follows from the heavy traffic limit theorems (and from the numerical data) that this is inconsequential for large enough B_+^0. In fact, the condition is useful only to simplify some of the analytic details for the ergodic cost criterion, and B_+^0 can be set equal to infinity for the discounted cost function. But we include the finite upper bound here to unify the development. Let $F^{0,N}(t)$ denote $1/\sqrt{N}$ times the number of class 0 customers not admitted by time t.

The time requirements for the members of class 0 are also exponentially distributed. But the rate at which a customer of class 0 departs from the system depends on the (random) resources available to it while it is in the system. The "conditional mean instantaneous departure rate" for a class 0 customer at time t is defined to be $\mu^{0,N}$ times the bandwidth available to that customer at that time; i.e., if at time t, $B(t)$ is the total bandwidth unused by members of class 1, and there are $A(t)$ members of class 0 in the system, then the probability (conditioned on the data up to that time) of a single departure of a member of class 0 in the time interval $[t, t+\delta)$ is $\mu^{0,N}B(t)\delta/A(t) + o(\delta)$, and the probability of more than one departure in that interval is $o(\delta)$. The set of interarrival times and the service times for class 1 are assumed to be mutually independent. This independence of the interarrival times is a good description of reality, since we are working at the session level, where this property has often been observed.[2] Long range dependence in the arrival process is not relevant as it might be at the packet level.

We assume that there is b (parameterizing the "excess capacity," which might be negative) such that the channel capacity is

$$C_N = N\left[\frac{\lambda^{0,N}}{\mu^{0,N}} + \frac{\lambda^{1,N}}{\mu^{1,N}}\right] + b\sqrt{N}. \tag{2.1}$$

Suppose that $\lambda^{i,N} \to \lambda_i$ and $\mu^{i,N} \to \mu_i$, all positive. Thus, the mean arrival rates and the channel capacity are all $O(N)$, and the channel capacity is $O(\sqrt{N})$ greater than the mean requirements. For appropriate b, this will be seen to be sufficient for good

[2] See, e.g., Liu (INRIA, Sophia Antipolis, France), *Measurements over the Web*, private communication, to be submitted.

behavior. We also make the innocuous assumption that

$$\left\{ \left| \alpha_l^{i,N} / \bar{\alpha}^{i,N} \right|^2 ; l, i, N \right\} \text{ is uniformly integrable,} \tag{2.2}$$

and that there are σ_i^2 such that

$$E \left[1 - \frac{\alpha_l^{i,N}}{\bar{\alpha}^{i,N}} \right]^2 \to \sigma_i^2. \tag{2.3}$$

In common cases, the arrival processes are assumed to be Poisson. Then all moments of the terms in (2.2) are uniformly (in i, l, n) bounded. If the interarrival intervals for class i are constant, then $\sigma_i^2 = 0$. If the arrival stream for class i is Poisson, then $\sigma_i^2 = 1$. Define the scaled mean arrival process

$$S^{i,N}(t) = \frac{1}{N} \left[\# \text{ of class } i \text{ arrivals by time } t \right]$$

and the normalized and centered sum

$$W^{i,N}(t) = \frac{1}{\sqrt{N}} \sum_{l=1}^{[Nt]} \left[1 - \frac{\alpha_l^{i,N}}{\bar{\alpha}^{i,N}} \right],$$

where $[Nt]$ denotes the integer part.

As usual in heavy traffic scaling, all of the basic system variables are scaled by $1/\sqrt{N}$. Define $X^{0,N}(t)$ to be $1/\sqrt{N}$ times the number of members of class 0 in the system at time t, and set

$$X^{1,N}(t) = \frac{1}{\sqrt{N}} \left[\# \text{ of class 1 in system at time } t - N \frac{\lambda^{1,N}}{\mu^{1,N}} \right].$$

Thus $X^{1,N}(t)$ is the scaled number of class 1 customers centered about the mean, if there are no rejections and an infinite channel capacity. Define $A^{i,N}(t)$ (resp., $D^{i,N}(t)$) to be $1/\sqrt{N}$ times the number of arrivals (resp., departures) of class i by time t, and $F^{1,N}(t)$ denotes $1/\sqrt{N}$ times the number of arrivals of class 1 up to time t which the controller did not admit. Thus, $A^{i,N}(t) = \sqrt{N} S^{i,N}(t)$. The (nondecreasing) control process $F^{1,N}(\cdot)$ is assumed to be *admissible* in that it is (ω, t) measurable and its value at time t depends only on the data which is available up to time t.

The system balance equations are

$$X^{0,N}(t) = X^{0,N}(0) + A^{0,N}(t) - D^{0,N}(t) - F^{0,N}(t), \tag{2.4a}$$

$$X^{1,N}(t) = X^{1,N}(0) + A^{1,N}(t) - D^{1,N}(t) - F^{1,N}(t) - U^{1,N}(t), \tag{2.4b}$$

where $U^{1,N}(t)$ is $1/\sqrt{N}$ times the number of class 1 customers that could not be admitted due the *entire* channel being occupied by class 1 customers. This last term will disappear in the limit. We will suppose that the set

$$\left\{ X^{1,N}(0), N \right\} \text{ is tight (bounded in probability).} \tag{2.5}$$

If this set is not tight, then there will be a long delay before the heavy traffic regime is entered. The set $\{X^{0,N}(0)\}$ is always tight, since $B_+^0 < \infty$. It is also supposed that the initial condition is independent of the subsequent arrival times and service times for class 1.

In order to simplify the convergence proofs, one needs to put the input and output processes into a more convenient form.

Representation of the input processes. Following a common practice in weak convergence analysis, we decompose the arrival process as

$$\begin{aligned} A^{i,N}(t) &= \frac{1}{\sqrt{N}} \sum_{l=1}^{NS^{i,N}(t)} 1 \\ &= \frac{1}{\sqrt{N}} \sum_{l=1}^{NS^{i,N}(t)} \left[1 - \frac{\alpha_l^{i,N}}{\bar{\alpha}^{i,N}}\right] + \frac{1}{\sqrt{N}} \sum_{l=1}^{NS^{i,N}(t)} \frac{\alpha_l^{i,N}}{\bar{\alpha}^{i,N}}. \end{aligned} \tag{2.6a}$$

Note that

$$\sum_{l=1}^{NS^{i,N}(t)} \alpha_l^{i,N}$$

equals t minus the time since the last arrival before or at t. Thus, by (2.6a),

$$A^{i,N}(t) = W^{i,N}(S^{i,N}(t)) + t\lambda^{i,N}\sqrt{N} - \frac{\rho^{i,N}(t)}{\sqrt{N}}, \tag{2.6b}$$

where $\rho^{i,N}(t)$ is the time since the last arrival before or at t, divided by the mean interarrival interval. The sequence of processes $\rho^{i,N}(\cdot)/\sqrt{N}$ converges weakly to the "zero" process. It does not affect any of the subsequent calculations and for the sake of notational simplicity, it will be *omitted* in all of the subsequent system equations after (2.11).

Representation of the output processes. Because of the exponential distribution of the service time for the class 1 customers, $D^{1,N}(\cdot)$ can be decomposed into the sum of the integral of the *instantaneous conditional mean rate* at which $D^{1,N}(\cdot)$ increases, and a martingale process $\tilde{D}^{i,N}(\cdot)$. To do this, first note that the conditional mean instantaneous rate of increase of $D^{1,N}(\cdot)$ at t is $\mu^{1,N}/\sqrt{N}$ times the number of class 1 customers in the system at t.

We have the decomposition

$$D^{1,N}(t) = \mu^{1,N} \int_0^t \left[X^{1,N}(s) + \sqrt{N}\frac{\lambda^{1,N}}{\mu^{1,N}}\right] ds + \tilde{D}^{1,N}(t). \tag{2.7}$$

The Doob–Meyer increasing process associated with the martingale is just $1/\sqrt{N}$ times the integral in (2.7); namely,

$$\langle \tilde{D}^{1,N} \rangle(t) = \frac{\mu^{1,N}}{\sqrt{N}} \int_0^t \left[X^{1,N}(s) + \sqrt{N}\frac{\lambda^{1,N}}{\mu^{1,N}}\right] ds. \tag{2.8}$$

The factor $1/\sqrt{N}$ appears due to the definition of $D^{i,N}(t)$ as $1/\sqrt{N}$ times the number of departures by time t. Similar decompositions were used in [22, 27].

Let $I^{0,N}(t)$ denote the indicator function of the event that there are class 0 customers in the system at t. The departure process for class 0 is similarly decomposed into the integral of the conditional mean instantaneous rate at which $D^{0,N}(\cdot)$ increases and a martingale $\tilde{D}^{0,N}(\cdot)$. In preparation for this, first note that the available bandwidth per class 0 customer at time t is

$$\frac{C_N - \left[\sqrt{N}X^{1,N}(t) + N\lambda^{1,N}/\mu^{1,N}\right]}{\sqrt{N}X^{0,N}(t)} I^{0,N}(t),$$

which equals

$$\frac{N\lambda^{0,N}/\mu^{0,N} + b\sqrt{N} - \sqrt{N}X^{1,N}(t)}{\sqrt{N}X^{0,N}(t)} I^{0,N}(t).$$

Thus the conditional mean instantaneous rate at which $D^{0,N}(\cdot)$ increases at t is

$$\left[\sqrt{N}\frac{\lambda^{0,N}}{\mu^{0,N}} + b - X^{1,N}(t)\right] I^{0,N}(t).$$

Hence,

$$D^{0,N}(t) = \mu^{0,N} \int_0^t \left[\sqrt{N}\frac{\lambda^{0,N}}{\mu^{0,N}} + b - X^{1,N}(s)\right] I^{0,N}(t)ds + \tilde{D}^{0,N}(t), \tag{2.9a}$$

which we rewrite as

$$D^{0,N}(t) = \mu^{0,N} \int_0^t \left[\sqrt{N}\frac{\lambda^{0,N}}{\mu^{0,N}} + b - X^{1,N}(s)\right] ds + \tilde{D}^{0,N}(t) - Y^{0,N}(t). \tag{2.9b}$$

The term $Y^{0,N}(\cdot)$ is a reflection term; it compensates for the difference in the integrals in (2.9a) and (2.9b), and it can increase only when $X^{0,N}(t) = 0$. The Doob–Meyer increasing process associated with $\tilde{D}^{0,N}(\cdot)$ is

$$\langle \tilde{D}^{0,N} \rangle(t) = \frac{\mu^{0,N}}{\sqrt{N}} \int_0^t \left[\sqrt{N}\frac{\lambda^{0,N}}{\mu^{0,N}} + b - X^{1,N}(s)\right] I^{0,N}(s)ds,$$

which is written more conveniently as

$$\mu^{0,N} \int_0^t \left[\frac{\lambda^{0,N}}{\mu^{0,N}} + \frac{b}{\sqrt{N}} - \frac{X^{1,N}(s)}{\sqrt{N}}\right] I^{0,N}(s)ds. \tag{2.10}$$

Now, putting all of the above representations together and canceling the $\pm\sqrt{N}\lambda^{i,N}$ terms yields the forms:

$$\begin{aligned} X^{0,N}(t) = X^{0,N}(0) + W^{0,N}(S^{0,N}(t)) - \mu^{0,N} \int_0^t g_0(X^N(s))ds \\ - \tilde{D}^{0,N}(t) - F^{0,N}(t) + Y^{0,N}(t) - \frac{\rho^{0,N}(t)}{\sqrt{N}}, \end{aligned} \tag{2.11a}$$

$$\begin{aligned} X^{1,N}(t) = X^{1,N}(0) + W^{1,N}(S^{1,N}(t)) - \mu^{1,N} \int_0^t X^{1,N}(s)ds \\ - \tilde{D}^{1,N}(t) - F^{1,N}(t) - U^{1,N}(t) - \frac{\rho^{1,N}(t)}{\sqrt{N}}, \end{aligned} \tag{2.11b}$$

where we define

$$g_0(x) = b - x^1, \tag{2.12}$$

and the $Y^{0,N}(t)$ term compensates for the fact that there are no departures of class 0 customers at time t if $X^{0,N}(t) = 0$. It assures that $X^{0,N}(t)$ will stay nonnegative.

Comments on weak convergence. The path space for all of the random processes is $D^k[0,\infty)$, the space of functions which are right continuous, have left hand limits, and take values in Euclidean k-space for appropriate integers k. The Skorohod topology is used [6, 9]. This is the most common and convenient choice in heavy traffic analysis. The following is a convenient criterion for tightness in this space. It will be used implicitly. Let $\{Y^n(\cdot)\}$ be a sequence of processes with paths in $D[0,\infty)$, with probability one. Let $\mathcal{T}^n(t)$ denote the stopping times with respect to the filtration engendered by $Y^n(\cdot)$ and which are no larger than t. If

$$\lim_{\delta\to 0} \sup_n \sup_{\tau\in\mathcal{T}^n(t)} E\left(1 \wedge |Y^n(\tau+\delta) - Y^n(\tau)|\right) = 0, \tag{2.13}$$

for each t, and

$$\{Y^n(t) : n, t\} \text{ is tight}, \tag{2.14}$$

then $\{Y^n(\cdot)\}$ is tight [9].

3. Discounted cost function and weak convergence.

The convergence theorem implies that only a bandwidth excess of order $O(\sqrt{N})$ is needed for good performance.

Although the set of applications is new, the setup so far is similar to that of many other problems of control of queues under heavy traffic. See, for example, [29, 22, 25, 23]. The main new questions concern the control functions $F^{1,N}(\cdot)$. We will be concerned with two types of cost functions: the discounted and the average cost per unit time (to be called the ergodic cost function). Until section 7, we concentrate on the discounted problem. Much of the analysis carries over to the ergodic case with little change. Also, analogous procedures are used for the convergence proofs for all of the problem formulations in the following sections.

Let $\beta > 0, c_i > 0$, where β can be as small as we wish, and let $k(\cdot)$ be a nonnegative continuous function with $k(0) = 0$. The discounted cost function is defined by

$$C_\beta^N(x, F^{1,N}) = E\int_0^\infty e^{-\beta s} k(X^{0,N}(s))ds + E\int_0^\infty e^{-\beta s} \sum_i c_i dF^{i,N}(s). \tag{3.1}$$

The second term penalizes the rejections. We do not penalize the loss $U^{1,N}(\cdot)$, since it is zero in the limit as $N \to \infty$ no matter what the controls are. The first term can be quite general. If $k(\cdot)$ is linear, then it simply penalizes the waiting time for class 0 customers. More generally, it can be nonlinear. For example, it might be zero for small values of the argument (if the delays at small values are considered to be unimportant), or it might increase superlinearly to discourage long delays. The allowed generality of the cost function is one of the advantages of the approach. Define $V_\beta^N(x) = \inf_{F^{1,N}} C_\beta^N(x, F^{1.N})$, where the inf is over the admissible controls.

The basic structure of the convergence proofs is similar to those in the cited references to controlled queues in heavy traffic, except for the questions of tightness and approximation of the control functions. After stating the convergence theorems, the proofs will be outlined and references given for many of the details.

Define $B^{i,N}(\cdot) = W^{i,N}(S^{i,N}(\cdot)) - \tilde{D}^{i,N}(\cdot)$.

THEOREM 3.1. *Let $\epsilon > 0$ be small but arbitrary, and let $F^{1,N}(\cdot)$ be ϵ-optimal controls. Suppose that $\{F^{1,N}(\cdot), N\}$ is tight.*[3] *Then the set*

$$\{X^{i,N}(\cdot), B^{i,N}(\cdot), F^{i,N}(\cdot), i = 1, 2; Y^{0,N}(\cdot), N\}$$

is tight. The weak sense limit of any weakly convergent subsequence satisfies

$$X^0(t) = X^0(0) - \mu_0 \int_0^t g_0(X(s))ds + B^0(t) + Y^0(t) - F^0(t), \tag{3.2a}$$

$$X^1(t) = X^1(0) - \mu_1 \int_0^t X^1(s)ds + B^1(t) - F^1(t), \tag{3.2b}$$

where the $B^i(\cdot)$ are mutually independent Wiener processes with variance parameters $\lambda_i(1 + \sigma_i^2)$. $0 \le X^0(t) \le B_+^0$, and $Y^0(\cdot)$ is the reflection term at zero. $F^0(\cdot)$ is the reflection term at the upper bound. The other processes are nonanticipative[4] *with respect to the Wiener processes.*

Comments on the proof. Details for similar results are in [22, 23, 25, 29, 27], and we will only outline the sequence of ideas. First, by the renewal theorem, $S^{i,N}(\cdot)$ converges weakly to the deterministic (limit scaled mean arrival rate) process with values $\lambda_i t$. By Donsker's Theorem [6, 9], $W^{i,N}(\cdot)$ converges weakly to mutually independent Wiener processes with variance parameters σ_i^2. Hence the $W^{i,N}(S^{i,N}(\cdot))$ converges to mutually independent Wiener processes $\tilde{W}^i(\cdot)$ with variance parameters $\lambda_i \sigma_i^2$.

The set of martingales $\{\tilde{D}^{i,N}(\cdot); i, N\}$ can readily be proved to be tight via the criterion (2.14), (2.15). This is done by a direct computation using the fact that their associated Doob–Meyer increasing processes (2.8) and (2.10) are bounded by a constant times t. The fact that the scaled discontinuities go to zero as $N \to \infty$ implies that their weak sense limits have continuous paths with probability one.

With the above tightness results available on the "driving terms" $W^{i,N}(S^{i,N}(\cdot))$, $\tilde{D}^{i,N}(\cdot)$, and the tightness assumption on the control terms, the tightness of $\{X^{1,N}(\cdot)$, $U^{1,N}(\cdot)$, $N\}$ can be readily proved if the $X^{1,N}$ in (2.11b) is replaced by some bounded functions. For the general case, the tightness follows by a standard truncation argument [21]. This tightness implies that the sequence $U^{1,N}(\cdot)$ has "zero" weak sense limits, since the "upper boundary" for $X^{1,N}$ gets pushed to infinity as $N \to \infty$. The tightness of $\{X^{1,N}(\cdot)\}$ also implies the tightness of $\{X^{0,N}(\cdot), Y^{0,N}(\cdot), F^{0,N}(\cdot), N\}$. In fact, the tightness of the $\{F^{0,N}(\cdot)\}$ can be shown even without tightness of the controls, since the controls only decrease $X^{1,N}(\cdot)$ (hence $X^{0,N}(\cdot)$) and $F^{0,N}(\cdot)$ is a reflection process at an upper boundary. Indeed, one can show that $F^{0,N}(\cdot)$ is asymptotically continuous with probability one. If it were not asymptotically continuous, then the asymptotic discontinuity and the properties of the other driving terms for $X^{0,N}(\cdot)$ would imply that asymptotically there is a jump to the interior of $[0, B_+^0]$ from the upper boundary, which is impossible, since the individual steps go to zero as $N \to \infty$ and $F^{0,N}(\cdot)$ can increase only on the boundary. The tightness of $\{Y^{0,N}(\cdot), N\}$ implies that $I^{0,N}(\cdot)$ can be nonzero only on set whose Lebesgue measure goes to zero as $N \to \infty$. Hence, $I^{0,N}(\cdot)$ has no asymptotic influence on the Doob–Meyer increasing process associated with the martingale $\tilde{D}^{0,N}(\cdot)$.

[3] By Theorem 3.3, this tightness assumption entails no loss of generality.

[4] The associated filtration here and in subsequent uses of "nonanticipative" is that generated by all of the processes $(X^i(\cdot), B^i(\cdot), F^i(\cdot), Y^i(\cdot), i = 0, 1)$.

Now, given the tightness, extract a weakly convergent subsequence, with the weak sense limits being denoted by $X^i(\cdot)$, etc. It can be shown that the weak sense limits $\tilde{D}^i(\cdot)$ are mutually independent Wiener processes which are independent of the $\tilde{W}^i(\cdot)$, and have the variance parameters λ_i. The Wiener property of the limits of the $\tilde{D}^{i,N}(\cdot)$ is proved as in [22, Theorem 3.1]. The mutual independence is a consequence of the independence conditions and the definition of "conditional mean instantaneous rate" via a conditional expectation, and an analogous computation is in [22, Theorem 3.1]. The fact that (3.2) holds follows from the weak convergence. The nonanticipativenss properties are proved by standard "martingale" means as in [29, Theorem 5.1], [22, Theorem 3.1], or [23, Theorem 2.1]

Definition. *The discounted cost function for* (3.2) *is*

$$C_\beta(x, F^1) = E\int_0^\infty e^{-\beta s} k(X^0(s))ds + E\int_0^\infty e^{-\beta s}\sum_i c_i dF^i(s). \tag{3.3}$$

Define an *admissible* control $F^1(\cdot)$ for (3.2) to be a nondecreasing process which is nonanticipative with respect to the Wiener processes. If $F^1(\cdot)$ is absolutely continuous, with derivative $u(\cdot)$, then we say that $u(\cdot)$ is admissible if it is nonnegative, measurable, and nonanticipative. Define $V_\beta(x) = \inf_{F^1} C_\beta^N(x, F^1)$, where the inf is over the admissible controls. We say that $F^{1,N}(\cdot)$ has a *derivative* which is bounded by R if for each t and $s > 0$, $F^{1,N}(t+s) - F^{1,N}(t) \leq Rs + 1/\sqrt{N}$. The $1/\sqrt{N}$ term is needed since $F^{1,N}(\cdot)$ is piecewise constant with jumps $1/\sqrt{N}$.

Theorem 3.2 (weak convergence). *Let* $X^N(0) \Rightarrow X(0)$ *Then*

$$V_\beta^N(X^N(0)) \to V_\beta(X(0)). \tag{3.4}$$

Discussion of the proof. By Theorem 3.3, it can be assumed without loss of generality that for each fixed $\epsilon > 0$, there is some set of ϵ-optimal controls for (2.11), (3.2), with uniformly bounded derivatives; hence the set is tight. As a result of the discounting, as noted in the proof of Theorem 3.3, we can suppose that there is $T_\epsilon < \infty$ such that the $F^{1,N}(\cdot)$ do not change after time T_ϵ. It is sufficient to work with this set below.

Given $\epsilon > 0$, let $F^{1,N}(\cdot)$ be a sequence of ϵ-optimal controls. Let $X^N(0) \to X(0)$. By Theorem 3.1, $\{X^{0,N}(\cdot), X^{1,N}(\cdot), F^{1,N}(\cdot)\}$ is tight. Let N_k index a weakly convergent subsequence with weak sense limits $(X^0(\cdot), X^1(\cdot), F^1(\cdot))$. Then, by the weak convergence and Fatou's Lemma

$$\liminf_k V_\beta^{N_k}(X^{N_k}(0)) \geq V_\beta(X(0)).$$

This and the ϵ-optimality of $F^{1,N}(\cdot)$ for each N implies that

$$\begin{aligned} \epsilon + \liminf_N V_\beta^N(X^N(0)) &\geq \liminf_N C_\beta^N(X^N(0), F^{1,N}) \\ &\geq C_\beta(X(0), F^1) \geq V_\beta(X(0)). \end{aligned} \tag{3.5}$$

Since ϵ is arbitrary, (3.5) implies that

$$\liminf_N V_\beta^N(X^N(0)) \geq V_\beta(X(0)). \tag{3.6}$$

Now we prove the reverse inequality to (3.5); namely,

$$\limsup_N V_\beta^N(X^N(0)) \leq V_\beta(X(0)). \tag{3.7}$$

To do this we apply the following widely useful approach. Note that, for each initial condition, the distribution of the limit process $X(\cdot)$ depends on the pair $(F^1(\cdot), B(\cdot))$. For each fixed $\epsilon > 0$, we first find an ϵ-optimal pair $(F^{\epsilon,1}(\cdot), B^{\epsilon}(\cdot))$ for the limit process with the property that there is a sequence of admissible controls $F^{\epsilon,1,N}(\cdot)$ for (2.11) such that (defining $B^{\epsilon,N}(\cdot)$ as $B^N(\cdot)$ was defined but under control $F^{\epsilon,1,N}(\cdot)$) $\{X^N(0), F^{\epsilon,1,N}(\cdot), B^{\epsilon,N}(\cdot))\}$ converges weakly to $(X(0), F^{\epsilon,1}(\cdot), B(\cdot))$ and also that

$$C_\beta^N(X^N(0), F^{\epsilon,1,N}) \to C_\beta(X(0), F^{\epsilon,1}) \leq V_\beta(X(0)) + \epsilon. \tag{3.8}$$

For any $\epsilon > 0$, there is an $F^{\epsilon,1}(\cdot)$ and a sequence of admissible controls $\{F^{\epsilon,1,N}(\cdot)\}$ satisfying the requirement (3.8). See, e.g., [28, section 5]. The sequence of controls given by the reference is admissible and the $F^{\epsilon,1,N}(T_\epsilon)$ are uniformly bounded. It does not necessarily have a bounded derivative, and might not even be tight in the Skorohod topology. (A time transformation method was used in the reference to circumvent the tightness problem. But the method used here is simpler in the current case, since Theorem 3.3 shows how to alter the control sequence (without loss of generality) so that we have bounded derivatives (hence tightness) with (3.8) holding (with perhaps ϵ replaced by 2ϵ).)

Now, (3.8) and the fact that (due to the nonoptimality of $F^{\epsilon,1,N}(\cdot)$)

$$V_\beta^N(X^N(0)) \leq C_\beta^N(X^N(0), F^{\epsilon,1,N})$$

yields (3.4).

Comment on the optimal controls. The comments on the form of the optimal control in section 7 also hold for the discounted cost problem. In particular, numerical data show that there is a piecewise linear or nearly linear switching curve such that the optimal control is to reject above and accept below, with any decision allowed when on the curve. For large N, this control will be nearly optimal for the physical system. These comments on the shape of the switching curves are based on (very consistent) numerical computations, but not on proofs.

Comment on tightness. An ϵ-optimal sequence $\{F^{1,N}(\cdot), N\}$ need not be tight in general. Let $X^{1,N}(t_0) > 0$. Consider the example of a control which rejects until $X^{1,N}(t)$ reaches the value zero, and then stops. Since the required time for $X^{1,N}(\cdot)$ to reach zero is of the order of $X^{1,N}(t_0)/\sqrt{N}$, the control clearly converges to a step function in an obvious way. But, owing to the fact that it increases in small steps (of size $1/\sqrt{N}$), the sequence is not tight in the Skorohod topology. There are many ways of dealing with this problem. We can adapt the time transformation method of [26, 24]. But, because of the relative simplicity of the dynamics and cost function, there is a simpler way, which avoids the extra notation and concepts.

THEOREM 3.3. *It can be supposed that the controls in Theorems* 3.1 *and* 3.2 *are tight. In fact, it can be supposed that they have uniformly bounded derivatives for each* $\epsilon > 0$.

Proof. Fix $\epsilon > 0$. Since $B_+^0 < \infty$, $\{k(X^{0,N}(t)), k(X^0(t)); N, t\}$ is uniformly bounded.[5] Thus, due to the discounting, there is $T_\epsilon < \infty$ and a sequence of $\epsilon/2$-optimal controls which do not change after time T_ϵ. Note that the control which is identically zero has uniformly (in N) bounded costs.

[5] Even if $B_+^0 = \infty$, by imposing a growth rate $k(x^0) = O(|x^0|^{1+\delta})$ for large x^0, $0 \leq \delta < 1$, and assuming that $\sup_N E|X^{0,N}(0)|^2 < \infty$, $Ek(X^{0,N}(t))$ is at most $O(t^2)$ for large t, uniformly in the control (and analogously for the limit system).

The existence of $T_\epsilon < \infty$ can be seen from the following argument. From any time T on, and with no control after that time, the limit cost is

$$E\int_T^\infty e^{-\beta t}k(X^0(t))dt + E\int_T^\infty e^{-\beta t}c_0 dF^0(t).$$

The first term obviously goes to zero as $T \to \infty$, uniformly in the past values of the control. The same thing can be said of the second term, owing to the properties of the solution to (3.2a). A similar argument applies to the physical system and controls $F^{1,N}(\cdot)$. We need only work with controls for which the sequence of costs is uniformly bounded.

For large enough $K < \infty$, the sequence defined by $F_K^{1,N}(\cdot) = F^{1,N}(\cdot) \wedge K$ will be $3\epsilon/4$-optimal, and similarly for $F_K^1(\cdot) = F^1(\cdot) \wedge K$. To see this, note that whatever the controls in the previous paragraph are, the boundedness of the costs imply that

$$\sup_N EF^{1,N}(T_\epsilon) < \infty, \quad EF^1(T_\epsilon) < \infty. \tag{3.9}$$

Furthermore,

$$\begin{aligned} \lim_K \sup_N P\left\{\left[F^{1,N}(T_\epsilon) - F^{1,N}(T_\epsilon) \wedge K\right] \neq 0\right\} = 0, \\ \lim\nolimits_K P\left\{\left[F^1(T_\epsilon) - F^1(T_\epsilon) \wedge K\right] \neq 0\right\} = 0. \end{aligned} \tag{3.10}$$

(3.9) and (3.10) and a straightforward analysis using Fatou's Lemma implies that the costs and systems associated with the use of $F^{i,N}(\cdot) \wedge K$ (resp., $F^i(\cdot) \wedge K$ for the limit system) are asymptotically (as $K \to \infty$) no worse (uniformly in N) than the costs for the original untruncated controls.

Now that we know that there are $3\epsilon/4$-optimal controls $F^{1,N}(\cdot), F^1(\cdot)$ which do not increase after T_ϵ and that are uniformly bounded, we can show that we can bound the derivative as well: More precisely, we can show that there are $R < \infty$ and ϵ-optimal controls which satisfy:

$$F_R^1(t+s) - F_R^1(t) \leq Rs, \tag{3.11}$$

$$F_R^{1,N}(t+s) - F_R^{1,N}(t) \leq Rs + 1/\sqrt{N}, \tag{3.12}$$

for all $t, s > 0$. In particular, let $F_R^1(\cdot)$ denote the largest control which satisfies (3.11), but is no greater than $F^1(\cdot)$, and let $F_R^{1,N}(\cdot)$ be the largest control which satisfies (3.12) and which is no greater than $F^{1,N}(\cdot)$.

Since $K < \infty$ and R is as large as desired, any jump in $F^{1,N}(\cdot)$ can be reached by $F_R^{1,N}(\cdot)$ in an arbitrarily short (uniformly in N and in the realization) time afterwards. Thus, excluding a set of measure which goes to zero (uniformly in N) as $R \to \infty$, $F^{1,N}(t) - F_R^{1,N}(t)$ goes to zero (uniformly in N) as $R \to \infty$. This and the boundedness of the F functions imply that $X^{1,N}(t) - X_R^{1,N}(t)$ is bounded and (excluding sets of arbitrarily small measure) converges to zero uniformly in N. Similar remarks hold for $F^{1,N}(\cdot), F_R^{1,N}(\cdot)$. These results imply that the costs converge as well.

The $1/\sqrt{N}$ term appears in (3.12) since $F^{1,N}(\cdot)$ is piecewise constant with an increment of $1/\sqrt{N}$ at each rejection.

4. Upper limit to the bandwidth for the BE sharing customers. In the model of the previous two sections, the class 0 customers shared the available bandwidth, whatever it was, and used *all* of it. In general, it might not be possible for all of the available bandwidth to be used. For example, there might be local restrictions on the rate at which data can enter the channel buffer (e.g., bounded modem speed, etc.). This possibility changes the problem a little, and we will indicate the few required adjustments. Such examples are further illustrations of the versatility of the approach.

Suppose that the *maximum* bandwidth that any single class 0 customer can use is C_0. The main difference in the development concerns the departure process for class 0 customers and the structure of the appropriate cost function. We now *redefine*

$$X^{0,N}(t) = \frac{\#\text{ of class 0 customers at time } t - N\lambda^{0,N}/(C_0\mu^{0,N})}{\sqrt{N}}.$$

Note that now $X^{0,N}(t)$ is centered around a *mean value,* assuming that each class 0 customer uses exactly C_0 units of bandwidth. In sections 2 and 3, the number of class 0 customers in the system was $O(\sqrt{N})$, and $X^{0,N}(t)$ measured that actual number, scaled by $1/\sqrt{N}$. Now, the number in the system will be $O(N)$, and $X^{0,N}(t)$ measures the deviation from the mean number, scaled by $1/\sqrt{N}$. We suppose that class 0 customers are rejected if $X^{0,N}(t) \geq B^0_+$, where $B^0_+ < \infty$. Theoretical and numerical results show that this will have negligible effect if B^0_+ is large.

The martingale decomposition (2.9) remains valid, but the instantaneous conditional mean departure rate is different, being determined by whether or not the available capacity per class 0 customer is greater than C_0. The conditional mean instantaneous rate at which $D^{0,N}(\cdot)$ increases at time t is

$$\frac{\mu^{0,N}}{\sqrt{N}}[\#\text{ of class 0 in system at } t] \times \min\left[\frac{\text{available BW at } t}{\#\text{ of class 0 in system at } t},\ C_0\right] I^{0,N}(t),$$

which equals

$$\begin{aligned}
&I^{0,N}(t)\frac{\mu^{0,N}}{\sqrt{N}} \times \min\left[\text{available BW at } t,\ C_0(\#\text{ of class 0 in system at } t)\right]\\
&= I^{0,N}(t)\frac{\mu^{0,N}}{\sqrt{N}}\times\\
&\min\left[\frac{N\lambda^{0,N}}{\mu^{0,N}} + b\sqrt{N} - \sqrt{N}X^{1,N}(t),\ C_0\left(\frac{\lambda^{0,N}N}{C_0\mu^{0,N}} + \sqrt{N}X^{0,N}(t)\right)\right]\\
&= I^{0,N}(t)\lambda^{0,N}\sqrt{N} + I^{0,N}(t)\mu^{0,N}g_1(X^N(t)),
\end{aligned} \tag{4.1}$$

where we define

$$g_1(x) = \min\left[b - x^1, C_0x^0\right]. \tag{4.2}$$

Now, analogously to what was done in section 2, we can write the decomposition as

$$D^{0,N}(t) = \lambda^{0,N}\sqrt{N}t + \int_0^t \mu^{0,N}g_1(X^N(s))ds + \tilde{D}^{0,N}(t) - Y^{0,N}(t),$$

where the Doob–Meyer increasing process associated with the martingale $\tilde{D}^{0,N}(\cdot)$ is

$$\langle\tilde{D}^{0,N}\rangle(t) = \int_0^t \left[\lambda^{0,N} + \frac{1}{\sqrt{N}}\mu^{0,N}g_1(X^N(s))\right] I^{0,N}(s)ds.$$

The dynamical equation is (2.11) with (2.11a) replaced by

$$\begin{aligned} X^{0,N}(t) &= X^{0,N}(0) + W^{0,N}(S^{0,N}(t)) \\ &\qquad - \mu^{0,N} \int_0^t g_1(X^N(s))ds - \tilde{D}^{0,N}(t) - F^{0,N}(t) + Y^{0,N}(t), \end{aligned} \tag{4.3}$$

and the limit equation is (3.2) with (3.2a) replaced by

$$X^0(t) = X^0(0) - \mu_0 \int_0^t g_1(X(s))ds + B^0(t) - F^0(t). \tag{4.4}$$

In the cost function (3.1), the function $k(\cdot)$ was assumed to be nonnegative. This made sense since $X^{0,N}(t)$ was nonnegative. Now, since $X^{0,N}(t)$ can take any sign, we suppose that $k(x^0)$ takes the sign of x^0 and is zero if $x^0 = 0$. Note that for large negative x^0, the departure rate is essentially limited by the C_0 limitation, and the control has little effect. Also, suppose that

$$|k(x^0)| = O(|x^0|), \quad \sup_N E\left|X^{0,N}(0)\right|^2 < \infty. \tag{4.5}$$

Under the given conditions Theorems 3.1 to 3.3 hold.

The savings in waiting time for the class 0 customers (for the controlled problem) are of the order of that of the model in sections 2 and 3. But, since there are many more customers in the system at any time, the savings per customer is less. We are essentially concerned with "marginal" savings, at a "marginal" cost.

Comments on the proofs. The proofs outlined in section 3 work here as well, with essentially the same details. The only differences are due to the fact that in the present case the $X^{0,N}(t)$ are not bounded below. But in the proofs, the second moment bound

$$\sup_{N,t,F^{1,N}} E\left|X^{0,N}(t)\right|^2 < \infty \tag{4.6}$$

is used in lieu of the zero lower bound. (4.6) is proved by use of a dominating system. The second moments are bounded by a constant plus the second moments of the following system, which is defined on the interval $(-\infty, 0]$, and the reflection term $F^{0,N}(\cdot)$ now acts at the origin, and keeps the state nonpositive:

$$X^{0,N}(t) = X^{0,N}(0) + W^{0,N}(S^{0,N}(t)) - \mu^{0,N} \int_0^t X^{0,N}(s)ds - \tilde{D}^{0,N}(t) - F^{0,N}(t). \tag{4.7}$$

The proof of (4.6) for the model (4.7) is done by a Liapunov function technique and can be found in [22, p. 771]

5. Extension. Several BE and GP subclasses. The developments in the previous sections can be extended to the case where there are multiple subclasses of any of the classes (and similarly for the models in the subsequent sections). We will illustrate only one of the many possibilities, working with the setup in sections 2 and 3. Suppose that there are now two types of BE (class 0) customers, called class 01 and class 02, with parameters $\lambda^{0i,N}, \mu^{0i,N}, i = 1, 2$. We suppose that the natural analogs of the conditions in sections 2 and 3 hold. There is a surprisng and very useful degeneracy, which simplifies both the analytical problem and the numerical analysis.

Analogous to (2.1), we let the channel capacity be

$$C_N = N\left[\sum_i \frac{\lambda^{0i,N}}{\mu^{0i,N}} + \frac{\lambda^{1,N}}{\mu^{1,N}}\right] + \sqrt{N}b. \tag{5.1}$$

The bandwidth available at time t for both subclasses 01 and 02 is

$$\sum_i N\frac{\lambda^{0i,N}}{\mu^{0i,N}} - \sqrt{N}X^{1,N}(t) + \sqrt{N}b, \quad b > 0.$$

The processes are defined analogously to what was done in section 2; e.g., $D^{0i,N}(\cdot)$ is the number of departures of subclass $0i$ by time t, divided by $\sqrt{N}$.

Until otherwise noted, let us assume that the available bandwidth is shared equally among all class 0 customers, irrespective of the subclass. Then the total conditional mean instantaneous rate at which $D^{0i,N}(\cdot)$ increases at t is (following the idea used in section 2 and assuming that there are class 0 customers in the system)

$$\frac{\mu^{0i,N}}{\sqrt{N}}\,[\text{\# of subclass } 0i \text{ in system at } t]\,\frac{\text{avail BW at } t}{\text{total \# of class 0 in system at } t},$$

which equals

$$\mu^{0i,N}X^{0i,N}(t)\frac{\sum_j N\frac{\lambda^{0j,N}}{\mu^{0j,N}} - \sqrt{N}X^{1,N}(t) + \sqrt{N}b}{\sqrt{N}\sum_j X^{0j,N}(t)}. \tag{5.2}$$

Define

$$\bar{a}^N = \frac{\lambda^{01,N}/\mu^{01,N}}{\sum_j \lambda^{0j,N}/\mu^{0j,N}}, \quad \bar{a} = \lim_N \bar{a}^N. \tag{5.3}$$

The system is degenerate in that if the costs are bounded in N, then the ratios $X^{01,N}(t)/(X^{01,N}(t)+X^{02,N}(t))$ converge to $\bar{a}$ as $N \to \infty$. Thus we need only analyze the system with class 1 and one of the subclasses $0i$. Only an informal argument will be given. We note in passing that this convergence relation is an example of what is called state space collapse in the heavy traffic analysis of queueing systems. It is not the usual type, which is concerned with multiclass queues under the workload formulation.

Let us examine the mean rates of increase of $A^{0i,N}(\cdot)$ and $D^{0i,N}(\cdot)$. The "mean rate" at which $A^{0i,N}(\cdot)$ increases is $\sqrt{N}\lambda^{0i,N}$. Define $B^N = \sum_i[\lambda^{0i,N}/\mu^{0i,N}]$. Set

$$a^N(t) = \frac{X^{01,N}(t)}{X^{01,N}(t) + X^{02,N}(t)}.$$

Using the fact that the available bandwidth is partitioned equally among all the class 0 customers, the conditional mean instantaneous rates at which $D^{0i,N}(\cdot), i = 1, 2$, resp., increase at t are, resp.,

$$\left[\sqrt{N}B^N - X^{1,N}(t) + b\right]\mu^{01,N}a^N(t),$$

$$\left[\sqrt{N}B^N - X^{1,N}(t) + b\right]\mu^{02,N}\left(1 - a^N(t)\right).$$

The differences of the dominant arrival and departure terms for the two subclasses are, resp.,

$$\sqrt{N}\left[\lambda^{01,N} - B^N \mu^{01,N} a^N(t)\right], \tag{5.4a}$$

$$\sqrt{N}\left[\lambda^{02,N} - B^N \mu^{02,N}\left(1 - a^N(t)\right)\right]. \tag{5.4b}$$

For the case of section 2, the analogs of these terms have the value zero.

Let $\epsilon > 0$ be small. If for large N, $a^N(t) \notin [\bar{a} - \epsilon, \bar{a} + \epsilon]$, then (5.4a) implies that there is a large force (of the order of $\sqrt{N}$) returning it to this interval. Similarly, (5.4b) implies that $(1 - \alpha^N(t))$ must be very close to $(1 - \bar{a})$. The contribution of the nondominant terms is relatively small in comparison.

The degeneracy situation is similar if there are more than two subclasses.

Many interesting variations of the multiple subclass problem can be analyzed. For example, we might wish to alter the above formulation to allow each of the $0i$ subclasses a different fraction of the available bandwidth. More concretely, suppose that there are positive numbers k_i such that for each unit of bandwidth allocated to a customer of subclass 01, we allocate k_2/k_1 units of bandwidth to each customer of subclass 02. Then the dominant term in the conditional mean instantaneous rate at which $D^{0i,N}(\cdot)$ increases at t is

$$\mu^{0i,N} k_i X^{0i,N}(t) \frac{B^N N}{\sqrt{N}\left(k_1 X^{01,N}(t) + k_2 X^{02,N}(t)\right)}. \tag{5.5}$$

Redefine $a^N(t)$:

$$a^N(t) = \frac{k_1 X^{01,N}(t)}{\sum_j k_i X^{0i,N}(t)}. \tag{5.6}$$

Then, the difference between the dominant input and output terms is (5.4), but with the new value of $a^N(t)$ used. Thus, we see that the new value of $a^N(t)$ is very close to $\bar{a}$ for large N.

Note that the weighted fair-share for different types of BE customers in the above equations is the one defined by the ATM forum for sharing bandwidth among ABR users (see [1, Section I.3]).

6. Multicast: The limit dynamical equations and the discounted cost function. Now consider the case where there are two channels. There are three classes of customers. Class 0 is as in section 2, but must be transmitted simultaneously and with the same instantaneous rate on both channels. Class i, $i = 1, 2$, is to be transmitted on channel i only. We make the natural analogs of the assumptions of sections 2 and 3, defining $A^{i,N}(\cdot), D^{i,N}(\cdot), X^{i,N}(\cdot), i = 0, 1, 2$, etc., analogously to what was done there. Analogous to (2.1), the capacity of channel i is assumed to be

$$C_N^i = N \sum_i \frac{\lambda^{i,N}}{\mu^{i,N}} + b_i \sqrt{N}, \quad b_i > 0. \tag{6.1}$$

Any number of channels and 0 subclasses could also be used, with an arbitrary assignment of the subclasses to the channels.

At time t, the bandwidth available for class 0 customers on channel i is

$$N \frac{\lambda^{0,N}}{\mu^{0,N}} + b^i \sqrt{N} - X^{i,N}(t)\sqrt{N}.$$

Thus, the conditional mean instantaneous rate at which $D^{0,N}(\cdot)$ increases at t is determined by the channel with the largest available bandwidth and (analogous to what was done in section 2) is

$$\text{(6.2)}\quad \frac{\mu^{0,N}\sqrt{N}X^{0,N}(t)}{\sqrt{N}} \times \min\left[\frac{N\frac{\lambda^{0,N}}{\mu^{0,N}}+b_1\sqrt{N}-X^{1,N}(t)\sqrt{N}}{\sqrt{N}X^{0,N}(t)},\ \frac{N\frac{\lambda^{0,N}}{\mu^{0,N}}+b_2\sqrt{N}-X^{2,N}(t)\sqrt{N}}{\sqrt{N}X^{0,N}(t)}\right] I^{0,N}(t).$$

This equals

$$\left[\sqrt{N}\lambda^{0,N}+\mu^{0,N}g_2(X^N(t))\right] I^{0,N}(t), \tag{6.3}$$

where we define

$$g_2(x)=\min\left[b_1-X^{1,N}(t),\ b_2-X^{2,N}(t)\right]. \tag{6.4}$$

Thus analogous to what was done in section 2, we can write

$$D^{0,N}(t)=\lambda^{0,N}\sqrt{N}t+\int_0^t g_2(X^N(s))ds+\tilde{D}^{0,N}(t)-Y^{0,N}(t),$$

where the Doob–Meyer increasing process associated with the martingale is

$$\langle\tilde{D}^{0,N}\rangle(t)=\int_0^t\left[\lambda^{0,N}+\frac{\mu^{0,N}}{\sqrt{N}}g_2(X^N(s))\right]I^{0,N}(s)ds.$$

The analog of (2.11) is

$$\begin{aligned} X^{0,N}(t)=X^{0,N}(0)+W^{0,N}(S^{0,N}(t))-\mu^{0,N}\int_0^t g_2(X^N(s))ds\\ -\tilde{D}^{0,N}(t)+Y^{0,N}(t)-F^{0,N}(t), \end{aligned} \tag{6.6a}$$

and for $i=1,2$,

$$\begin{aligned} X^{i,N}(t)=X^{i,N}(0)+W^{i,N}(S^{i,N}(t))-\mu^{i,N}\int_0^t X^{i,N}(s)ds\\ -\tilde{D}^{i,N}(t)-F^{i,N}(t)-U^{i,N}(t), \end{aligned} \tag{6.6b}$$

where the $Y^{0,N}(t)$ term compensates for the fact that there are no departures of class 0 customers at time t if $X^{0,N}(t)=0$, and $U^{i,N}(t)$ compensates for the class i arrivals lost due to a full system (when the entire channel is occupied by class i customers).

The discounted cost function is still (3.1), but now the sum has three terms. The analysis given in section 3 holds here in the same way and the limit equations are

$$X^0(t)=X^0(0)-\mu_0\int_0^t g_2(X(s))ds+B^0(t)+Y^0(t)-F^0(t), \tag{6.7a}$$

$$X^i(t)=X^i(0)-\mu_i\int_0^t X^i(s)ds+B^i(t)-F^i(t),\quad i=1,2, \tag{6.7b}$$

where the $B^i(\cdot)$ are mutually independent Wiener processes with variance parameters $\lambda_i(\sigma_i^2+1)$. The associated cost function is (3.3), and we still have

$$V_\beta^N(X^N(0)) \to V_\beta(X(0)), \tag{6.8}$$

if $X^N(0) \Rightarrow X(0)$.

The model of section 4. Now, suppose that each class 0 customer can use at most a bandwidth C_0. Then define $X^{0,N}(t)$ as in section 4, and let $B_+^0 < \infty$. The development is a combination of those of sections 3 and 4. Now, the conditional mean instantaneous rate at which $D^{0,N}(\cdot)$ increases at t is obtained as the minimum of three terms, depending on whether the available capacities in channels 1, 2, or C_0 are the limiting factor. It is

$$\begin{aligned}&\frac{\mu^{0,N}}{\sqrt{N}}[\#\text{ of class 0 in system at } t]\\ &\min\left[\frac{\text{avail BW in ch 1 at } t}{\#\text{ of class 0 in system at } t}, \frac{\text{avail BW in ch 2 at } t}{\#\text{ of class 0 in system at } t}, C_0\right] I^{0,N}(t).\end{aligned} \tag{6.9}$$

Define

$$g_3(x) = \min\left[b_1 - x^1, b_2 - x^2, C_0 x^0\right]. \tag{6.10}$$

Then (6.9) can be written as

$$\left[\sqrt{N}\lambda^{0,N} + \mu^{0,N} g_3(X^N(t))\right] I^{0,N}(t).$$

All of the previous results continue to hold with $g_3(\cdot)$ replacing $g_2(\cdot)$.

7. The ergodic cost function: The basic model: Bounded control rate.

For concreteness, we work with the system model and assumptions of sections 2 and 3, although all of the results hold for all of the other models. In this section, we will suppose that the controls $F^{1,N}(\cdot)$ and $F^1(\cdot)$ (for the limit system) have bounded derivatives in the sense used in Theorem 3.3, as follows: There is a constant R, which can be as large as we wish, such that $\dot{F}^1(t) \le R$ for all t, and for all $t, s > 0$, $F^{1,N}(t+s) - F^{1,N}(t) \le Rs + 1/\sqrt{N}$. Thus, the maximum "rate" of refusing admission to class 1 customers is bounded by $\sqrt{N}R$.

The reasonableness of the bounded derivative assumption is also seen from the form of the limit equation (3.2), which (informally) suggests that one loses very little by bounding the derivative of $F^1(\cdot)$. Furthermore, it is completely borne out by our numerical data. The next section shows that we can make this assumption in the proofs with no loss of generality.

Although useful in applications, the mathematical reason for the assumption of bounded control "derivatives" concerns the mathematics of the ergodic cost problem. Little is known about the ergodic cost problem for the limit system when the control functions are arbitrary right continuous functions. But a great deal is known when they have uniformly bounded derivatives. In that case, for the current nondegenerate model (3.2), there is an optimal feedback control which is time independent and the optimal value $\bar{\gamma}_R(x)$ (defined below) does not depend on x. More importantly, for our purposes, for any $\epsilon > 0$, there is an ϵ-optimal time independent feedback control $u^\epsilon(\cdot)$ such that $u^\epsilon(\cdot) = \dot{F}^{\epsilon,1}(\cdot)$ is arbitrarily smooth, and under which there is a unique stationary measure. The $F^{\epsilon,1}(\cdot)$ plays the role of the $F^{1,\epsilon}(\cdot)$ in Theorem 3.2. The basic convergence results are quite technical. They are in [20] for the unconstrained (no reflecting boundaries) problem, with extensions to the constrained problem being in [22, 23]. Indeed, under our basic setup, the needed convergence results can be obtained from [23] by appropriate identification of terms.

Define the cost functions

$$\mathcal{C}^N(X^N(0),T,F^{1,N}) = \int_0^T k(X^{0,N}(s))ds + \sum_i c_i F^{i,N}(T),$$

$$\bar{\gamma}_R^N(X^N(0),T) = \inf_{F^{1,N}} E\mathcal{C}^N(X^N(0),T,F^{1,N})/T,$$

$$\bar{\gamma}_R^N(X^N(0)) = \limsup_T \bar{\gamma}_R^N(X^N(0),T).$$

For the limit system, define the analogous quantities, with the N dropped. If there is no rate R restriction, we drop the subscript R. We also suppose that (with little loss of generality)

$$\sup_N E|X^{1,N}(0)|^2 < \infty. \tag{7.1}$$

In section 3 it was shown that, for the discounted cost problem and large enough R, we can get as close to optimality as we wish. The proof in Theorem 3.3 used the fact that the discounting implied that we need concern ourselves only with a finite time interval. The proof is more subtle for the ergodic cost problem and is given in the next section.

The results in [23] will apply if we have tightness of the doubly indexed (both t and N are indices now) set of processes

$$\left\{X^N(t+\cdot), B^N(t+\cdot) - B^N(t), F^{1,N}(t+\cdot) - F^{1,N}(t); N, t\right\}. \tag{7.2}$$

The tightness holds for the set of $F^{1,N}(t+\cdot) - F^{1,N}(t)$ processes by the assumption on boundedness of the derivative. The set $\{W^{i,N}(S^{i,N}(t+\cdot)) - W^{i,N}(S^{i,N}(t); N, t\}$ is tight due to the independence properties of the interarrival intervals, (2.2), the weak convergence of the $S^{i,N}(t+\cdot) - S^{i,N}(t)$, as $N \to \infty$ and for any sequence t, and the use of the criterion (2.14), (2.15).

A standard Liapunov function argument (using the Liapunov function $|X^{1,N}|^2$) and the "R-derivative" restrictions on the controls can be used to prove directly that

$$\sup_{t,N,F^{1,N}} E\left|X^{1,N}(t)\right|^2 < \infty. \tag{7.3}$$

Here, the sup is over the $F^{1,N}$ which satisfy the R-derivative restriction. Then, tightness can be shown for the set of $\tilde{D}^{i,N}(t+\cdot) - \tilde{D}^{i,N}(t)$ processes by a direct application of the criterion (2.14), (2.15) and the use of (7.3) to bound the expectation of the Doob–Meyer processes associated with $\tilde{D}^{i,N}(t+\cdot) - \tilde{D}^{i,N}(t)$. The proof of the tightness of the doubly indexed sequence $\{X^{i,N}(t+\cdot); N, t\}$ is then the same as the proof of tightness of $\{X^{i,N}(\cdot), t\}$, where the initial conditions vary over a tight set (the $X^{0,N}(t)$ are bounded by B_0^+).

Given the tightness of (7.2) and the nondegeneracy of the limit system (3.2) (the set of driving Wiener processes is nondegenerate; in fact, the components are mutually independent with positive variances), the following results follow by a direct application of the results and ideas in [23]:

$$\bar{\gamma}_R^N(X^N(0),T) \to \bar{\gamma}_R, \tag{7.4}$$

as $T \to \infty$ and $N \to \infty$ in any way at all, where $\bar{\gamma}_R$ is the infimum of the costs for the limit system over controls with derivatives bounded by R, and it does not depend on the initial condition.

Furthermore, for any $\epsilon > 0$,

$$\lim_{N,T} P\left\{\frac{\mathcal{C}^N(X^N(0),T,F^{1,N})}{T} \le \bar{\gamma}_R - \epsilon\right\} = 0, \tag{7.5}$$

where N, T can go to their limits in any way at all, and $F^{1,N}(\cdot)$ is an arbitrary sequence of controls. There is a converse to (7.5) which says that a good control for the limit system is a good control for the physical system. Given $\epsilon > 0$, let $F^{\epsilon,1}(\cdot)$ be an ϵ-optimal control with smooth derivative $u^\epsilon(\cdot)$ and an adaptation $F^{\epsilon,1,N}(\cdot)$ to the physical system such that

$$\lim_{N,T} P\left\{\frac{\mathcal{C}^N(X^N(0),T,F^{\epsilon,1,N})}{T} \ge \bar{\gamma}_R + 2\epsilon\right\} = 0, \tag{7.6}$$

The results of the next section imply that we can replace $\bar{\gamma}_R$ in (7.5) and (7.6) by $\bar{\gamma}$.

The control $u^\epsilon(\cdot)$ can be adapted for use on the physical system in many ways, for large N. For example, by rejecting an arrival of class 1 at t with probability (conditioned on the past system data) $u^\epsilon(X^N(t))/[\lambda^{1,N}\sqrt{N}]$. Alternatively, we need not have the rejection choices being random, provided that $\sqrt{N}$ times the number rejected when the state is "near" x converges to $u^\epsilon(x)$ as $N \to \infty$.

Comments on the controls. Numerical data show that the derivative $\dot{F}(t) = u(t)$ of the optimal control takes either the value R or zero, with the regions separated by a piecewise linear or nearly linear switching curve. One applies this control to the physical system as in the last paragraph. (This procedure is asymptotically equivalent to rejecting all arrivals when the state is above the switching curve.) Equation (7.6) holds for such discontinuous controls as well. This is important in applications since such controls are easily implemented. Numerical data show that the switching curves converge nicely to piecewise smooth (or even linear) curves as $R \to \infty$. (7.6) holds for this curve a well. Then we reject all arrivals of class 1 when the state is above the switching curve. Analogous remarks hold for the discounted cost problem.

Comment. Note that both (7.5) and (7.6) deal with pathwise average costs, not with average costs. Since any application is a single realization, the convergence of pathwise average costs is more important than the convergence of expectations. The inequalities (7.5) and (7.6) say that for large N, the optimal controls for the physical problem are (asymptotically) only negligibly better than the use of a nice almost optimal control for the limit system.

Finally, we simply note without further comment that the methods in [20, 22, 23] can be adapted to prove that $\lim_{N,\beta\to 0} \beta V_\beta^N(X^N(0)) = \bar{\gamma}$.

8. The ergodic cost problem: The basic model: Arbitrary controls.

Now, return to the problem of bounded derivative controls. We will show that we can approximate any optimal or nearly optimal control by a control which has a "derivative" bounded by R, for large enough R.

THEOREM 8.1.

$$\lim_R [\bar\gamma - \bar\gamma_R] = 0. \tag{8.1a}$$

$$\lim_R \limsup_N \left[\bar\gamma^N(x) - \bar\gamma_R^N(x)\right] = 0. \tag{8.1b}$$

Proof. A detailed outline of the steps will be given. Unlike as in Theorem 3.3, we cannot restrict ourselves to a finite interval. We need to show that for any $\delta > 0$, there is $R_\delta < \infty$ such that there are δ-optimal controls for both the physical and the limit system with bounded rate R_δ.

The development proceeds in several steps. The steps will be outlined (informally to save space) for the physical system. The details are a little simpler for the limit system.

1. Given $\epsilon > 0$, show that there is a $B_\epsilon < \infty$ such that the optimal cost will change by no more than ϵ if we do not reject when $X^{1,N}(t) < -B_\epsilon$.

2. Let $\epsilon > 0$. Allowing only controls which do not reject if $X^{1,N}(t) < -B$, for some given $0 < B < \infty$, show that there is $K_\epsilon < \infty$ such that if we further restrict the controls such that the increments $F^{1,N}(n+1) - F^{1,N}(n)$ are at most K_ϵ for all n, N, then the optimal cost will change by no more than ϵ.

3. Let $\epsilon > 0$. Allowing only controls satisfying the restrictions of the first two steps for some finite B, K, show that there is $R_\epsilon < \infty$ such that the optimal cost will change by at most ϵ if we further restrict the controls to have maximum "derivative" R_ϵ.

Step 1 is the least difficult to accept even without a proof, since it is quite reasonable that there is a $B < \infty$ such that an optimal or nearly optimal control that would not reject if $X^{1,N}(t) \le -B$. The proof is a formalization of the following idea. Given a control $F^{1,N}(\cdot)$ and a $B > 0$, define another control $F_B^{1,N}(t) \le F^{1,N}(t)$, where $F_B^{1,N}(t)$ is as close as possible to $F^{1,N}(t)$, but acts only when $X^{1,N}(t) \ge -B$. For large B, the change in $X^{0,N}(\cdot)$ is slight. To save space, we concentrate on the outline for the other steps.

Thus, we start by supposing that there is $0 < B < \infty$ such that there are no rejections if $X^{1,N}(t) \le -B$. We will show that, given $\epsilon > 0$, there is $K_\epsilon < \infty$, such that we lose less than ϵ in the cost if we restrict the control to satisfy $F^{1,N}(n+1) - F^{1,N}(n) \le K_\epsilon$ for all N, n, ω.

Because we do not reject if $X^{1,N}(t) \le -B$, a Liapunov function argument can be used to get that there is $C < \infty$ such that

$$\sup_{N,t,F^{1,N}} E\left|X^{1,N}(t)\right|^2 \le C. \tag{8.2}$$

Also, the same $-B$ restriction and (8.2) can be used to show that

$$\sup_{n,N,F^{1,N}} E\left[F^{1,N}(n+1) - F^{1,N}(n)\right]^2 < \infty. \tag{8.3}$$

The sup in (8.2) is over all controls satisfying our $-B$ restriction. The proof of (8.3) computes a worst case on each interval, which is a control taking $X^{1,N}(n)$ satisfying only (8.2) to $-B$ as quickly as possible, then keeping it there until the end of the interval, and repeating on the next interval, etc. The uniform mean square boundedness of the part due to keeping $X^{1,N}(\cdot)$ at $-B$ on $[n, n+1]$ follows from the reflection mapping and the mean square bounds on the martingales driving (2.11b). For the reflection mapping and the Lipschitz continuity of the reflection term as a function of the driving processes, see [8], [7, Proposition 2.1].

Given any $F^{1,N}(\cdot)$ satisfying our restriction, we proceed to approximate it by bounding the increments by K. The approximation will be denoted by $F_K^{1,N}(\cdot)$, and the associated processes denoted by $X_K^N(\cdot)$. Define $F_K^{1,N}(\cdot)$ such that it satisfies the restriction $F_K^{1,N}(n+1) - F_K^{1,N}(n) \leq K$, it is no greater than $F^{1,N}(\cdot)$, and tries to keep $X_K^{1,N}(\cdot)$ as close as possible to $X^{1,N}(\cdot)$.

We always have $X_K^{1,N}(t) \geq X^{1,N}(t)$ and hence $X_K^{0,N}(t) \geq X^{0,N}(t)$. It is not hard to see that, for large enough K, $X_K^{1,N}(\cdot)$ will repeatedly catch up to and equal $X^{1,N}(\cdot)$. We can decompose time into successive intervals where $X^{1,N}(t) < X_K^{1,N}(t)$ and where $X^{1,N}(t) = X_K^{1,N}(t)$. The key to the proof of step 2 is the observation that, as $K \to \infty$, a larger percentage of time will be taken up by the latter intervals. More precisely, for any $T_0 < \infty$, it can be shown that

$$\lim_K \limsup_N \sup_{t,F^{1,N}} P\left\{X^{1,N}(s) \neq X_K^{1,N}(s) \text{ for some } s \in [t, t+T_0]\right\} = 0. \tag{8.4}$$

Equation (8.4) follows from the observations made before it. Choose the control $F_K^{1,N}(\cdot)$ as described. Then, starting at time $t-k$ for large k, the probability that $X_K^{1,N}(\cdot)$ catches up to $X^{1,N}(\cdot)$ and equals it on $[t, t+T_0]$ goes to unity as $K \to \infty$.

Note that if $X_K^{0,N}(t) = 0$, then $X^{0,N}(t) = 0$. We next bound the "return times" to the boundary $x^0 = 0$. Indeed, it can be shown that

$$\lim_{T\to\infty} \limsup_N \sup_{t,F_K^{1,N},K} \sup_\omega P\left\{X_K^{0,N}(t+s) \neq 0, \text{ for some } s \leq T \middle| \text{data to } t\right\} = 0. \tag{8.5}$$

This can be shown by a weak convergence argument, using the fact that it holds for the limit process, as follows. The worst case for proving (8.5) is where there is no control since the control only decreases $X^{0,N}(t)$. Thus, suppose that there are $\rho > 0$, t_n and $N_n \to \infty, T_n \to \infty$ such that (no control)

$$\lim_n \sup_\omega P\left\{X^{0,N_n}(t_n+s) \neq 0, \text{ for some } s \leq T_n \middle| \text{data to } t_n\right\} \geq \rho. \tag{8.6a}$$

We will show a contradiction to (8.6a). Actually, it is more direct to show that the assertion

$$\lim_n \sup_\omega P\left\{Y^{0,N_n}(t_n+T_n) - Y^{0,N_n}(t_n) = 0 \middle| \text{data to } t_n\right\} \geq \rho. \tag{8.6b}$$

is false. The falsity of (8.6b) implies the falsity of (8.6a).

A Liapunov function argument using (7.1) and the fact that there is no control can be used to prove that

$$\sup_n E|X^{1,N_n}(t_n)|^2 < \infty. \tag{8.7}$$

Now, extract a weakly convergent subsequence of $X^{N_n}(t_n + \cdot)$, and note that its limit $X(\cdot)$ satisfies (3.2). The distribution of $X^1(0)$ depends on the selected convergent subsequence. But, due to (8.7), $E|X^1(0)|^2$ is bounded uniformly in the selected convergent subsequence. Using this last fact, the properties of (3.2) and the weak convergence now imply that (8.6b) cannot hold unless $\rho = 0$. Now note that (a key point) if $X_K^{0,N}(t) = 0$ and $X^{1,N}(t) = X_K^{1,N}(t)$, then the two processes start again at t with equal initial values.

The above results imply the following. For any $\delta > 0$, with a probability arbitrarily close to one, the fraction of time that $|X^{0,N}(t) - X_K^{0,N}(t)| \geq \delta$ on any time interval goes to zero as $K \to \infty$, uniformly in the time interval and in (large) N. This implies that the change in the $k(\cdot)$ part of the cost can be made as small as desired by making K large enough. By construction, $F^{1,N}(t) \geq F_K^{1,N}(t)$, hence the control cost is no greater for the approximating control. We omit the details of the fact that

$$E[F_K^{0,N}(T) - F^{0,N}(T)]/T$$

can be made as small as desired by making K large. But it can be proved by a weak convergence argument and the facts established above.

Thus for large enough K, we lose as little as desired by restricting $F^{1,N}(\cdot)$ such that $F^{1,N}(n+1) - F^{1,N}(n) \leq K$ for all N, n. This completes step 2.

Now, we turn to step 3 and make a few comments concerning the $k(\cdot)$ component of the cost. Given a control $F^{1,N}(\cdot)$ satisfying the restrictions of steps 1 and 2 (with constants B and K, resp.), find a suitable approximation with a bounded "derivative." Let R denote the derivative bound. Define a control $F_R^{1,N}(\cdot)$ with derivative bounded by R, such that $F_R^{1,N}(t) \leq F^{1,N}(t)$, but where the associated process $X_R^{1,N}(\cdot)$ is allowed to catch up with $X^{1,N}(\cdot)$ when possible. It will catch up repeatedly, for large enough R. This is because the maximum number of rejects on any time interval of unit length is $2K/\sqrt{N}$. Since K is bounded and R large, except for an arbitrarily small time subinterval the number of rejects on any time interval $[n, n+1]$ can be made as close as desired to what is needed, uniformly in N, n. Additionally, $|X^{1,N}(t) - X_R^{1,N}(t)|$ is uniformly (in N, t) bounded. Thus, the values of $X^{1,N}(t)$ and $X_R^{1,N}(t)$ will be arbitrarily close when $X_R^{0,N}(t)$ (hence, $X^{0,N}(t)$) hits zero, or very shortly thereafter (at most a time $K/R + O(1/\sqrt{N})$ later).

Note that the approximation problem is more subtle than in step 2, since we cannot guarantee that $X_R^{1,N}(\cdot)$ will equal $X^{1,N}(\cdot)$ on longer and longer intervals. For example, if $F^{1,N}(\cdot)$ jumps periodically, or if the limit is singular with respect to Lebesgue measure.

The following properties can be proved. First, by a Liapunov function argument, it can be shown that

$$\limsup_N \sup_{t,R,F_R^{1,N}} E\left|X_R^{1,N}(t)\right|^2 < \infty.$$

Using this, it can be shown that

$$\limsup_N \sup_{n,R,F_R^{1,N}} E \sup_{n \leq s \leq n+1} \left|X_R^{1,N}(s)\right|^2 < \infty, \tag{8.8}$$

with a similar estimate holding for the $X^{1,N}(\cdot)$. The above comments imply that

$$\lim_R \limsup_N \sup_\tau E \sup_{t \leq T} \int_\tau^{\tau+t} \left|X^{1,N}(s) - X_R^{1,N}(s)\right| ds = 0, \tag{8.9}$$

for any $T < \infty$, and where τ are stopping times. Now use an argument based on recurrence to $X^{0,N}(\cdot)$ to zero analogous to what was done in step 2 to get that the

$k(\cdot)$-costs are close for large R. Obviously, the component of the cost due to $F_R^{1,N}(\cdot)$ is no greater than that due to $F^{1,N}(\cdot)$. Again, by a weak convergence argument, it can be shown that the overflow costs also converge, and the details are omitted.

9. Data. Some typical data is given in Table 1 below. The cost function is $c_0 EX^0(1) + EF^1(1) + 5EF^0(1)$, all stationary values. $B_+^0 = 6.4$, and larger values made little difference. The individual components of the cost function are tabulated, and we write $EF^0(1) = 0$ if it is less than 10^{-4}. The fraction of lost class 1 customers is $EF^1(1)/[\lambda_1\sqrt{N}]$, so it depends on N. The tables indicate the potential tradeoffs between the time gained for class 0 and the lost class 1 customers.

TABLE 1.
Components of optimal cost for limit system.

Table A: $\mu_1 = .5, \mu_0 = 1, \lambda_1 = 1, \lambda_0 = 1, \sigma_1^2 = 1, \sigma_0^2 = 1, b = 2.5$							
	$EX^0(1)$	$EF^1(1)$	$EF^0(1)$	% savings	% Rejection of class 1		
					N=100	N=10^3	N=10^4
no cont.	.555	0	.0032	na	0	0	0
$c_0 = 5$	.267	.489	0	52%	4.89	1.55	.489
$c_0 = 10$	.184	1.01	0	67%	1.01	.319	.101

Table B: $\mu_1 = .25, \mu_0 = .25, \lambda_1 = .5, \lambda_0 = 1, \sigma_1^2 = 1, \sigma_0^2 = 1, b = 2.5$							
	$EX^0(1)$	$EF^1(1)$	$EF^0(1)$	% savings	% Rejection of class 1		
					N=100	N=10^3	N=10^4
no cont.	1.57	0	.002	na	0	0	0
$c_0 = 5$	.424	1.463	0	73%	1.463	.386	.1463
$c_0 = 10$	.299	2.34	0	81%	2.34	.740	.234

Table C: $\mu_1 = 1, \mu_0 = 1, \lambda_1 = 1, \lambda_0 = 1, \sigma_1^2 = 1, \sigma_0^2 = 3, b = 2.5$							
	$EX^0(1)$	$EF^1(1)$	$EF^0(1)$	% savings	% Rejection of class 1		
					N=100	N=10^3	N=10^4
no cont.	.853	0	.004	na	0	0	0
$c_0 = 5$	.557	.750	0	35%	.750	.237	.0750
$c_0 = 10$	.415	1.77	0	51%	1.77	.316	.177

In the above examples and in all other cases that we tested numerically, a considerable saving in the global performance is obtained. The price paid for this saving is the rejection of class 1 customers. However, the fraction of rejected class 1 customers is acceptable for large N. As is seen in the tables, it is of the order of 1% for $N = 1000$, and less than 0.5% for $N = 10000$. We thus conclude that for large systems operating at a heavy traffic regime, we may gain considerably in overall performance of the system at the cost of rejection of a very small fraction of GP calls.

REFERENCES

[1] THE ATM FORUM TECHNICAL COMMITTEE, *Traffic Management Specification*, Version 4.0, af-tm-0056, 1996.

[2] E. ALTMAN, D. ARTIGES, AND K. TRAORE, *On the Integration of Best-Effort and Guaranteed Performance Services*, INRIA Research report RR-3222, INRIA, Le Chesnay cedex, France, 1997.

[3] E. ALTMAN, F. BOCCARA, J. BOLOT, P. NAIN, P. BROWN, D. COLLANGE, AND C. FENZY, *Analysis of the TCP/IP flow control mechanism in high-speed wide-area networks*, European Trans. Telecommunications, 10 (1999), pp. 125–134.

[4] E. Altman, A. Orda, and N. Shimkin, *Bandwidth allocation for guaranteed versus best effort service categories*, in Proceedings of the IEEE Infocom '98, San Fransisco, CA, 1998, pp. 617–624.

[5] J. C. Bolot and A. Vega-Garcma, *Control mechanisms for packet audio in the internet*, in Proceedings of the IEEE Infocom '96, San Francisco, CA, 1996, pp. 232–239.

[6] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[7] H. Chen and W. Whitt, *Diffusion approximations for open queueing networks with service interruption*, Queueing Systems Theory Appl., 13 (1993), pp. 335–359.

[8] P. Dupuis and H. Ishii, *On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications*, Stochastics Stochastics Rep., 35 (1991), pp. 31–62.

[9] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.

[10] A. I. Elwalid and D. Mitra, *Effective bandwidth of general Markovian traffic sources and admission control of high speed networks*, IEEE/ACM Trans. on Networking, 1 (1993), pp. 329–343.

[11] E. A. Feinberg and M. I. Reiman, *Optimality of randomized trunk reservation*, Probab. Engrg. Inform. Sci., 8 (1994), pp. 463–489.

[12] E. Gelenber, X. Mang, and R. Onvural, *Diffusion based statistical call admission in ATM networks*, Performance Evaluation, 27/28 (1996), pp. 411–436.

[13] R. J. Gibbens and P. J. Hunt, *Effective bandwidth for the multi-type UAS channel*, Queueing Systems, 9 (1991), pp. 17–28.

[14] M. Grossglauser and D. Tse, *A framework for robust measurement-based admission control*, IEEE/ACM Trans. Networking, 7 (1999), pp. 293–309.

[15] R. Guérin, H. Ahmadi, and M. Naghshineh, *Equivalent capacity and its application to bandwidth allocation in high-speed networks.* IEEE J. Selected Areas Communication, 9 (1991), pp. 968–981.

[16] A. Hordijk and F. Spieksma, *Constrained admission control to a queuing system*, Adv. Appl. Probab., 21 (1989), pp. 409–431.

[17] V. Jacobson, *Congestion avoidance and control*, ACM SIGCOMM, 88 (1988), pp. 273–288.

[18] F. P. Kelly, *Effective bandwidth at multi-class queues*, Queueing Sytems, 9 (1991), pp. 5–16.

[19] G. Kesidis, J. Walrand, and C.-S. Chang, *Effective bandwidth for multiclass Markov fluids and other ATM sources*, IEEE/ACM Trans. Networking, 1 (1993), pp. 424–428.

[20] H. J. Kushner, *Optimality conditions for the average cost per unit time problem with a diffusion model*, SIAM J. Control Optim., 16 (1978), pp. 330–346.

[21] H. J. Kushner *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.

[22] H. J. Kushner, *Control of trunk line systems in heavy traffic*, SIAM J. Control Optim., 33 (1995), pp. 765–803.

[23] H. J. Kushner, *Heavy traffic analysis of controlled multiplexing systems*, Queueing Systems Theory Appl., 28 (1998), pp. 79–107.

[24] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, Berlin, New York, 1992.

[25] H. J. Kushner, D. Jarvis, and J. Yang, *Controlled and optimally controlled multiplexing systems: A numerical exploration*, Queueing Systems Theory Appl., 20 (1995), pp. 255–291.

[26] H. J. Kushner and L. F. Martins, *Numerical methods for stochastic singular control problems*, SIAM J. Control Optim., 29 (1991), pp. 1443–1475.

[27] H. J. Kushner and L. F. Martins, *Heavy traffic analysis of a data transmission system with many independent sources*, SIAM J. Appl. Math., 53 (1993), pp. 1095–1122.

[28] H. J. Kushner and L. F. Martins, *Heavy traffic analysis of a controlled multiclass queueing network via weak convergence theory*, SIAM J. Control Optim., 34 (1996), pp. 1781–1797.

[29] H. J. Kushner and K. M. Ramachandran, *Optimal and approximately optimal control policies for queues in heavy traffic*, SIAM J. Control Optim., 27 (1989), pp. 1293–1318.

[30] H. J. Kushner and J. Yang, *Numerical methods for controlled routing in large trunk line systems via stochastic control theory*, ORSA J. Computing, 6 (1994), pp. 300–316.

[31] R. Mazumdar, *On call admission control*, 2nd IFIP Workshop on Traffic Management and Synthesis of ATM Networks, Montreal, September 24–26, 1997.

[32] Z. Liu, P. Nain, and D. Towsley, *Exponential bounds with application to call admission*, J. ACM, 44 (1997), pp. 366–394.

[33] A. Mandlebaum and G. Pats, *State-dependent stochastic networks. Part* I: *Approximations and applications with continuous diffusion limits*, Ann. Appl. Prob., 8 (1998), pp. 569–646.

[34] K. W. Ross and D. Tsang, *Optimal Circuit Access Control Policies in an ISDN Environment: A Markov Decision Approach*, IEEE Trans. Commun., 37 (1989), pp. 934–939.

[35] D. Tse and M. Grossglauser, *Measurement-based call admission control: Analysis and simulation*, in Proceedings of IEEE Infocom '97, Kobe, Japan, April 1997.

[36] J. Walrand, *Measurement and control of ATM networks*, in 2nd IFIP Workshop on Traffic Management and Synthesis of ATM Networks, Montreal, September 24–26, 1997.

[37] W. Whitt, *Tail probabilities with statistical multiplexing and effective bandwidth in multi class queues*, Telecommunication Systems, 2 (1993), pp. 71–107.

SIAM J. Control Optim.
Vol. 37, No. 6, pp. 1808–1821

ON CONCEPTS OF CONTROLLABILITY FOR DETERMINISTIC AND STOCHASTIC SYSTEMS*

AGAMIRZA E. BASHIROV[†] AND NAZIM I. MAHMUDOV[†]

Abstract. The new necessary and sufficient conditions, which are formulated in terms of convergence of a certain sequence of operators involving the resolvent of the negative of the controllability operator, are found for deterministic linear stationary control systems to be completely and approximately controllable, respectively. These conditions are applied to study the S-controllability (a property of attaining an arbitrarily small neighborhood of each point in the state space with a probability arbitrarily close to one) and C-controllability (the S-controllability fortified with some uniformity) of stochastic systems. It is shown that the S-controllability (the C-controllability) of a partially observable linear stationary control system with an additive Gaussian white noise disturbance on all the intervals $[0,T]$ for $T>0$ is equivalent to the approximate (complete) controllability of its deterministic part on all the intervals $[0,T]$ for $T>0$.

Key words. complete controllability, approximate controllability, stochastic controllability, deterministic linear system, partially observable linear system

AMS subject classifications. Primary, 93B, 93E; Secondary, 60G

PII. S036301299732184X

1. Introduction.

Theory of controllability originates from the famous work [1] done by Kalman. At present this theory is almost complete for deterministic linear control systems (see, for example, Curtain and Pritchard [2]; Curtain and Zwart [3]; Balakrishnan [4]; Bensoussan et al. [5]; Zabczyk [6]).

The natural extension of the complete and approximate controllability concepts to stochastic control systems is meaningless. In Bashirov [7] and Bashirov and Kerimov [8] these concepts were weakened and, for stochastic control systems, the concept of S-controllability was defined. Briefly, an S-controllable stochastic control system is a system attaining an arbitrarily small neighborhood of each point in the state space with a probability arbitrarily close to one. We also found it useful to define the concept of C-controllability for stochastic control systems as S-controllability fortified with some uniformity.

The main results of [7, 8] concern a partially observable linear stationary control system with an additive Gaussian white noise disturbance (the system (S)) and its deterministic part (the system (D)). From the results of [8] (the necessity part of Theorems 4 and 5(b)), it follows that if the system (S) is C-controllable (S-controllable) on the interval $[0,T]$, where $T>0$ is fixed, then the system (D) is completely (approximately) controllable on the same interval $[0,T]$. A sufficient condition of C-controllability (which is a sufficient condition of S-controllability as well) for the system (S) on the fixed interval $[0,T]$ is also found in [8]. This sufficient condition is based on Lemma 7 in [8] which is proved under the complete controllability condition of the system (D) on all the intervals $[0,t]$ for $0<t\le T$. Thus, more precisely than in [8], this sufficient condition is the complete controllability of the system (D) on all the intervals $[0,t]$ for $0<t\le T$.

*Received by the editors May 27, 1997; accepted for publication (in revised form) September 23, 1998; published electronically October 4, 1999.

http://www.siam.org/journals/sicon/37-6/32184.html

[†]Department of Mathematics, Eastern Mediterranean University, Gazimagusa, Mersin 10, Turkey, and Institute of Cybernetics, Azerbaijan Academy of Sciences, F. Agayev St. 9, Baku 370141, Azerbaijan (aga.as@mozart.emu.edu.tr, nazim.as@mozart.emu.edu.tr).

A discussion of an example presented in [8] leads us to expect that a weaker sufficient condition of S-controllability on the interval $[0, T]$ for the system (S) could exist as the approximate controllability of the system (D) on all the intervals $[0, t]$ for $0 < t \leq T$. This was conjectured in [7, 8], wherein this conjecture is settled positively.

Discussing the S- and C-controllability concepts, we found the new necessary and sufficient conditions for deterministic linear stationary control systems to be completely and approximately controllable. These conditions are formulated in terms of uniform and strong convergence of a certain sequence of operators involving the resolvent of the negative of the controllability operator and clearly distinguish complete and approximate controllabilities.

Studying sources about theory of controllability, we did not find the analogues of these conditions which prompted us to consider them as new. For convenience, we call the above-mentioned conditions the resolvent conditions of complete and approximate controllabilities.

A verification of the resolvent conditions for a concrete control system requires a computation of the respective resolvent and then studying the convergence of the above-mentioned sequence involving this resolvent. This is illustrated in the examples of controlled one-dimensional heat and wave equations. We expect that the resolvent conditions will play a significant role in theoretical investigations of theory of controllability because after a first application, they have allowed us to settle the conjecture mentioned above.

2. General notations. In this paper X and Y are real separable Hilbert spaces. R^k denotes the k-dimensional real Euclidean space. As usual, $R^1 = R$. The closure of the set D is denoted by $\overline{D}$. The space of all linear bounded operators from X to Y is denoted by $\mathcal{L}(X, Y)$. The brief notation $\mathcal{L}(X) = \mathcal{L}(X, X)$ is used as well. A^* denotes the adjoint of the operator A. The trace of the operator A is denoted by $\operatorname{tr} A$. If $A \in \mathcal{L}(X)$ is self-adjoint and $\langle h, Ah \rangle \geq 0$ (respectively, $\langle h, Ah \rangle \geq c\|h\|^2$, where $c = \text{const.} > 0$) for all $h \in X$, then we write $A \geq 0$ (respectively, $A > 0$), where $\langle \cdot, \cdot \rangle$ is an inner product and $\|\cdot\|$ is a norm. For $A \geq 0$, the square root of A is denoted by $A^{1/2}$. The symbol I denotes an identity operator. Zero operator, zero vector, and the number zero are denoted by 0; it is clear which is meant from the context.

It is always supposed that two time moments are given. The initial time moment is identified with zero and is fixed. The terminal moment is denoted by T $(T > 0)$ and is considered variable. The notation $\mathbf{T}$ is used for the finite time interval $[0, T]$. $L_2(\mathbf{T}, X)$ and $L_2(0, T; X)$ denote the space of equivalence classes of all functions from $\mathbf{T} = [0, T]$ to X that are Lebesgue measurable and square integrable with respect to the Lebesgue measure. As usual, we use the brief notation $L_2(0, T) = L_2(0, T; R)$. The notation $\Delta = \{(t, s) : 0 \leq s \leq t \leq T\}$ is used for the triangular set over $\mathbf{T}$. $B_2(\Delta, \mathcal{L}(X, Y))$ denotes the class of all $\mathcal{L}(X, Y)$-valued functions on Δ that are strongly measurable and square integrable with respect to the Lebesgue measure on Δ (see, for example, [2, 3]).

All integrals of vector-valued functions are considered in the Bochner sense. For probability, expectation, and conditional expectation, the notations $\mathbf{P}$, $\mathbf{E}$, and $\mathbf{E}(\cdot \,|\, \cdot)$, respectively, are used. $\operatorname{cov}(x, y)$ is the covariance operator of the random variables x and y. The brief notation $\operatorname{cov} x = \operatorname{cov}(x, x)$ is used as well. The integrals of operator-valued functions (except stochastic integrals) are in the strong Bochner sense.

3. Main definitions. Consider a control system on $\mathbf{T}$. Let x_t^u be its (random or not) state value at time $t \in \mathbf{T}$ corresponding to the control u taken from the set of the admissible controls U. If the control system under consideration is stochastic, then by $\mathcal{F}^u$ we denote the smallest σ-algebra generated by the observations on the time interval $\mathbf{T}$ corresponding to the control u. Suppose that X is the state space. For $0 \le \varepsilon < \infty$ and for $0 \le p \le 1$, introduce the sets

$$D = \{x_T^u : u \in U\}, \tag{1}$$

$$S(\varepsilon, p) = \big\{h \in X : \exists u \in U\ \mathbf{P}\big(\|\mathbf{E}(x_T^u|\mathcal{F}^u) - h\|^2 > \varepsilon\big) \le 1 - p\big\}, \tag{2}$$

$$C(\varepsilon, p) = \big\{h \in X : \exists u \in U\ h = \mathbf{E}x_T^u,\ \mathbf{P}\big(\|\mathbf{E}(x_T^u|\mathcal{F}^u) - h\|^2 > \varepsilon\big) \le 1 - p\big\}. \tag{3}$$

Definition 1. *A deterministic control system will be called*

(a) D^c-controllable on $\mathbf{T}$ *if* $D = X$;

(b) D^a-controllable on $\mathbf{T}$ *if* $\overline{D} = X$.

It is clear that the D^c- and D^a-controllabilities are the well-known complete and approximate controllabilities for deterministic control systems, respectively. Originally, the D^c-controllability was introduced in Kalman [1] as a concept for finite dimensional deterministic control systems, so the natural extension of this concept is too strong for many infinite dimensional control systems. Therefore, the D^a-controllability was introduced as a weakened version of the D^c-controllability. It is also clear that neither D^c- nor D^a-controllabilities can be a property of stochastic control systems, so there is a need to further weaken these concepts in order to extend them to stochastic control systems.

The following definition will be used as a step in discussing the main concepts of controllability for stochastic systems.

Definition 2. *Given* $\varepsilon \ge 0$ *and* $0 \le p \le 1$, *a control system will be called*

(a) $S^c_{\varepsilon,p}$-controllable on $\mathbf{T}$ *if* $S(\varepsilon, p) = X$;

(b) $S^a_{\varepsilon,p}$-controllable on $\mathbf{T}$ *if* $\overline{S(\varepsilon, p)} = X$;

(c) $C^c_{\varepsilon,p}$-controllable on $\mathbf{T}$ *if* $C(\varepsilon, p) = X$;

(d) $C^a_{\varepsilon,p}$-controllable on $\mathbf{T}$ *if* $\overline{C(\varepsilon, p)} = X$;

(e) $S^0_{\varepsilon,p}$-controllable on $\mathbf{T}$ *if* $0 \in S(\varepsilon, p)$.

The geometric interpretation of the $S^c_{\varepsilon,p}$-controllability ($S^a_{\varepsilon,p}$-controllability) is as follows. If a control system with the initial state x_0 is $S^c_{\varepsilon,p}$-controllable ($S^a_{\varepsilon,p}$-controllable) on $\mathbf{T}$, then with probability not less than p it can pass from x_0 for the time T into the $\sqrt{\varepsilon}$-neighborhood of an arbitrary point in the state space (in a set that is dense in the state space). The geometric interpretation of the $C^c_{\varepsilon,p}$- and $C^a_{\varepsilon,p}$-controllabilities differs from the same of the $S^c_{\varepsilon,p}$- and $S^a_{\varepsilon,p}$-controllabilities since among the controls, with the help of which the $\sqrt{\varepsilon}$-neighborhood of any point h is achieved, there exists one with a property that the expectation of the target state, corresponding to this control, coincides with h. Obviously, a $C^c_{\varepsilon,p}$-controllable ($C^a_{\varepsilon,p}$-controllable) control system is $S^c_{\varepsilon,p}$-controllable ($S^a_{\varepsilon,p}$-controllable), but the converse is not true.

The smaller ε is and the larger p is for a control system, the more controllable it is; i.e., it is possible to hit into a smaller neighborhood with a higher probability. One can observe that all control systems are $S^c_{\varepsilon,p}$-, $S^a_{\varepsilon,p}$-, $C^c_{\varepsilon,p}$-, and $C^a_{\varepsilon,p}$-controllable on any interval with $\varepsilon \ge 0$ and $p = 0$ or $\varepsilon = \infty$ and $0 \le p \le 1$, if we admit ∞ as a value for ε. At the same time, it is clear that a D^c-controllable (D^a-controllable) deterministic system is $S^c_{0,1}$- and $C^c_{0,1}$-controllable ($S^a_{0,1}$- and $C^a_{0,1}$-controllable) with parameters $\varepsilon = 0$ and $p = 1$ since, for deterministic systems, $D = S(0, 1) = C(0, 1)$.

Also, each kind of controllability from Definition 2 with a smaller ε and a greater p implies the same kind of controllability with a greater ε and a smaller p.

Summarizing, we can give the following easy necessary and sufficient conditions for the D^c- and D^a-controllabilities.

PROPOSITION 1. *For a deterministic control system the following three conditions are equivalent*:

(a) D^c*-controllability on* $\mathbf{T}$;

(b) $S^c_{\varepsilon,p}$*-controllability on* $\mathbf{T}$ *for all* $\varepsilon \geq 0$ *and for all* $0 \leq p \leq 1$;

(c) $C^c_{\varepsilon,p}$*-controllability on* $\mathbf{T}$ *for all* $\varepsilon \geq 0$ *and for all* $0 \leq p \leq 1$.

PROPOSITION 2. *For a deterministic control system the following three conditions are equivalent*:

(a) D^a*-controllability on* $\mathbf{T}$;

(b) $S^a_{\varepsilon,p}$*-controllability on* $\mathbf{T}$ *for all* $\varepsilon \geq 0$ *and for all* $0 \leq p \leq 1$;

(c) $C^a_{\varepsilon,p}$*-controllability on* $\mathbf{T}$ *for all* $\varepsilon \geq 0$ *and for all* $0 \leq p \leq 1$.

Excepting the limit values $\varepsilon = 0$ and $p = 1$ from the above-mentioned necessary and sufficient conditions for the D^c- and D^a-controllabilities, one can obtain the weakened versions of these concepts. For a moment call a given stochastic system

(a) S^c*-controllable on* $\mathbf{T}$ if it is $S^c_{\varepsilon,p}$-controllable on $\mathbf{T}$ for all $\varepsilon > 0$ and for all $0 \leq p < 1$;

(b) S^a*-controllable on* $\mathbf{T}$ if it is $S^a_{\varepsilon,p}$-controllable on $\mathbf{T}$ for all $\varepsilon > 0$ and for all $0 \leq p < 1$;

(c) C^c*-controllable on* $\mathbf{T}$ if it is $C^c_{\varepsilon,p}$-controllable on $\mathbf{T}$ for all $\varepsilon > 0$ and for all $0 \leq p < 1$;

(d) C^a*-controllable on* $\mathbf{T}$ if it is $C^a_{\varepsilon,p}$-controllable on $\mathbf{T}$ for all $\varepsilon > 0$ and for all $0 \leq p < 1$.

In [7, 8] it is shown that the concepts of S^c- and S^a-controllabilities are equivalent. It will also be shown that for partially observable linear stationary control systems with additive Gaussian white noise disturbance the C^a-controllability on all the intervals $[0, T]$ with $T > 0$ is equivalent to the S^c-and S^a-controllabilities on all the intervals $[0, T]$ with $T > 0$. Thus, we can define two basic and one additional concepts of controllability for stochastic systems.

DEFINITION 3. *A control system will be called*

(a) S-controllable on $\mathbf{T}$ *if it is* $S^c_{\varepsilon,p}$*-controllable on* $\mathbf{T}$ *or, equivalently,* $S^a_{\varepsilon,p}$*-controllable on* $\mathbf{T}$ *for all* $\varepsilon > 0$ *and for all* $0 \leq p < 1$;

(b) C-controllable on $\mathbf{T}$ *if it is* $C^c_{\varepsilon,p}$*-controllable on* $\mathbf{T}$ *for all* $\varepsilon > 0$ *and for all* $0 \leq p < 1$;

(c) S^0-controllable on $\mathbf{T}$ *if it is* $S^0_{\varepsilon,p}$*-controllable on* $\mathbf{T}$ *for all* $\varepsilon > 0$ *and for all* $0 \leq p < 1$.

Geometrically, the S-controllability can be interpreted as follows: an S-controllable on $\mathbf{T}$ control system can attain for the time T an arbitrarily small neighborhood of each point in the state space with a probability arbitrarily close to one. The C-controllability is the S-controllability fortified with some uniformity. The S^0-controllability is useful in discussing S- and C-controllabilities.

Finally, notice that the abbreviations D, S, C, c, and a in the previously introduced controllability concepts mean deterministic, stochastic, combined, complete, and approximate, respectively.

4. Preliminaries.

In this paper it is always supposed that A is the infinitesimal generator of a strongly continuous semigroup $\mathcal{U}$, $B \in \mathcal{L}(Y, X)$, $C \in \mathcal{L}\big(X, R^k\big)$; x_0 is a Gaussian random variable with $\operatorname{cov} x_0 = P_0$; m and n are X- and R^k-valued Wiener processes, respectively; $n_0 = 0$, $m_0 = 0$, $\mathbf{E}n_t = 0$, $\mathbf{E}m_t = 0$, $\operatorname{cov} n_t = It$, $\operatorname{cov} m_t = Mt$, M is a nuclear operator on X; and x_0, n, m are mutually independent. Let $f \in L_2(\mathbf{T}, X)$ and consider the linear partially observable stochastic control system

$$\begin{cases} dx_t^u = (Ax_t^u + Bu_t + f_t)dt + dm_t, \ 0 < t \le T, \ x_0^u = x_0, \\ d\xi_t^u = Cx_t^u dt + dn_t, \ 0 < t \le T, \ \xi_0^u = 0, \end{cases} \tag{4}$$

where x, u, and ξ are state, control, and observation processes. Under the set U of admissible controls we consider the set of all controls in the linear feedback form

$$u_t = \bar{u}_t + \int_0^t K_{t,s}\, d\xi_s^u,$$

where $K \in B_2\big(\Delta, \mathcal{L}(R^k, Y)\big)$ and $\bar{u} \in L_2(\mathbf{T}, Y)$.

To the system (4) one can associate two control systems. The first is the deterministic control system

$$\frac{d}{dt} y_t^v = Ay_t^v + Bv_t + f_t, \ 0 < t \le T, \ y_0^v = y_0 = \mathbf{E}x_0, \tag{5}$$

where v is a control in $V = L_2(\mathbf{T}, Y)$. The second is the partially observable stochastic control system

$$\begin{cases} dz_t^w = (Az_t^w + Bw_t)dt + dm_t, \ 0 < t \le T, \ z_0^w = z_0 = x_0 - \mathbf{E}x_0, \\ d\eta_t^w = Cz_t^w dt + dn_t, \ 0 < t \le T, \ \eta_0^w = 0, \end{cases} \tag{6}$$

where w is a control in W consisting of all controls in the linear feedback form

$$w_t = \int_0^t K_{t,s}\, d\eta_s^w,$$

where $K \in B_2\big(\Delta, \mathcal{L}(R^k, Y)\big)$.

Note that solutions of the equations in (4), (5), and (6) are meant in the mild sense, i.e.,

$$\begin{aligned} x_t^u &= \mathcal{U}_t x_0 + \int_0^t \mathcal{U}_{t-s}(Bu_s + f_s)\, ds + \int_0^t \mathcal{U}_{t-s}\, dm_s, \ 0 \le t \le T, \\ y_t^v &= \mathcal{U}_t y_0 + \int_0^t \mathcal{U}_{t-s}(Bv_s + f_s)\, ds, \ 0 \le t \le T, \\ z_t^w &= \mathcal{U}_t z_0 + \int_0^t \mathcal{U}_{t-s} Bw_s\, ds + \int_0^t \mathcal{U}_{t-s}\, dm_s, \ 0 \le t \le T. \end{aligned}$$

Denote

$$\Gamma_{T-t} = \int_t^T \mathcal{U}_{T-s} BB^* \mathcal{U}_{T-s}^*\, ds, \ 0 \le t \le T. \tag{7}$$

For $0 \le t < T$, the operator Γ_{T-t} is called a controllability operator. One can see that $\Gamma_{T-t} \ge 0$ and, hence, the resolvent $R(\lambda, -\Gamma_{T-t}) = (\lambda I + \Gamma_{T-t})^{-1}$ is well defined for all $\lambda > 0$ and for all $0 \le t \le T$. If $\Gamma_{T-t} > 0$, then $R(\lambda, -\Gamma_{T-t})$ is defined for $\lambda = 0$ as well.

We will use the following operator Riccati equations:

$$\frac{d}{dt} Q_t + Q_t A + A^* Q_t - \lambda^{-1} Q_t BB^* Q_t = 0, \ 0 \le t < T, \ Q_T = I, \ \lambda > 0, \tag{8}$$

$$\frac{d}{dt}P_t - AP_t - P_tA^* - M + P_tC^*CP_t = 0,\ 0 < t \leq T,\ P_0 = \operatorname{cov} z_0. \tag{9}$$

LEMMA 1. *There exist the unique strongly continuous solutions* (*in scalar product sense*) Q^λ *and* P *of* (8) *and* (9), *respectively, satisfying* $Q_t^\lambda \geq 0$ *and* $P_t \geq 0$ *for all* $t \in \mathbf{T}$. *Moreover, the solution of* (8) *has the explicit form*

$$Q_t^\lambda = \lambda \mathcal{U}_{T-t}^* R(\lambda, -\Gamma_{T-t})\mathcal{U}_{T-t},\ 0 \leq t \leq T,\ \lambda > 0. \tag{10}$$

For the proof of existence and uniqueness part of this lemma, see [2]. For the proof of representation (10), see [8, 9].

Consider the linear regulator problem consisting of minimizing the cost functional

$$J(v) = \|y_T^v - h\|^2 + \lambda \int_0^T \|v_t\|^2\, dt, \tag{11}$$

where y^v is a state process, defined by (5); v is a control in $V = L_2(\mathbf{T}, Y)$; and $h \in X$ and $\lambda > 0$ are parameters.

LEMMA 2. *For given* $h \in X$ *and* $\lambda > 0$, *there exists a unique optimal control* v^λ *in* $L_2(\mathbf{T}, Y)$ *at which the functional* (11) *takes on its minimum value and*

$$v_t^\lambda = -B^*\mathcal{U}_{T-t}^* R(\lambda, -\Gamma_T)(\mathcal{U}_T y_0 - h + g) \textit{ almost everywhere (a.e.) on } \mathbf{T}, \tag{12}$$

$$y_T^{v^\lambda} - h = \lambda R(\lambda, -\Gamma_T)(\mathcal{U}_T y_0 - h + g), \tag{13}$$

where

$$g = \int_0^T \mathcal{U}_{T-t} f_t\, dt.$$

Proof. The existence and the uniqueness of an optimal control follows from a general theorem about linear regulator problems (see [2]). We will prove the formulae (12) and (13). By computing the variation of the functional (11), one can easily obtain

$$v_t^\lambda = -\lambda^{-1} B^* \mathcal{U}_{T-t}^* \left(y_T^{v^\lambda} - h\right) \text{ a.e. on } \mathbf{T}. \tag{14}$$

Using this in (5), we have

$$\begin{aligned} y_T^{v^\lambda} &= \mathcal{U}_T y_0 + \int_0^T \mathcal{U}_{T-t}\left(Bv_t^\lambda + f_t\right) dt \\ &= \mathcal{U}_T y_0 + g - \lambda^{-1} \int_0^T \mathcal{U}_{T-t} BB^* \mathcal{U}_{T-t}^* \left(y_T^{v^\lambda} - h\right) dt \\ &= \mathcal{U}_T y_0 + g - \lambda^{-1} \Gamma_T \left(y_T^{v^\lambda} - h\right). \end{aligned}$$

Hence,

$$\lambda y_T^{v^\lambda} = \lambda(\mathcal{U}_T y_0 + g) - \Gamma_T\left(y_T^{v^\lambda} - h\right),$$

which implies

$$(\lambda I + \Gamma_T) y_T^{v^\lambda} = \lambda(\mathcal{U}_T y_0 + g) + \Gamma_T h$$

and, consequently,

$$\begin{aligned} y_T^{v^\lambda} &= \lambda(\lambda I + \Gamma_T)^{-1}(\mathcal{U}_T y_0 + g) + (\lambda I + \Gamma_T)^{-1}(\lambda I + \Gamma_T - \lambda I)h \\ &= \lambda R(\lambda, -\Gamma_T)(\mathcal{U}_T y_0 + g - h) + h. \end{aligned}$$

Thus, (13) holds. Substituting (13) in (14), we obtain (12). The lemma is proven.

5. Resolvent conditions for D^c- and D^a-controllabilities. In this section, the necessary and sufficient conditions of D^c - and D^a -controllabilities are discussed.

THEOREM 1. *The following statements are equivalent*:

(1a) *the control system* (5) *is D^c-controllable on* $\mathbf{T}$;

(1b) *the complete controllability condition for the system* (5) *holds, i.e.,* $\Gamma_T > 0$;

(1c) $R(\lambda, -\Gamma_T)$ *converges as* $\lambda \to 0$ *in uniform operator topology*;

(1d) $R(\lambda, -\Gamma_T)$ *converges as* $\lambda \to 0$ *in strong operator topology*;

(1e) $R(\lambda, -\Gamma_T)$ *converges as* $\lambda \to 0$ *in weak operator topology*;

(1f) $\lambda R(\lambda, -\Gamma_T)$ *converges to zero operator as* $\lambda \to 0$ *in uniform operator topology.*

Proof. The equivalence (1a) ⇔ (1b) is well known. For the implication (1b) ⇒ (1c), suppose $\Gamma_T > 0$. Then for all $x \in X$ and for all $\lambda \geq 0$,

$$\langle x, (\lambda I + \Gamma_T)x\rangle \geq (\lambda + k)\|x\|^2,$$

where $k > 0$ is a constant. Therefore, for all $\lambda \geq 0$,

$$\|R(\lambda, -\Gamma_T)\| = \left\|(\lambda I + \Gamma_T)^{-1}\right\| \leq \frac{1}{\lambda + k} \leq \frac{1}{k}.$$

We obtain that $\|R(\lambda, -\Gamma_T)\|$ is bounded with respect to $\lambda \geq 0$. Furthermore,

$$\begin{aligned} \left\|R(\lambda, -\Gamma_T) - \Gamma_T^{-1}\right\| &= \left\|(\lambda I + \Gamma_T)^{-1} - \Gamma_T^{-1}\right\| \\ &= \left\|\Gamma_T^{-1}(\Gamma_T - \lambda I - \Gamma_T)(\lambda I + \Gamma_T)^{-1}\right\| \\ &\leq \lambda\left\|\Gamma_T^{-1}\right\| \left\|(\lambda I + \Gamma_T)^{-1}\right\| \\ &\leq \lambda k^{-2}. \end{aligned}$$

Thus, $R(\lambda, -\Gamma_T)$ converges to Γ_T^{-1} as $\lambda \to 0$ in uniform operator topology. The implications (1c) ⇒ (1d) ⇒ (1e) are obvious. The implication (1e) ⇒ (1f) follows from the boundedness of a weakly convergent sequence of operators. For the implication (1f) ⇒ (1b), suppose

$$\lambda\|R(\lambda, -\Gamma_T)\| = \lambda \left\|(\lambda I + \Gamma_T)^{-1}\right\| \to 0, \ \lambda \to 0.$$

Then $\lambda^{1/2} \left\|(\lambda I + \Gamma_T)^{-1/2}\right\| \to 0$ as $\lambda \to 0$. For sufficiently small $\lambda_0 > 0$, we can write

$$\lambda_0^{1/2} \left\|(\lambda_0 I + \Gamma_T)^{-1/2}\right\| \leq 1/\sqrt{2}.$$

Thus, for all $x \in X$ we have

$$\begin{aligned} \|x\|^2 &= \left\|\left(\lambda_0^{1/2}(\lambda_0 I + \Gamma_T)^{-1/2}\right)\left(\lambda_0^{-1/2}(\lambda_0 I + \Gamma_T)^{1/2}\right)x\right\|^2 \\ &\leq \frac{1}{2}\left\|\lambda_0^{-1/2}(\lambda_0 I + \Gamma_T)^{1/2}x\right\|^2 \\ &= \frac{1}{2}\left\langle \lambda_0^{-1}(\lambda_0 I + \Gamma_T)x, x\right\rangle, \end{aligned}$$

which implies

$$\left\langle \lambda_0^{-1}(\lambda_0 I + \Gamma_T)x, x \right\rangle \geq 2\|x\|^2$$

and, consequently,

$$\langle \Gamma_T x, x \rangle \geq \lambda_0 \|x\|^2.$$

Thus, $\Gamma_T > 0$. The theorem is proven.

THEOREM 2. *The following statements are equivalent*:

(2a) *the control system* (5) *is* D^a*-controllable on* $\mathbf{T}$;

(2b) *the approximate controllability condition for the system* (5) *holds, i.e., if* $B^*\mathcal{U}_t^* x = 0$ *for all* $t \in \mathbf{T}$*, then* $x = 0$;

(2c) $\lambda R(\lambda, -\Gamma_T)$ *converges to zero operator as* $\lambda \to 0$ *in strong operator topology*;

(2d) $\lambda R(\lambda, -\Gamma_T)$ *converges to zero operator as* $\lambda \to 0$ *in weak operator topology.*

Proof. The equivalence (2a) ⇔ (2b) is well known. For the implication (2c) ⇒ (2a), suppose $\lambda R(\lambda, -\Gamma_T) \to 0$ as $\lambda \to 0$ in strong operator topology. Consider arbitrary $h \in X$ and the functional (11) with this h. By (13), selecting λ sufficiently small, we can make $y_T^{v^\lambda}$ close to h, so the control system (5) is D^a-controllable. For the implication (2a) ⇒ (2c), let the control system (5) be D^a-controllable. Then for arbitrary $h \in X$, there exists a sequence $\{\bar{v}^i\}$ in $L_2(\mathbf{T}, Y)$ such that $\left\|y_T^{\bar{v}^i} - h\right\| \to 0$ as $i \to \infty$. We have

$$\left\|y_T^{v^\lambda} - h\right\|^2 \leq \left\|y_T^{v^\lambda} - h\right\|^2 + \lambda \int_0^T \left\|v_t^\lambda\right\|^2 dt \leq \left\|y_T^{\bar{v}^i} - h\right\|^2 + \lambda \int_0^T \left\|\bar{v}_t^i\right\|^2 dt,$$

where v^λ is the control at which the functional (11) takes on its minimum value. If $\varepsilon > 0$ is given, then we can make $\|y_T^{\bar{v}^i} - h\| < \varepsilon/\sqrt{2}$ for some sufficiently large i and then we can select $\delta > 0$ to be sufficiently small so that for all $0 < \lambda < \delta$,

$$\lambda \int_0^T \left\|\bar{v}_t^i\right\|^2 dt < \frac{\varepsilon^2}{2}.$$

Thus, $\left\|y_T^{v^\lambda} - h\right\| < \varepsilon$ for all $0 < \lambda < \delta$. By (13) and the arbitrariness of h, the convergence of $\lambda R(\lambda, -\Gamma_T)$ to zero operator is implied as $\lambda \to 0$ in strong operator topology. Finally, the equivalence (2c) ⇔ (2d) is a consequence of $\lambda R(\lambda, -\Gamma_T) \geq 0$. The theorem is proven.

The conditions (1f) and (2c) in Theorems 1 and 2 clearly distinguish the D^c- and D^a-controllabilities of the control system (5) showing that the distinction between them is in a kind of convergence of $\lambda R(\lambda, -\Gamma_T)$ to zero operator as $\lambda \to 0$. We call these conditions the resolvent conditions for the control system (5) to be D^c- and D^a-controllable, respectively.

An application of the resolvent conditions to a concrete system requires a computation of the respective resolvent and then a verification of the respective convergence. These are illustrated below in the examples of controlled one-dimensional heat and wave equations.

Example 1. Consider a controlled one-dimensional heat equation

$$\frac{\partial}{\partial t} y_{t,\theta} = \frac{\partial^2}{\partial \theta^2} y_{t,\theta} + v_{t,\theta},\ 0 \leq \theta \leq 1,\ 0 < t \leq T, \tag{15}$$

with the initial and boundary conditions

$$y_{0,\theta} = y_{t,0} = y_{t,1} = 0,\ 0 \leq \theta \leq 1,\ 0 \leq t \leq T. \tag{16}$$

Let $X = L_2(0,1)$. In the system (15)–(16), the second-order differential operator $d^2/d\theta^2$ stands for the operator A with the domain

$$D(A) = \bigl\{h \in X : \bigl(d^2/d\theta^2\bigr)h \in X,\ h_0 = h_1 = 0\bigr\},$$

and it generates the strongly continuous semigroup $\mathcal{U}$ defined by

$$[\mathcal{U}_t h]_\theta = \sum_{i=1}^{\infty} 2e^{-i^2\pi^2 t}\sin(i\pi\theta)\int_0^1 h_\alpha \sin(i\pi\alpha)\,d\alpha,\ 0 \le \theta \le 1,\ t \ge 0,\ h \in X.$$

If v is considered as a control action taken from the set of admissible controls $V = L_2(\mathbf{T}, X)$, then it is easily shown that $B = B^* = I$ and, since $\mathcal{U}_t$ is self-adjoint,

$$\Gamma_T = \int_0^T \mathcal{U}_{T-s}BB^*\mathcal{U}_{T-s}^*\,ds = \int_0^T \mathcal{U}_{2s}\,ds.$$

Therefore, for $h \in X$,

$$\begin{aligned}
[\Gamma_T h]_\theta &= \left[\int_0^T \mathcal{U}_{2s}h\,ds\right]_\theta \\
&= \sum_{i=1}^{\infty}\int_0^T 2e^{-2i^2\pi^2 s}\sin(i\pi\theta)\int_0^1 h_\alpha \sin(i\pi\alpha)\,d\alpha\,ds \\
&= \sum_{i=1}^{\infty}\frac{1-e^{-2i^2\pi^2 T}}{i^2\pi^2}\sin(i\pi\theta)\int_0^1 h_\alpha\sin(i\pi\alpha)\,d\alpha.
\end{aligned}$$

The half-range Fourier sine expansion of $h \in X$ is

$$h_\theta = \sum_{i=1}^{\infty} 2\sin(i\pi\theta)\int_0^1 h_\alpha\sin(i\pi\alpha)\,d\alpha,\ 0 \le \theta \le 1.$$

Using this, we obtain

$$[(\lambda I + \Gamma_T)h]_\theta = \sum_{i=1}^{\infty}\frac{2i^2\pi^2\lambda + 1 - e^{-2i^2\pi^2 T}}{i^2\pi^2}\sin(i\pi\theta)\int_0^1 h_\alpha\sin(i\pi\alpha)\,d\alpha.$$

Let $(\lambda I + \Gamma_T)h = g$. If we use the half-range Fourier sine expansion of $g \in X$, then

$$\begin{aligned}
&\sum_{i=1}^{\infty}\frac{2i^2\pi^2\lambda + 1 - e^{-2i^2\pi^2 T}}{i^2\pi^2}\sin(i\pi\theta)\int_0^1 h_\alpha\sin(i\pi\alpha)\,d\alpha \\
&\quad = \sum_{i=1}^{\infty} 2\sin(i\pi\theta)\int_0^1 g_\alpha\sin(i\pi\alpha)\,d\alpha,
\end{aligned}$$

which implies

$$\begin{aligned}
&\int_0^1 h_\alpha\sin(i\pi\alpha)\,d\alpha \\
&= \frac{2i^2\pi^2}{2i^2\pi^2\lambda + 1 - e^{-2i^2\pi^2 T}}\int_0^1 g_\alpha\sin(i\pi\alpha)\,d\alpha,\ i = 1, 2, \ldots.
\end{aligned}$$

Therefore,

$$\begin{aligned} h_\theta &= \left[(\lambda I + \Gamma_T)^{-1} g\right]_\theta = [R(\lambda, -\Gamma_T) g]_\theta \\ &= \sum_{i=1}^{\infty} \frac{4i^2\pi^2}{2i^2\pi^2\lambda + 1 - e^{-2i^2\pi^2 T}} \sin(i\pi\theta) \int_0^1 g_\alpha \sin(i\pi\alpha)\, d\alpha. \end{aligned}$$

If $g_\alpha \equiv 1$, then by Parseval's identity,

$$\begin{aligned} \|R(\lambda, -\Gamma_T) g\|_X^2 &= \frac{1}{2} \sum_{i=1}^{\infty} \frac{(4i^2\pi^2)^2}{(2i^2\pi^2\lambda + 1 - e^{-2i^2\pi^2 T})^2} \left(\int_0^1 \sin(i\pi\alpha)\, d\alpha \right)^2 \\ &= \sum_{i=1}^{\infty} \frac{8i^2\pi^2 (1 - (-1)^i)^2}{(2i^2\pi^2\lambda + 1 - e^{-2i^2\pi^2 T})^2} \\ &\geq \sum_{i=1}^{\infty} \frac{8i^2\pi^2 (1 - (-1)^i)^2}{(2i^2\pi^2\lambda + 1)^2} = \sum_{i=1,3,5,\ldots} \frac{32 i^2\pi^2}{(2i^2\pi^2\lambda + 1)^2}. \end{aligned}$$

One can verify that the inequality

$$\frac{i}{2i^2\pi^2\lambda + 1} > \frac{i+1}{2(i+1)^2\pi^2\lambda + 1}$$

holds whenever i is an integer that is greater than the number $1/\sqrt{2\lambda}\pi$. Let N_λ be the smallest odd integer that is greater than $1/\sqrt{2\lambda}\pi$. Then the sequence

$$\left\{ i^2\pi^2 / \left(2i^2\pi^2\lambda + 1\right)^2 \right\}_{i=1,2,\ldots}$$

is decreasing for $i \geq N_\lambda$. The following limits are obvious:

$$N_\lambda \to \infty \text{ and } \lambda N_\lambda^2 \to \frac{1}{2\pi^2} \text{ as } \lambda \to 0.$$

Using these, for $g_\alpha \equiv 1$, we obtain

$$\begin{aligned} \|R(\lambda, -\Gamma_T) g\|_X^2 &\geq \sum_{i=N_\lambda}^{\infty} \frac{16 i^2\pi^2}{(2i^2\pi^2\lambda + 1)^2} \geq \int_{N_\lambda}^{\infty} \frac{16\pi^2 t^2}{(2\pi^2\lambda t^2 + 1)^2}\, dt \\ &\geq \int_{N_\lambda}^{\infty} \frac{4\pi^2 t}{(2\pi^2\lambda t^2 + 1)^2}\, dt = \frac{1}{\lambda(2\pi^2\lambda N_\lambda^2 + 1)} \to \infty \end{aligned}$$

as $\lambda \to 0$. Therefore, by (1a) $\Leftrightarrow$ (1d) in Theorem 1, the system (15)–(16) is not D^c-controllable. At the same time, for all $g \in X$,

$$\begin{aligned} &\|\lambda R(\lambda, -\Gamma_T) g\|_X^2 \\ &= \sum_{i=1}^{\infty} \frac{8 i^4\pi^4\lambda^2}{(2i^2\pi^2\lambda + 1 - e^{-2i^2\pi^2 T})^2} \left(\int_0^1 g_\alpha \sin(i\pi\alpha)\, d\alpha \right)^2 \to 0 \end{aligned}$$

as $\lambda \to 0$ and hence, by (2a) $\Leftrightarrow$ (2c) in Theorem 2, the system (15)–(16) is D^a-controllable.

Example 2. Consider a controlled wave equation

$$\text{(17)} \qquad \frac{\partial^2}{\partial t^2} \xi_{t,\theta} = \frac{\partial^2}{\partial \theta^2} \xi_{t,\theta} + b_\theta v_t, \ 0 \leq \theta \leq 1, \ 0 < t \leq T,$$

with the initial and boundary conditions

$$\xi_{0,\theta} = f_\theta,\ \frac{\partial}{\partial t}\xi_{t,\theta}\Big|_{t=0} = g_\theta,\ \xi_{t,0} = \xi_{t,1} = 0,\ 0 \le \theta \le 1,\ 0 \le t \le T, \tag{18}$$

where v is a control action taken from the set of admissible controls $V = L_2(0,T)$, i.e., $Y = R$. We assume that f, g, and b are functions in $L_2(0,1)$. For these functions, we will use the half-range Fourier sine expansions

$$f_\theta = \sum_{i=1}^{\infty} \alpha_i \sin(i\pi\theta),\ g_\theta = \sum_{i=1}^{\infty} \beta_i \sin(i\pi\theta),\ b_\theta = \sum_{i=1}^{\infty} \gamma_i \sin(i\pi\theta)$$

and suppose that

$$\sum_{i=1}^{\infty} i^2\alpha_i^2 < \infty.$$

Let X be a Hilbert space of all functions

$$h = \begin{bmatrix} f \\ g \end{bmatrix} : [0,1] \to R,$$

where f and g satisfy the above-mentioned conditions endowed with the scalar product

$$\langle h, \tilde{h}\rangle = \left\langle \begin{bmatrix} f \\ g \end{bmatrix}, \begin{bmatrix} \tilde{f} \\ \tilde{g} \end{bmatrix} \right\rangle = \sum_{i=1}^{\infty} \left(i^2\pi^2\alpha_i\tilde{\alpha}_i + \beta_i\tilde{\beta}_i \right),$$

where $\tilde{\alpha}_i$ and $\tilde{\beta}_i$ are the respective Fourier coefficients of $\tilde{f}$ and $\tilde{g}$. This space X is suitable for the problem (17)–(18) (see Curtain and Zwart [3, p. 149] and Zabczyk [6, p. 180]). For the operator

$$A = \begin{bmatrix} 0 & I \\ d^2/d\theta^2 & 0 \end{bmatrix}, \tag{19}$$

where I is the identity operator on $L_2(0,1)$ and $d^2/d\theta^2$ has the domain

$$D\big(d^2/d\theta^2\big) = \big\{\eta \in L_2(0,1) : \big(d^2/d\theta^2\big)\eta \in L_2(0,1),\ \eta_0 = \eta_1 = 0\big\},$$

and for $B \in \mathcal{L}(R, X)$ defined by

$$[Bv]_\theta = \begin{bmatrix} 0 \\ b_\theta v \end{bmatrix},\ 0 \le \theta \le 1,\ v \in R,$$

the problem (17)–(18) can be formulated in the abstract form

$$\frac{d}{dt}y_t = Ay_t + Bv_t,\ t > 0, \tag{20}$$

where

$$[y_t]_\theta = \begin{bmatrix} \xi_{t,\theta} \\ (\partial/\partial t)\xi_{t,\theta} \end{bmatrix}, 0 \le \theta \le 1,\ 0 < t \le T;\ y_0 = \begin{bmatrix} f \\ g \end{bmatrix}.$$

It is known that the operator A defined by (19) generates a continuous group $\mathcal{U}$ (see Curtain and Zwart [3] and Zabczyk [6]) as defined by

$$[\mathcal{U}_t h]_\theta = \sum_{i=1}^{\infty} \begin{bmatrix} \cos(i\pi t) & (i\pi)^{-1}\sin(i\pi t) \\ -i\pi\sin(i\pi t) & \cos(i\pi t) \end{bmatrix} \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \sin(i\pi\theta),\ 0 \le \theta \le 1,\ t \in R,$$

where

$$h = \begin{bmatrix} f \\ g \end{bmatrix} \in X$$

and α_i and β_i are Fourier coefficients of f and g, respectively. Since $\mathcal{U}$ is a group, we have $\mathcal{U}_t^* = \mathcal{U}_{-t}$. Therefore, the controllability operator Γ_T of the system (20) is

$$\Gamma_T h = \int_0^T \mathcal{U}_{T-t} B B^* \mathcal{U}_{T-t}^* h\, dt = \int_0^T \mathcal{U}_t B B^* \mathcal{U}_{-t} h\, dt,\ h \in X.$$

We have

$$[\mathcal{U}_{-t} h]_\theta = \sum_{i=1}^{\infty} \begin{bmatrix} \alpha_i \cos(i\pi t) - \beta_i (i\pi)^{-1}\sin(i\pi t) \\ \alpha_i i\pi \sin(i\pi t) + \beta_i \cos(i\pi t) \end{bmatrix} \sin(i\pi\theta).$$

One can calculate that

$$B^* h = \sum_{i=1}^{\infty} \gamma_i \beta_i,\ h \in X.$$

Hence,

$$B^* \mathcal{U}_{-t} h = \sum_{i=1}^{\infty} \gamma_i (\alpha_i i\pi \sin(i\pi t) + \beta_i \cos(i\pi t))$$

and, consequently,

$$[\mathcal{U}_t B B^* \mathcal{U}_{-t} h]_\theta = \sum_{i=1}^{\infty} \begin{bmatrix} \gamma_i (i\pi)^{-1}\sin(i\pi t) \\ \gamma_i \cos(i\pi t) \end{bmatrix} \sin(i\pi\theta) \\ \times \sum_{j=1}^{\infty} \gamma_j (\alpha_j j\pi \sin(j\pi t) + \beta_j \cos(j\pi t)).$$

Thus, for $T = 2$,

$$[\Gamma_2 h]_\theta = \int_0^2 [\mathcal{U}_t B B^* \mathcal{U}_{-t} h]_\theta\, dt = \sum_{i=1}^{\infty} \begin{bmatrix} \gamma_i^2 \alpha_i \\ \gamma_i^2 \beta_i \end{bmatrix} \sin(i\pi\theta).$$

We obtain that

$$[(\lambda I + \Gamma_2) h]_\theta = \sum_{i=1}^{\infty} \left(\lambda + \gamma_i^2\right) \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \sin(i\pi\theta),$$

which implies

$$[R(\lambda, -\Gamma_2) h]_\theta = \left[(\lambda I + \Gamma_2)^{-1} h\right]_\theta = \sum_{i=1}^{\infty} \left(\lambda + \gamma_i^2\right)^{-1} \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \sin(i\pi\theta).$$

Finally, for all $h \in X$,

$$\|\lambda R(\lambda, -\Gamma_2)h\|^2 = \sum_{i=1}^{\infty} \frac{\lambda^2}{\left(\lambda + \gamma_i^2\right)^2} \left(i^2\pi^2\alpha_i^2 + \beta_i^2\right) \to 0$$

as $\lambda \to 0$ if $\gamma_i \neq 0$ for all $i = 1, 2, \ldots$. Thus, by (2a) $\Leftrightarrow$ (2c) in Theorem 2, we obtain the following sufficient condition for the approximate controllability of the system (17)–(18) which agrees with Theorem 2.10 in Zabczyk [6]: if $T \geq 2$ and b is such that

$$\gamma_i = 2\int_0^1 b_\theta \sin(i\pi\theta)\, d\theta \neq 0,\ i = 1, 2, \ldots,$$

then the system (17)–(18) is D^a-controllable.

6. Necessary and sufficient conditions for C- and S-controllabilities. The following two lemmas are proven in [7, 8].

Lemma 3. *For $\varepsilon > 0$ and for $0 \leq p < 1$, the control system* (4) *is $C^c_{\varepsilon,p}$-controllable* ($C^a_{\varepsilon,p}$*-controllable*) *on* $\mathbf{T}$ *if and only if the control system* (5) *is D^c-controllable* (D^a*-controllable*) *on* $\mathbf{T}$ *and the control system* (6) *is $S^0_{\varepsilon,p}$-controllable on* $\mathbf{T}$.

By this lemma, the study of the $C^c_{\varepsilon,p}$-controllability (the $C^a_{\varepsilon,p}$-controllability) of the control system (4) is separated into the study of the D^c-controllability (D^a-controllability) and the $S^0_{\varepsilon,p}$-controllability of the control systems (5) and (6), respectively.

Lemma 4. *The following statements hold:*

(a) *there exists a finite limit*

$$a_T = \lim_{\lambda \to 0} \int_0^T \operatorname{tr} CP_sQ_s^\lambda P_sC^* ds,$$

where Q^λ and P are solutions of (8) *and* (9)*, respectively*;

(b) *the control system* (6) *is $S^0_{\varepsilon,p}$-controllable on* $\mathbf{T}$ *if $a_T < \varepsilon(1-p)$*;

(c) *the system* (6) *is S^0-controllable on* $\mathbf{T}$ *if $a_T = 0$*.

It turns out that the condition $a_T = 0$, which is sufficient for the system (6) to be S^0-controllable, is weaker than the D^a-controllability (particularly, the D^c-controllability) of the control system (5) on all the intervals $[0, t]$ with $0 < t \leq T$.

Lemma 5. *If the control system* (5) *is D^a-controllable on all the intervals* $[0, t]$ *with $0 < t \leq T$, then $a_T = 0$, where a_T is defined in Lemma* 4(a).

Proof. From (2a) $\Leftrightarrow$ (2c) (see Theorem 2), we obtain that $\lambda R(\lambda, -\Gamma_{T-t})$ strongly converges to zero operator as $\lambda \to 0$ for all $0 \leq t < T$. Hence, by Lemma 1, Q_t^λ strongly converges to zero operator as $\lambda \to 0$ for all $0 \leq t < T$. Furthermore, substituting $h = \lambda^{1/2}(\lambda I + \Gamma_{T-t})^{-1/2}x$ in

$$\left\langle \lambda^{-1}(\lambda I + \Gamma_{T-t})h, h\right\rangle \geq \langle h, h\rangle,$$

we obtain

$$\left\langle \lambda(\lambda I + \Gamma_{T-t})^{-1}x, x\right\rangle \leq \|x\|^2.$$

Thus, $\lambda R(\lambda, -\Gamma_{T-t}) \leq I$ and by Lemma 1, $Q_t^\lambda \leq \mathcal{U}_{T-t}^*\mathcal{U}_{T-t}$ for all $\lambda > 0$ and for all $0 \leq t \leq T$. Hence, we can change the places of limit, integral, and trace in definition of the number a_T in Lemma 4(a) to obtain $a_T = 0$. The lemma is proven.

THEOREM 3. *The control system* (4) *is* C*-controllable on all the intervals* $[0,T]$ *with* $T > 0$ *if and only if the control system* (5) *is* D^c*-controllable on all the intervals* $[0,T]$ *with* $T > 0$.

This follows from Lemmas 3, 4(c), and 5.

THEOREM 4. *The control system* (4) *is* S*-controllable on all the intervals* $[0,T]$ *with* $T > 0$ *if and only if the control system* (5) *is* D^a*-controllable on all the intervals* $[0,T]$ *with* $T > 0$.

The necessity follows from Theorem 5(b) in [8]. The sufficiency follows from Lemmas 3, 4(c), and 5.

COROLLARY 1. *The control system* (4) *is* S*-controllable on all the intervals* $[0,T]$ *with* $T > 0$ *if it is* C^a*-controllable on all the intervals* $[0,T]$ *with* $T > 0$.

Proof. From Lemmas 3 and 5 and Theorem 4, one can see that each of the S- and C^a- controllabilities of the control system (4) on all the intervals $[0,T]$ with $T > 0$ is equivalent to the D^a-controllability of the control system (5) on all the intervals $[0,T]$ with $T > 0$.

Example 3. Consider the control system (4) with the operators A and B as defined in Example 1. It was shown that the deterministic part of this system is D^a-controllable on all the intervals $[0,T]$ with $T > 0$. Hence, this system is S-controllable on all the intervals $[0,T]$ with $T > 0$.

Example 4. Consider the control system (4) with the operators A and B as defined in Example 2. It was shown that the deterministic part of this system is D^a-controllable on all the intervals $[0,T]$ with $T > 2$ if some additional condition holds. However, Theorem 4 does not guarantee the S-controllability of this system.

REFERENCES

[1] R. E. KALMAN, *A new approach to linear filtering and prediction problems*, Trans. ASME Ser. D, J. Basic Engineering, 82 (1960), pp. 35–45.

[2] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Inform. Sci. 8, Springer-Verlag, New York, 1978, pp. 1–297.

[3] R. F. CURTAIN AND H. J. ZWART, *An Introduction to Infinite Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.

[4] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, New York, 1976.

[5] A. BENSOUSSAN, G. DA PRATO, M. C. DELFOUR, AND S. K. MITTER, *Representation and Control of Infinite Dimensional Systems*, Vol. 2, Birkhauser, Berlin, 1993.

[6] J. ZABCZYK, *Mathematical Control Theory: An Introduction*, Birkhauser, Berlin, 1992.

[7] A. E. BASHIROV, *On weakening of the controllability concepts*, in Proceedings of the 35th IEEE Conference on Decision and Control, Kobe, Japan, 1996, pp. 640–645.

[8] A. E. BASHIROV AND K. R. KERIMOV, *On controllability conception for stochastic systems*, SIAM J. Control Optim., 35 (1997), pp. 384–398.

[9] G. DA PRATO AND V. BARBU, *A representation formula for the solutions to the operator Riccati equation*, Differential Integral Equations, 5 (1992), pp. 821–829.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 6, pp. 1822–1847

ASYMPTOTIC EFFICIENCY OF PERTURBATION-ANALYSIS-BASED STOCHASTIC APPROXIMATION WITH AVERAGING[*]

QIAN-YU TANG[†], PIERRE L'ECUYER[†], AND HAN-FU CHEN[‡]

Abstract. Central limit theorems are obtained for the perturbation analysis Robbins–Monro single run (PARMSR) algorithm updated either after every regenerative cycle or after every fixed-length observation period, and with averaging of the iterates, for one-dependent regenerative processes. When the convergence to the optimizer is expressed in terms of the total observation time of the system (or the total computing budget in the case of a simulation), the convergence rate and the limit covariance matrix turn out to be the same for all updating schemes and are optimal within the class of stochastic approximation-algorithms under certain assumptions. A bound on the strong convergence rate of the usual PARMSR algorithm updated after every fixed-length observation period is established using a limit theorem on double array martingales. This is the key step for obtaining central limit theorems for the algorithms with averaging and has interest in its own right.

Key words. perturbation analysis, asymptotic efficiency, central limit theorems, stochastic approximation, recursive estimation, queueing theory

AMS subject classifications. 62L20, 93E12, 93E23, 93E30, 60F05, 60K25

PII. S0363012997321243

1. Introduction.

Consider a discrete-time stochastic process $\{X_i,\ i \geq 0\}$ which follows the stochastic recurrence

$$X_{i+1} = h(X_i,\ \tilde{\theta}_i,\ U_i) \tag{1.1}$$

with initial state X_0, where h is a measurable function, each $\tilde{\theta}_i \in D$ is a *decision* (or control), D is a compact subset of the l-dimensional real space $\mathbb{R}^l$, and $\{U_i,\ i \geq 0\}$ is a sequence of i.i.d. random vectors defined on a probability space $\{\Omega,\ \mathcal{F},\ P\}$. In the case of a simulation model, $\{U_i,\ i \geq 0\}$ can be interpreted as the sequence of $U(0,1)$ random numbers that drive the simulation. A real-valued *cost process* $\{J_i,\ i \geq 0\}$ is defined by $J_i = \phi(X_i, \tilde{\theta}_i)$ for some measurable mapping ϕ. Denote (X_i, J_i) by $(X_i(\theta), J_i(\theta))$ when the control parameter $\tilde{\theta}_i$ is fixed at θ for all $i \geq 0$. The objective function is the steady-state average cost

$$\overline{J}(\theta) = \lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} \mathsf{E}[J_i(\theta)], \tag{1.2}$$

which is to be minimized with respect to θ for $\theta \in D$.

To insure that (1.2) exists and to prove our results, we assume certain stability conditions on the process $\{J_i,\ i \geq 0\}$, expressed in the form of a regenerative

[*] Received by the editors May 9, 1997; accepted for publication (in revised form) January 11, 1999; published electronically October 4, 1999.
http://www.siam.org/journals/sicon/37-6/32124.html

[†] Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, Québec, H3C 3J7, Canada (qytang@inrs-telecom.uquebec.ca, lecuyer@iro.umontreal.ca). The research of the second author was supported by NSERC-Canada grants ODGP0110050 and SMF0169893 and by FCAR-Québec grant 93ER1654.

[‡] Laboratory of Systems and Control, Institute of Systems Science, Academia Sinica, Beijing 100080, People's Republic of China (hfchen@iss03.iss.ac.cn). The research of this author was partially supported by the National Natural Science Foundation of China.

structure. We suppose (roughly) that this process, as well as its stochastic gradient (perturbation analysis) process, is one-dependent regenerative in an extended sense (see the next section for details). One-dependent regenerative is a weakened version of the classical notion of a regenerative stochastic process and covers a broad class of systems [1, 29, 36]. The weakening is that adjacent regenerative cycles are allowed to be dependent. The stochastic recursion model (1.1) is very general (it suffices to put enough information into the state X_i), and it will permit us to define the perturbation analysis gradient estimators (under some additional conditions).

We assume that no closed-form expression for $\bar{J}$ is available, and we consider a stochastic approximation (SA) approach to find a root of $f(\theta) \overset{\triangle}{=} \overline{J}_\theta(\theta) \overset{\triangle}{=} d\overline{J}(\theta)/d\theta$. (In this paper, a θ subscript means the gradient with respect to θ.) We suppose that the root is the unique optimizer θ^0 of $\overline{J}(\theta)$. The Robbins–Monro [33] SA algorithm has the general form

$$\theta_{n+1} = \Pi_D(\theta_n - a_n f_{n+1}) \tag{1.3}$$

with an initial value θ_0, where $\{a_n,\ n \geq 0\}$ is a deterministic sequence of matrices called the *step sizes*, $f_{n+1} = f(\theta_n) + \varepsilon_{n+1}$ is an unbiased estimator of $f(\theta_n)$, ε_{n+1} is the observation *noise* at the $(n+1)$th step, $\Pi_D : \mathbb{R}^l \to D$ is a projection operator (which would disappear if D was $\mathbb{R}^l$, as in the original Robbins–Monro formulation), and θ_n is the nth estimate for the optimizer θ^0. See, e.g., [7, 22, 24, 28]. It is well known that under certain conditions on the regression function $f(\theta)$, on the noise sequence $\{\varepsilon_n,\ n \geq 1\}$, and on the matrix A^* assuming that $a_n = A^*/n$, $\sqrt{n}(\theta_n - \theta^0)$ is asymptotically $N(0,\ S^*)$, where S^* is a limiting covariance matrix. The trace of S^* is minimized by taking $A^* = M_1^{-1}$, where $M_1 = f_\theta(\theta^0)$ is the Hessian matrix of $\overline{J}(\theta)$ at θ^0. This optimal covariance matrix is $S^* = M_1^{-1}S_0^*(M_1^{-1})'$, where the prime means "transpose" and S_0^* is the asymptotic covariance matrix of $(1/\sqrt{n})\sum_{j=1}^n \varepsilon_j$.

However, since M_1 is generally unknown, this optimal SA scheme is usually impractical. This has motivated the introduction of SA algorithms with averaging of the iterates, using a sequence of step sizes which decreases at a rate slower than $1/n$ (see [3, 5, 24, 31, 32, 43]). One of these algorithms uses (1.3) as usual, then retains the following estimator of θ^0 at step n:

$$\overline{\theta}_n = \frac{1}{n}\sum_{j=1}^n \theta_j = \frac{1}{n}(\theta_n + (n-1)\bar{\theta}_{n-1}). \tag{1.4}$$

Under certain conditions, $\sqrt{n}(\overline{\theta}_n - \theta^0)$ is asymptotically $N(0,\ S^*)$. The major advantage of this averaging algorithm is that there is no need to know M_1^{-1}.

Under appropriate smoothness conditions, *infinitesimal perturbation analysis* (IPA) offers a viable means of estimating $f(\theta)$ by observing a single sample path of the system (see section 2.2 and, e.g., [15, 21, 25] and other references therein). Going one step further, one can simultaneously optimize along that single sample path: At successive epochs along the path, a (generally biased) estimator of $f(\theta)$ for the current θ is put in the SA algorithm to recursively estimate θ^0 while the system is running (either on-line or via simulation). The general idea is as follows: Run the system for a short period of time, say, L_{n+1} steps of (1.1) with θ fixed at θ_n, and use the collected information to compute the estimator f_{n+1} of $f(\theta_n)$; put f_{n+1} in (1.3) to get θ_{n+1}; continue running the system (from its current state) for another L_{n+2} steps with $\theta = \theta_{n+1}$ to compute f_{n+2}; and so on. Denoting $N_n = \sum_{j=1}^n L_j$ (with $N_0 = 0$), one has $\tilde{\theta}_i = \theta_n$ for $N_n \leq i < N_{n+1}$. In [38, 39], this is called the perturbation analysis Robbins–Monro single run (PARMSR) algorithm. Considerable effort has been devoted, in recent years, to studying the convergence of the PARMSR algorithm in the field of discrete event dynamic systems (DEDSs); see, e.g., [8, 9, 10, 14, 23, 26, 27, 40, 41], among others. The convergence rate results mentioned previously for the SA algorithm, however, have been proved under conditions that do not hold in the PARMSR setup.

In this paper we study the convergence rate, in the sense of central limit theorems, for the PARMSR algorithm with averaging. Let N_n represent the cumulative *computing budget* for the first n steps of the PARMSR algorithm. We express the convergence speed of the algorithm in terms of N_n (as done, e.g., in [28]) instead of n. When $\sqrt{N_n}(\overline{\theta}_n - \theta^0)$ is asymptotically $N(0,\ S^*)$, with $S^* = M_1^{-1} S_0^* (M_1^{-1})'$, where S_0^* is the asymptotic covariance matrix of $(1/\sqrt{N_n}) \sum_{j=1}^{n} \varepsilon_j$, we say that the PARMSR algorithm is *asymptotically optimal* (see [3, 5, 43] for similar definitions in terms of n). This must be understood as optimal only within the class of SA algorithms and for the particular gradient estimator that is used. Another type of convergence rate, in the sense of large-deviation bounds on the probability of exit of a neighborhood of θ^0 by the tail process $\{\theta_n,\ n \geq n_0\}$ for large n_0, is studied in [12].

We analyze the following two cases:

- (R) The parameter θ_n is updated after each regenerative cycle of the process $\{J_i,\ i \geq 0\}$ (so L_n is random and represents the length of the nth regenerative cycle). In this regenerative case, we use θ_{n-1} instead of θ_n to obtain f_{n+1}, as explained in section 3.
- (F) θ_n is updated after every L steps of the recurrence (1.1), so $L_n = L$, a positive constant.

For case (R), the limiting behavior is relatively easy to analyze since the main part of the observation noise can be decomposed into two martingale difference sequences. Then, standard SA results are applicable. For case (F), the analysis is more difficult, primarily because the standard conditions on the observation noise, assumed in, e.g., [3, 24, 31, 32, 43], do not hold. These authors require the observation noise to satisfy the properties of martingale differences, or of stationary ϕ-mixing processes, or of the infinite sum of a martingale difference sequence. For the PARMSR algorithm with fixed-length observation period, the observation noise has a very complicated dynamic, as shown in previous convergence studies; see, e.g., [9, 10, 23, 26, 40, 41]. In this paper, we first obtain a bound on the almost sure (a.s.) convergence rate of the usual PARMSR algorithm (without averaging), using a limit theorem on double array martingales borrowed from [6] and [19]. We then apply this result to obtain a central limit theorem showing the asymptotic optimality of the PARMSR algorithm with averaging for the case (F).

Our results imply that the PARMSR algorithm with averaging has the same asymptotic efficiency, i.e., the same convergence rate and the same limiting covariance matrix in terms of N_n, for both (R) and (F). Moreover, for (F), this limit covariance is independent of the updating frequency L. Our emphasis in this paper is on the case (F). This case is more difficult to analyze than (R), but its implementation is much easier, because there is no need to recognize the regeneration points, so the implementation depends much less on the structure of the system. For case (R), the algorithm must identify the regeneration points explicitly, which is usually hard for complex systems. See [34, 35] on the identification of regeneration points for closed and open queueing networks.

A third approach, not considered here, is to have $L_n \to \infty$ with n. This was studied in [28] for decreasing step sizes and in [13] for constant step sizes, without the averaging of the iterates.

The rest of the paper is organized as follows. In the next section we specify the regenerative structure of the model. We also define the IPA estimators and introduce some assumptions to make sure that the IPA process shares a common regenerative structure with the original process and is strongly consistent. The asymptotic efficiency of the PARMSR algorithm with averaging for cases (R) and (F) is analyzed in sections 3 and 4, respectively. All the proofs are collected in the appendix.

2. The model.

2.1. The regenerative structure.

We now expand on the model introduced at the beginning of section 1. The state process $\{X_i,\, i \geq 0\}$ takes values in some Borel space $\mathcal{X}$, and the initial state $X_0 \in \mathcal{X}$ is a random variable. The stochastic process $\{X_i, \tilde{\theta}_i, J_i),\, i \geq 0\}$ is defined on the same probability space $\{\Omega,\, \mathcal{F},\, P\}$ as $\{X_0,\, U_i,\, i \geq 0\}$, via (1.1) and (1.3). Let $\mathcal{F}_i = \sigma(X_0, \ldots, X_i)$, the sigma field generated by $X_0, \ldots, X_i$. We suppose that there is a sequence of integers $0 = k_0 < k_1 < k_2 < \cdots$ such that each k_m is a stopping time with respect to the filtration $\{\mathcal{F}_i,\, i \geq 0\}$ and such that the following condition (A0) is satisfied.

Suppose for a moment that until time k_m, the process is run as usual with $\tilde{\theta}_i$ varying according to the SA algorithm, but that from then on, the parameter is fixed at θ; i.e., $\tilde{\theta}_i = \theta$ for $i \geq k_m$. We denote this modified process starting at time k_m by $\{X_i^{(m+1)}(\theta),\, i \geq 0\}$, that is, $X_i^{(m+1)}(\theta) = X_{k_m+i}$ for $i \geq 0$, and put $J_i^{(m+1)}(\theta) = \phi(X_i^{(m+1)}(\theta), \theta)$ for $i \geq 0$ and $k_r^{(m+1)}(\theta) = k_{m+r} - k_m$ for $r \geq 0$. We also denote $X_i(\theta) = X_i^{(1)}(\theta)$, $J_i(\theta) = J_i^{(1)}(\theta)$, and $k_m(\theta) = k_m^{(1)}(\theta)$, which correspond to a process where θ is fixed from the beginning.

(A0) For any fixed $\theta \in D$ and $m \geq 0$, $\{J_i^{(m+1)}(\theta),\, i \geq 0\}$ and $\{k_r^{(m+1)}(\theta),\, r \geq 0\}$ have the same joint distribution as $\{J_i(\theta),\, i \geq 0\}$ and $\{k_r(\theta),\, r \geq 0\}$ and are independent of $\mathcal{F}_{k_{m-1}-1}$ and k_m.

When this condition (A0) is satisfied, the controlled process $\{J_i,\, i \geq 0\}$ is called (nondelayed) *one-dependent regenerative*. This extends the usual definition found, e.g., in [29, 36]: For the process $\{J_i(\theta),\, i \geq 0\}$, for which θ is fixed from the beginning, we get the usual definition of one-dependent regenerative, which we call in this paper *one-dependent regenerative for fixed* θ. The k_m's are called the *regeneration points*. Denote $\eta_m = k_m - k_{m-1}$, the length of the mth regenerative cycle, $\eta_r^{(m+1)}(\theta) = k_r^{(m+1)}(\theta) - k_{r-1}^{(m+1)}(\theta)$, and $\eta_m(\theta) = k_m(\theta) - k_{m-1}(\theta)$. If (A0) holds with $\mathcal{F}_{k_{m-1}-1}$ replaced by $\mathcal{F}_{k_m-1}$, the process is called *classically regenerative*. From the definition of J_i and by Proposition V.1.1 of [1], if (A0) holds for $\{X_i,\, i \geq 0\}$, it also holds for $\{J_i,\, i \geq 0\}$ with the same regeneration points.

It is well known that Harris-recurrent Markov chains (HRMCs) are one-dependent regenerative processes for fixed θ; this can be seen by applying the splitting technique due to [2]. HRMCs cover a very large class of models. We refer the reader to [1, 29, 30, 36] for appropriate background.

If $\mathsf{E}[\eta_1(\theta)] < \infty$ and $\mathsf{E}[\sum_{i=1}^{\eta_1(\theta)} |J_i(\theta)|] < \infty$ for all $\theta \in D$, then $\overline{J}(\theta)$ in (1.2) is well defined on D and the renewal-reward theorem (see, e.g., [1] and [42]) implies that

$$\overline{J}(\theta) = \frac{1}{\mathsf{E}[\eta_1(\theta)]} \mathsf{E}\left[\sum_{i=1}^{\eta_1(\theta)} J_i(\theta)\right]. \tag{2.1}$$

2.2. Gradient estimation via IPA and the PARMSR algorithm. We now explain how IPA is used to estimate the gradient in this context. For fixed θ, the IPA estimators are constructed from the recurrence

$$\begin{aligned} J_{\theta,i}(\theta) &= \phi_x(X_i(\theta),\theta)X_{\theta,i}(\theta) + \phi_\theta(X_i(\theta),\theta), \\ X_{\theta,i+1}(\theta) &= h_x(X_i(\theta),\theta,U_i)X_{\theta,i}(\theta) + h_\theta(X_i(\theta),\theta,U_i), \end{aligned}$$

for $i \geq 0$, where $X_{\theta,0}(\theta) = 0$, ϕ_x and ϕ_θ denote the partial derivatives of ϕ with respect to its first and second components, and similarly for h. Assuming that these stochastic derivatives exist and under appropriate uniform integrability conditions (see [15, 25]), $J_{\theta,i}(\theta)$ is an unbiased IPA estimator of the gradient of $\mathsf{E}[J_i(\theta)]$.

Suppose now that θ is *not* fixed but takes the successive values $\{\tilde{\theta}_i,\ i \geq 0\}$ as in (1.1). This is what happens when the PARMSR algorithm is applied. We then define the IPA estimator as follows. Let

$$\begin{aligned} J_{\theta,i} &= \phi_x(X_i,\tilde{\theta}_i)X_{\theta,i} + \phi_\theta(X_i,\tilde{\theta}_i), \\ X_{\theta,i+1} &= h_x(X_i,\tilde{\theta}_i,U_i)X_{\theta,i} + h_\theta(X_i,\tilde{\theta}_i,U_i). \end{aligned}$$

The gradient estimator at the nth step of SA has the general form

$$f_{n+1} = K \sum_{i=1}^{L_{n+1}} J_{\theta,N_n+i} \tag{2.2}$$

for some constant K. In our context, we have $L_{n+1} = \eta_{n+1}$ and $K = 1$ for case (R), and $L_{n+1} = L$ and $K = 1/L$ for case (F).

Equations (1.3) and (2.2) give the general form of the PARMSR algorithm. In this paper, we define the projection operator Π_D of (1.3) as

$$\Pi_D(\theta_n - a_n f_{n+1}) = \begin{cases} \theta_n - a_n f_{n+1} & \text{if } \theta_n - a_n f_{n+1} \in D, \\ \\ \theta_n & \text{if } \theta_n - a_n f_{n+1} \notin D. \end{cases}$$

If D is a convex set, we could also define $\Pi_D(x)$ to be the nearest boundary point of D whenever $x \notin D$.

For the analysis of the PARMSR algorithm, we need to assume that $\{J_{\theta,i},\ i \geq 0\}$ is also one-dependent regenerative with the same regeneration points as $\{J_i,\ i \geq 0\}$. Glasserman [15, 16, 17] and Glasserman, Hu, and Strickland [18] give several sets of sufficient conditions for this property to hold for fixed θ, for classically regenerative systems (see, e.g., Theorem 8.3 in [15], Theorem 5.4 in [16], and Theorem 5.6 in [17]). The following conditions ensure, among other things, that the IPA estimators are *strongly consistent* for $f(\theta)$, in the sense that $\lim_{t\to\infty}(1/t)\sum_{i=1}^{t} J_{\theta,i}(\theta) = f(\theta)$ a.s. We denote $J_{\theta,i}^{(m+1)}(\theta) = dJ_i^{(m+1)}(\theta)/d\theta$. Later on, for our theorems, we will require values of ξ_0 larger than 2 in (A3).

(A1) For $i \geq 0$, $J_i(\theta)$ is absolutely continuous with respect to θ on D.

(A2) $\{J_i,\ i \geq 0\}$ and $\{J_{\theta,i},\ i \geq 0\}$ are one-dependent regenerative in the sense of (A0), with common regeneration points $\{k_m,\ m \geq 0\}$.

(A3) There are two sequences of one-dependent and identically distributed random variables (r.v.'s) $\{Z_m,\ m \geq 1\}$ and $\{\eta_m^*,\ m \geq 1\}$, and a constant $\xi_0 \geq 2$, such that $\mathsf{E}[Z_1^{\xi_0}] < \infty$ and $\mathsf{E}\left[(\eta_1^*)^{\xi_0}\right] < \infty$, and such that for all $m \geq 0$, $1 \leq i \leq \eta_{m+1}$, and $\theta \in D$, $\max(\|J_{\theta,k_m+i}\|,\ \|J_{\theta,k_m+i}(\theta)\|,\ \|J_{\theta,i}^{(m+1)}(\theta)\|)$ $\leq Z_{m+1}$ and $\max(\eta_{m+1},\ \eta_{m+1}(\theta),\ \eta_1^{(m+1)}(\theta)) \leq \eta_{m+1}^*$.

A classical example where (A1) to (A3) are verified in the literature is that of the $GI/G/1$ queue considered in [8, 9, 24, 26, 41].

LEMMA 2.1. *If conditions* (A0)–(A3) *hold, then for each* $\theta \in D$, *the IPA derivative estimator is strongly consistent for* $f(\theta)$ *and*

$$f(\theta) = \frac{1}{\mathsf{E}[\eta_1(\theta)]} \mathsf{E}\left[\sum_{i=1}^{\eta_1(\theta)} J_{\theta,i}(\theta)\right]. \tag{2.3}$$

Proof. The proof is along the same lines as in [15, 16, 17, 18]. In fact (A2) and (A3) are needed only for fixed θ (uniformly in θ) to prove this lemma. □

3. The PARMSR algorithm with averaging for case (R).

Since $\eta_1(\theta) \geq 1$, finding a root of $f(\theta)$ is equivalent to finding a root of $f(\theta)\overline{\eta}(\theta)$, where $\overline{\eta}(\theta) = \mathsf{E}[\eta_1(\theta)]$, i.e., to finding a root of the second expectation in (2.3). If regeneration points for the model under consideration can be identified explicitly, one can use the sum of the $J_{\theta,i}(\theta)$ over any regenerative cycle as an unbiased estimator of $f(\theta)\overline{\eta}(\theta)$, where θ is the (fixed) value of the parameter during that cycle. This motivates taking $L_{n+1} = \eta_{n+1}$ and $K = 1$ in (2.2).

The PARMSR algorithm *without averaging* for case (R) is composed of (1.3) and (2.2) (cf. [8, 14, 26, 27]), with $\tilde{\theta}_i = \theta_{n-1}$ for $k_n < i \leq k_{n+1}$, and where $\theta_{-1},\ \theta_0 \in D$ are initial values. Thus, f_{n+1} is viewed as an estimator of $f(\theta_{n-1})\overline{\eta}(\theta_{n-1})$. It is worth noticing that the decision parameter throughout the evolution of the $(n+1)$th regenerative cycle is fixed at θ_{n-1} rather than θ_n. This is because of the one-dependent nature of our model, and it simplifies the convergence analysis.

The *averaging* version of this PARMSR algorithm works the same, but retains $\overline{\theta}_n$ instead of θ_n as an estimator of θ^0. Our aim is to obtain a central limit theorem for $\{\bar{\theta}_n,\ n \geq 1\}$. For this, we introduce the following conditions on the step sizes a_n and the functions f and $\overline{\eta}$.

(A4) There are constants $\overline{a} > 0$ and $\nu \in (1/2,\ 1)$ such that $0 < a_n \leq \overline{a} n^{-\nu}$ for all $n \geq 1$; $a_j = \overline{a}$ for all $j \leq 0$; $\sum_{n=1}^{\infty} a_n = \infty$; $0 \leq a_{n+1}^{-1} - a_n^{-1} \to 0$ as $n \to \infty$

(A5) The set D is compact and convex, and $f(\theta)$ is bounded on D. The optimizer θ^0 is an interior point of D. There are a stable matrix $-M_1$ (all eigenvalues of M_1 have positive real parts) and positive constants r_0 and c_1 such that $\|f(\theta) - M_1(\theta - \theta^0)\| \leq c_1 \|\theta - \theta^0\|^2$, whenever $\|\theta - \theta^0\| \leq r_0$.

(A6) There exists a continuously differentiable function $v : R^l \to R$ such that $v(\theta^0) = 0$ and for all $\Delta_1 > 0$, $\inf\{(f(\theta))' v_\theta(\theta) : \theta \in D,\ \|\theta - \theta^0\| \leq \Delta_1\} > 0$.

(A7) $f(\theta)$ is Lipschitz with modulus B_1 on D; i.e., $\|f(x_1) - f(x_2)\| \leq B_1 \|x_1 - x_2\|\ \forall\ x_1,\ x_2 \in D$.

(A8) $\overline{\eta}(\theta)$ and $H(\theta)$ are continuous at θ^0, where $H(\theta)$ is defined as

$$H(\theta) = \mathsf{E}\left[\left(\sum_{i=1}^{\eta_1(\theta)} J_{\theta,i}(\theta)\right)\left(\sum_{i=1}^{\eta_1(\theta)} J_{\theta,i}(\theta)\right)'\right].$$

Let us comment on conditions (A4)–(A8). Observe that $\nu < 1$ in (A4), so the classical choice of $a_n = O(1/n)$ is not allowed. This (A4) is the usual condition of *slowly decreasing step sizes* (e.g., as in [3, 24, 32]). Conditions (A5) and (A6) are standard in the context of SA. (A5) asks the objective function to be locally quadratic around the optimizer θ^0, while (A6) says that θ^0 should be an attractor point from every other point of D. In the context of queueing systems, for example, the set D may represent a stability region where the standard load conditions are fulfilled. In the case where D is unbounded, an SA procedure with randomly varying truncations can be employed (see, e.g., [3, 4, 5] and [7]). Condition (A7) requires that the regression function $f(\theta)$ is sufficiently smooth. We note that if the model is classically regenerative, then the Lipschitz conditions on $f(\theta)$ and $\overline{\eta}(\theta)$ required in Theorem 3.1 can be dropped since in this situation, $\tilde{\theta}_i = \theta_{n-1}$ can be replaced by $\tilde{\theta}_i = \theta_n$ for $k_n < i \le k_{n+1}$, and we do not need to decompose the noise ε_{n+1} into two parts (see the proof of Theorem 3.1(i)). Condition (A8) requires the continuity of $\overline{\eta}(\theta)$ and $H(\theta)$ at the optimizer θ^0. This is a mild condition. See [26] for the verification in the context of a $GI/G/1$ queue.

The following theorem contains our main results for case (R).

THEOREM 3.1.

(i) *Suppose that the conditions* (A0)–(A7) *hold with* $\xi_0 \ge 4$ *in* (A3) *and that* $\overline{\eta}(\theta)$ *is Lipschitz with modulus* B_2. *Then* $\|\theta_n - \theta^0\| = o(a_n^\delta)$ *a.s. for all* $\delta \in [0,\ 1 - 1/(2\nu))$, *where* ν *is given by condition* (A4).

(ii) *Suppose that conditions* (A0)–(A8) *hold with some* $\xi_0 > 4/\nu$ *and that* $\overline{\eta}(\theta)$ *is Lipschitz on* D. *Then* $(\theta_n - \theta^0)/\sqrt{a_n} \to N(0,\ S_1)$ *in distribution as* $n \to \infty$, *where*

$$S_1 = \int_0^\infty e^{-\overline{\eta}(\theta^0)M_1 t} S e^{-\overline{\eta}(\theta^0)M_1' t} dt, \quad S = H(\theta^0).$$

(iii) *Under the same conditions as in* (ii) *with* $\xi_0 > 4/\nu$ *replaced by* $\xi_0 > 4$, *the PARMSR algorithm with averaging for case* (R) *satisfies* $\sqrt{n}(\overline{\theta}_n - \theta^0) \to N(0,\ S_2)$ *in distribution as* $n \to \infty$, *where* $S_2 = \overline{\eta}(\theta^0)^{-2} M_1^{-1} S (M_1^{-1})'$.

For the first n iterations of the PARMSR algorithm for case (R), the total computing budget is $N_n = k_n = \sum_{i=1}^n \eta_i$. By the continuity of $\overline{\eta}(\theta)$ at θ^0 and the law of large numbers for martingales, it is seen that $N_n/n \to \overline{\eta}(\theta^0)$ a.s. as $n \to \infty$. Combining this with Theorem 3.1, we obtain Corollary 3.1.

COROLLARY 3.1. *Under the assumptions of Theorem* 3.1 (iii), *one has* $\sqrt{N_n}(\overline{\theta}_n - \theta^0) \to N(0,\ S^*)$ *in distribution as* $n \to \infty$, *where* $S^* = \overline{\eta}(\theta^0)^{-1} M_1^{-1} S (M_1^{-1})'$.

4. Convergence of the PARMSR algorithm for case (F).

The PARMSR algorithm *without averaging*, with updating period L, is composed of (1.3) and (2.2) with $L_n = L$, $K = 1/L$, and $\tilde{\theta}_{nL+i} = \theta_n$ for $0 \le i \le L-1$ and $n \ge 0$. For $L = 1$, we have $\tilde{\theta}_i = \theta_i$ for all $i \ge 0$. The *averaging* version uses $\overline{\theta}_n$ as a final estimate of θ^0.

(A9) There exists a sequence of one-dependent and identically distributed r.v.'s $\{Z_{m+1}^{(0)},\ m \ge 0\}$, with $\mathsf{E}[(Z_1^{(0)})^{6/\nu}] < \infty$, such that for all $m \ge 0$ and $1 \le i \le \eta_{m+1}$, $\|J_{\theta,k_m+i} - J_{\theta,i}^{(m+1)}(\tilde{\theta}_{k_m})\| \le a_{k_m} Z_{m+1}^{(0)}$ a.s.

(A10) There are two positive constants α_0 and γ_1 such that

$$P\{\eta_{m+1} \ne k_1^{(m+1)}(\tilde{\theta}_{k_m}) \mid \mathcal{F}_{k_m}\} \le \alpha_0 a_{k_m}^{\gamma_1}.$$

(A11) The constants ξ_0 in (A3) and ν in (A4) satisfy $\xi_0 \ge \max\{2/\zeta,\ 4/\nu,\ 2p_1\}$, where $p_1 > 1$ and $\zeta > 0$ are other constants such that for γ_1 in (A10),

$$\nu\,(1 + \gamma_1(1 - 1/p_1)) > 1,\ 0 < \zeta < \delta'\nu,\ \gamma_1(1 - 1/p_1) \ge 1/2,\ \delta' \in (0,\ 1/2].$$

Chong and Ramadge [10] have checked conditions (A9) and (A10) for certain classically regenerative systems, although the convergence rates of the PARMSR algorithms have not been studied there. For example, (A9) and (A10) have been verified in [9, 10, 40, 41] in the context of a $GI/G/1$ queue, with some conditions on the service time and the interarrival time distributions.

THEOREM 4.1. *For the PARMSR algorithm with fixed updating period $L_n = L$, we have the following:*

(i) *If conditions* (A0)–(A10) *hold with $\xi_0 \geq \max\{4,\ 2p_1\}$ and $\nu(1+\gamma_1(1-1/p_1)) > 1$, where $p_1 > 1$ is some constant and γ_1 is given in* (A10)*, then $\theta_n \to \theta^0$ a.s. as $n \to \infty$.*

(ii) *If* (A0)–(A11) *hold and the constant δ' satisfies* (A11)*, one has $\|\theta_n - \theta^0\| = o(a_n^{1/2-\delta'})$ a.s.*

(iii) *If* (A0)–(A11) *hold with $\nu\gamma_1(1-1/p_1) > 1/2$ and $0 < \delta' < 1/2(1-1/(2\nu))$, then for the algorithm with averaging, one has $\sqrt{n}(\overline{\theta}_n - \theta^0) \to N(0,\ S_3)$ in distribution as $n \to \infty$, where $S_3 = L^{-1}\overline{\eta}(\theta^0)^{-1}M_1^{-1}S(M_1^{-1})'$.*

We note that δ' in Theorem 4.1(ii) is independent of ν. Thus $1/2 - \delta'$ can be arbitrarily close to $1/2$ if condition (A3) is fulfilled with a very large ξ_0. This is different from our Theorem 3.1 and from Theorem 2.2 in [41], where $\|\theta_n - \theta^0\| = o(a_n^\delta)$ a.s. with $\delta \in [0,\ 1-1/(2\nu))$ depending on ν. In the latter setup, if ν is close to $1/2$, δ must be close to zero.

For the first n SA iterations, the total computing budget of the PARMSR algorithm with updating period L is $N_n = nL$. The following corollary provides a central limit theorem in terms of N_n.

COROLLARY 4.1. *Suppose that the conditions of Theorem* 4.1(ii) *are satisfied. Then the PARMSR algorithm with averaging and with updating period L satisfies $\sqrt{N_n}(\overline{\theta}_n - \theta^0) \to N(0,\ S^*)$ in distribution as $n \to \infty$, where S^* is defined in Corollary* 3.1.

It follows from Corollaries 3.1 and 4.1 that the PARMSR algorithms with averaging, updated either after every regenerative cycle or after every L steps of the process $\{J_i,\ i \geq 0\}$, have the same convergence rate and the same limit covariance matrix, for arbitrary $L \geq 1$, as a function of the computing budget. It should be pointed out that the small-sample or transient behavior of these algorithms generally differ. We also recall that our analysis of the asymptotic behavior is based on what happens after no projection of θ_n on D occurs anymore. Thus, our results do not apply if the optimizer θ^0 lies on the boundary of D. If θ^0 is very close to the boundary, it may take a long while before no projection occurs, and this may affect the convergence speed.

The choices of $K = 1$ for (R) and $K = 1/L$ for (F) in (2.2) appear natural but are not necessary for the convergence results to hold. The two theorems and their corollaries are still valid for an arbitrary constant $K > 0$ in (2.2).

5. Appendix.

Proof of Theorem 3.1. We use the notation $\mathcal{F}^{(m)} = \mathcal{F}_{k_m}$ for $m > 0$ and $\mathcal{F}^{(m)} = \sigma\{X_0,\ \theta_0,\ \theta_{-1}\}$ for $m \leq 0$. For case (R), the observation noise is $\varepsilon_{n+1} = f_{n+1} - f(\theta_n)\overline{\eta}(\theta_n)$.

(i) We first decompose ε_{n+1} as $\varepsilon_{n+1} = \varepsilon_{n+1}^{(1)} + \varepsilon_{n+1}^{(2)}$, where

$$\varepsilon_{n+1}^{(1)} = \sum_{i=1}^{\eta_{n+1}} J_{\theta,k_n+i} - f(\theta_{n-1})\overline{\eta}(\theta_{n-1}), \tag{5.1}$$

$$\varepsilon_{n+1}^{(2)} = (f(\theta_{n-1}) - f(\theta_n))\overline{\eta}(\theta_{n-1}) + f(\theta_n)(\overline{\eta}(\theta_{n-1}) - \overline{\eta}(\theta_n)). \tag{5.2}$$

By Lemma 2.1 and the one-dependence assumption, it is seen that $\{\varepsilon_{2n}^{(1)},\ \mathcal{F}^{(2n)},\ n \geq$

$1\}$ and $\{\varepsilon^{(1)}_{2n-1},\, \mathcal{F}^{(2n-1)},\, n \geq 1\}$ are martingale difference sequences. By (5.1) and condition (A3),

$$\|\varepsilon^{(1)}_{n+1}\| \leq \eta^*_{n+1} Z_{n+1} + \sup_{\theta\in D} \|f(\theta)\| \sup_{\theta\in D} \overline{\eta}(\theta), \tag{5.3}$$

which yields

$$\sup_n \mathsf{E}\left[\|\varepsilon^{(1)}_{n+1}\|^2 \mid \mathcal{F}^{(n-1)}\right] < \infty \tag{5.4}$$

by the Schwarz inequality and (A3). Then, using the local convergence theorem of martingales (see, e.g., [6, 20, 37]), it is seen that $\sum_{n=1}^{\infty} a^{(1-\delta)}_{2n-1}\varepsilon^{(1)}_{2n} < \infty$ and $\sum_{n=1}^{\infty} a^{(1-\delta)}_{2n-2}\varepsilon^{(1)}_{2n-1} < \infty$ a.s., which implies that

$$\sum_{i=1}^{\infty} a^{1-\delta}_{i-1}\varepsilon^{(1)}_i < \infty \quad \text{a.s.} \tag{5.5}$$

By Lemma 2 of [40] and condition (A3), one derives

$$a^{(1-\delta)/2}_{n-1} Z_n \xrightarrow[n\,\to\,\infty]{} 0 \text{ and } a^{(1-\delta)/2}_{n-1}\eta^*_n \xrightarrow[n\,\to\,\infty]{} 0 \text{ a.s.} \quad \forall\, \delta \in [0,\, 1 - 1/(2\nu)). \tag{5.6}$$

Using condition (A3) and (5.6), it follows from (1.3) and (2.2) that

$$\|\theta_n - \theta_{n-1}\| \leq a_{n-1}\|f_n\| \leq a^{\delta}_{n-1} a^{1-\delta}_{n-1}\eta^*_n Z_n = o(a^{\delta}_n) \quad \text{a.s.},$$

which gives

$$\begin{aligned} \|\varepsilon^{(2)}_{n+1}\| &\leq \|\theta_n - \theta_{n-1}\| \left(B_1 \sup_{\theta\in D} \overline{\eta}(\theta) + B_2 \sup_{\theta\in D} \|f(\theta)\| \right) \\ &= O(\|\theta_n - \theta_{n-1}\|) = o(a^{\delta}_n) \quad \text{a.s.} \end{aligned} \tag{5.7}$$

via (5.2) and the Lipschitz conditions.

By (5.5) and (5.7), the a.s. convergence of the algorithm follows from Theorem 3.1 in [4] (see also Theorem 2.4.1 in [7]). Note that we are using a projection algorithm in (1.3), rather than a randomly varying truncation procedure as in [4, 7]. But the convergence analysis works the same way, since θ^0 is assumed in (A5) to be in the interior of D. After establishing the convergence, by (5.5), (5.7), and Theorem 3.2.1 of [7], it follows that $\|\theta_n - \theta\| = o(a^{\delta}_n)$ a.s.

(ii) Since $\theta_n \to \theta^0$ a.s. as $n \to \infty$, there is a finite time n_0 (which may depend on the sample path) such that

$$\theta_{n+1} = \theta_n - a_n f_{n+1} \quad \forall\, n \geq n_0. \tag{5.8}$$

Define $\varphi_{0,\,0} = I$, $\varphi_{n,n+1} = I$, and

$$\varphi_{n,\,k} = (I + a_n A_n)\cdots(I + a_k A_k) \quad \forall\, n \geq k, \tag{5.9}$$

where $\{A_n,\ n \geq 0\}$ is a sequence of deterministic matrices such that $\lim_{n\to\infty} A_n = -\overline{\eta}(\theta^0) M_1$. Then it is standard to derive that

$$\|\varphi_{n,\,k}\| \leq c_0 \exp\left(-c\sum_{j=k}^{n} a_j\right) \text{ and } \sup_n \sum_{i=1}^{n} a_i\|\varphi_{n,\,i+1}\|^{r_1} < \infty \quad \forall\, r_1 > 0, \tag{5.10}$$

where c_0 and c are some positive constants. From (5.8) one derives

$$\begin{aligned}(5.11)\quad &\frac{\theta_{n+1}-\theta^0}{\sqrt{a_{n+1}}}\\ &= \varphi_{n,n_0}\frac{\theta_{n_0}-\theta^0}{\sqrt{a_{n_0}}} - \sum_{i=n_0}^{n}\varphi_{n,i+1}\sqrt{a_i}\varepsilon_{i+1}^{(1)} - \sum_{i=n_0}^{n}\varphi_{n,i+1}\sqrt{a_i}\varepsilon_{i+1}^{(2)}\\ &\quad - \sum_{i=n_0}^{n}\varphi_{n,i+1}o(a_i)\sqrt{a_i}\varepsilon_{i+1} - \sum_{i=n_0}^{n}\varphi_{n,i+1}\sqrt{a_i}(1+o(a_i))O(\|\theta_i-\theta^0\|^2).\end{aligned}$$

In the proof of Theorem 2 of [3], it is shown that $\mathsf{E}[\|\theta_n-\theta^0\|^2]=O(a_n)$, so the last term on the right side of (5.11) converges to zero in probability via (5.10). By the conditions (A2), (A3), and Lemma 2 of [40], it is derived that $\sqrt{a_n}\varepsilon_{n+1}\to 0$ a.s. as $n\to\infty$, which yields that the fourth term on the right side of (5.11) converges to zero a.s. Similarly, we can prove that the first and the third terms converge to zero a.s. Then, the result follows from standard martingale arguments.

(iii) To prove (iii), we verify that all conditions of Theorem 2 in [3] are fulfilled. Using (2.9) in [3], it is easy to see that condition (A4) on the step sizes is equivalent to (1.5) and (1.6) in [3]. The condition on the existence of the Lyapunov function $v(\theta)$ is put in (A6), which is weaker than (A1) in [3] but is consistent with (A2) in [4] and (A4.3.3) in [7]. Note that we use a projection algorithm in (1.3) rather than a randomly varying truncation procedure as in [3, 4, 7]. Condition (A2) in [3] is implied by our condition (A5).

We now check conditions (A3) and (A4) in [3], concerning the observation noise $\varepsilon_n=\varepsilon_n^{(1)}+\varepsilon_n^{(2)}$. Letting $\delta=0$ in (5.5), by the Kronecker lemma, we have that as $n\to\infty$, $a_n\sum_{i=1}^n\varepsilon_{i+1}^{(1)}\to 0$ a.s., which is (2.5) in [3]. Since $\{\varepsilon_{2n}^{(1)},\ \mathcal{F}^{(2n)},\ n\ge 1\}$ and $\{\varepsilon_{2n-1}^{(1)},\ \mathcal{F}^{(2n-1)},\ n\ge 1\}$ are martingale difference sequences, by (5.3) and (5.4) it is easy to derive that (2.6) and (2.7) in [3] are satisfied. From (5.7) it follows that $\|\varepsilon_{n+1}^{(2)}\|=O(\|\theta_n-\theta_{n-1}\|)=O(a_{n-1}\|f_n\|)$. Then we derive that $\mathsf{E}[\|\varepsilon_{n+1}^{(2)}\|^2]=O(a_n^2)$, which satisfies (A4) in [3]. Thus we can apply Theorem 2 in [3] and this concludes the proof. □

Proof of Theorem 4.1. For Theorem 4.1, we first give the proof for $L=1$ for simplicity of writing, then extend the results to $L>1$. As in the proof of Theorem 3.1, for the convergence of $\sum_{n=1}^\infty a_n^{1-\delta}\varepsilon_{n+1}^{(1)}$, it suffices to prove the convergence of $\sum_{n=1}^\infty a_{2n-1}^{1-\delta}\varepsilon_{2n}^{(1)}$ and $\sum_{n=1}^\infty a_{2n-2}^{1-\delta}\varepsilon_{2n-1}^{(1)}$, where $\{\varepsilon_{2n}^{(1)},\ \mathcal{F}^{(2n)},\ n\ge 1\}$ and $\{\varepsilon_{2n-1}^{(1)},\ \mathcal{F}^{(2n-1)},\ n\ge 1\}$ are martingale difference sequences. Such a technique is standard when one wants to extend some results on classically regenerative processes to the one-dependent regenerative processes. Hence, no generality is lost by supposing that $\{J_n(\theta),\ n\ge 1\}$ is a classically regenerative process, which we now do for simplicity of writing. We first give two lemmas.

Lemma 5.1. *Suppose that $\{z_i\}$ is a sequence of r.v.'s with the same distribution. Then for any $r>0$, $\mathsf{E}|z_1|^r<\infty$ implies*

$$\lim_{n\to\infty} n^{-1/r}z_n=0\quad a.s.$$

Furthermore, if $\{z_i,\ i\ge 1\}$ are mutually independent, then the converse is true.

Proof. The proof follows from the Borel–Cantelli lemma and Corollary 4.1.3 in [11, pp. 90–91]. See also Lemma 2 in [40]. □

LEMMA 5.2. *Suppose that* $\{z_n, \mathcal{B}_n^*\}$ *is an* l*-dimensional martingale difference sequence satisfying*

$$\sup_n \mathsf{E}[\|z_{n+1}\|^2 | \mathcal{B}_n^*] < \infty, \quad \|z_n\| = o(h(n)) \text{ a.s.}, \; h(n) \le h(n+1) \quad \forall\, n \ge 0,$$

and that $g_{n,i}$ *is a* $\mathcal{B}_i^*$*-measurable* $l \times l$*-dimensional random matrix, for* $1 \le i \le n$*, which satisfies*

$$\sum_{i=1}^{n} \|g_{n,i}\|^2 \le \overline{g} < \infty \quad \text{a.s. } \forall\, n \ge 1,$$

where $h(n)$ *and* $\overline{g}$ *are positive constants. Then, as* $n \to \infty$,

$$\max_{1 \le i \le n} \left\| \sum_{j=1}^{i} g_{n,j} z_{j+1} \right\| = o(h(n+1) \log n) \quad \text{a.s.}$$

Proof. The proof is in Guo, Huang, and Hannan [19]. See also [6]. □

Proof of Theorem 4.1(i) for $\boldsymbol{L = 1}$**.** The key idea lies in verifying that $\sum_{n=1}^{\infty} a_n \varepsilon_{n+1}$ converges a.s. Then the desired result follows from Theorem 3.1 in [4] (see also Theorem 2.4.1 in [7]). Recall that the observation noise here is

$$\varepsilon_{n+1} = f_{n+1} - f(\theta_n). \tag{5.12}$$

Denote $D_{m,i} = J_{\theta,i}^{(m+1)}(\theta_{k_m})$, the value of f_{k_m+i} obtained if θ_{k_m+j} is fixed at θ_{k_m} for $j \ge 0$.

(a) We first show that $\sum_{m=1}^{\infty} \sum_{i=1}^{\eta_{m+1}} a_{k_m+i-1} \varepsilon_{k_m+i}$ converges a.s. From (2.2) and (5.12) it is easy to see that

$$\begin{aligned} &\sum_{m=1}^{\infty} \sum_{i=1}^{\eta_{m+1}} a_{k_m+i-1} \varepsilon_{k_m+i} \\ &= \sum_{m=1}^{\infty} \sum_{i=1}^{\eta_{m+1}} (a_{k_m+i-1} - a_{k_m}) \varepsilon_{k_m+i} + \sum_{m=1}^{\infty} a_{k_m} \sum_{i=1}^{\eta_1^{(m+1)}(\theta_{k_m})} (D_{m,i} - f(\theta_{k_m})) \\ &\quad + \sum_{m=1}^{\infty} a_{k_m} \sum_{i=1}^{\eta_{m+1}} (f(\theta_{k_m}) - f(\theta_{k_m+i-1})) + \sum_{m=1}^{\infty} a_{k_m} \sum_{i=1}^{\eta_{m+1}} (f_{k_m+i} - D_{m,i}) \\ &\quad + \sum_{m=1}^{\infty} a_{k_m} \sum_{i=\eta_1^{(m+1)}(\theta_{k_m})+1}^{\eta_{m+1}} (D_{m,i} - f(\theta_{k_m})) I\{\eta_{m+1} > \eta_1^{(m+1)}(\theta_{k_m})\} \\ &\quad - \sum_{m=1}^{\infty} a_{k_m} \sum_{i=\eta_{m+1}+1}^{\eta_1^{(m+1)}(\theta_{k_m})} (D_{m,i} - f(\theta_{k_m})) I\{\eta_{m+1} < \eta_1^{(m+1)}(\theta_{k_m})\}. \end{aligned} \tag{5.13}$$

As in Lemma 3.4 of [41], we can prove that each term on the right-hand side of (5.13) converges a.s.

(b) From the result of step (a), analogous to Lemma 3.5 of [41], we obtain the a.s. convergence of $\sum_{n=1}^{\infty} a_n \varepsilon_{n+1}$. □

Proof of Theorem 4.1(ii) for $L = 1$. We need the following definitions and lemma. For $n \geq 0$, let

$$\psi_{n,i} = \begin{cases} (I - a_n M_1) \cdots (I - a_i M_1) & \text{for } i \leq n, \\ I & \text{for } i = n+1, \\ 0 & \text{for } i \geq n+2, \end{cases} \tag{5.14}$$

$$\sigma(n) = \max\{m : \quad k_m \leq n\}, \quad \tau(n) = \sigma(n) + 1. \tag{5.15}$$

Lemma 5.3. *If conditions* (A0)–(A11) *are satisfied, with* δ' *satisfying* (A11), *then*

$$a_{n+1}^{\delta'-1/2} \sum_{j=0}^{n} a_j \psi_{n,j+1} \varepsilon_{j+1} \xrightarrow[n \to \infty]{} 0 \quad a.s.$$

Proof. Since $-M_1$ is stable, it is standard to derive that (see, e.g., [3] and [7])

$$\|\psi_{n,i}\| \leq c_0 \exp\left(-c \sum_{j=i}^{n} a_j\right), \quad \frac{a_i}{a_n} = \exp\left(o(1) \sum_{s=i}^{n-1} a_s\right) \quad \forall\, n \geq i, \tag{5.16}$$

$$\sup_n \sum_{i=1}^{n} a_i \exp\left(-rc \sum_{j=i+1}^{n} a_j\right) < \infty, \quad \sup_n \sum_{i=1}^{n} a_i \|\psi_{n,i+1}\|^r < \infty \quad \forall\, r > 0, \tag{5.17}$$

where c_0 and c are some constants and $o(1) \to 0$ as $i \to \infty$.

Using (2.2), (5.12), and (5.15), we have

$$\begin{aligned} &a_{n+1}^{\delta'-1/2} \sum_{j=0}^{n} a_j \psi_{n,j+1} \varepsilon_{j+1} \\ &= a_{n+1}^{\delta'-1/2} \sum_{j=k_{\sigma(n)}+1}^{n} a_j \psi_{n,j+1} \varepsilon_{j+1} \\ &\quad + a_{n+1}^{\delta'-1/2} \sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_{m+1}} a_{k_m+i-1} \psi_{n,k_m+i} (f_{k_m+i} - f(\theta_{k_m+i-1})) \\ &= a_{n+1}^{\delta'-1/2} \sum_{j=k_{\sigma(n)}+1}^{n} a_j \psi_{n,j+1} \varepsilon_{j+1} \\ &\quad + a_{n+1}^{\delta'-1/2} \sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_{m+1}} a_{k_m+i-1} \psi_{n,k_m+i} (f(\theta_{k_m}) - f(\theta_{k_m+i-1})) \\ &\quad + a_{n+1}^{\delta'-1/2} \sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_{m+1}} a_{k_m+i-1} \psi_{n,k_m+i} (f_{k_m+i} - D_{m,i}) \\ &\quad + a_{n+1}^{\delta'-1/2} \sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_{m+1}} (a_{k_m+i-1} - a_{k_m}) \psi_{n,k_m+i} (D_{m,i} - f(\theta_{k_m})) \end{aligned} \tag{5.18}$$

$$+a_{n+1}^{\delta'-1/2}\sum_{m=0}^{\sigma(n)-1}\sum_{i=1}^{\eta_{m+1}} a_{k_m}(\psi_{n,k_m+i}-\psi_{n,k_m})(D_{m,i}-f(\theta_{k_m}))$$

$$+a_{n+1}^{\delta'-1/2}\sum_{m=0}^{\sigma(n)-1} a_{k_m}\psi_{n,k_m}w_{m+1}$$

$$+a_{n+1}^{\delta'-1/2}\sum_{m=0}^{\sigma(n)-1} a_{k_m}\psi_{n,k_m}\sum_{i=\eta_1^{(m+1)}(\theta_{k_m})+1}^{\eta_{m+1}}(D_{m,i}-f(\theta_{k_m}))I\{\eta_{m+1}>\eta_1^{(m+1)}(\theta_{k_m})\}$$

$$-a_{n+1}^{\delta'-1/2}\sum_{m=0}^{\sigma(n)-1} a_{k_m}\psi_{n,k_m}\sum_{i=\eta_{m+1}+1}^{\eta_1^{(m+1)}(\theta_{k_m})}(D_{m,i}-f(\theta_{k_m}))I\{\eta_{m+1}<\eta_1^{(m+1)}(\theta_{k_m})\},$$

where

$$w_{m+1}=\sum_{i=1}^{\eta_1^{(m+1)}(\theta_{k_m})}(D_{m,i}-f(\theta_{k_m})). \tag{5.19}$$

We will show that each term on the right-hand side of the second equality in (5.18) converges to zero a.s. Then the desired result follows from (5.18).

(i) By condition (A4), there is a constant α_1 such that

$$\max_{1\le i\le\eta_{m+1}}\left(\frac{a_{k_m}}{a_{k_m+i}}\right)\le 1+\alpha_1 a_{k_m}\eta_{m+1}^* \text{ and } \max_{1\le i\le\eta_{m+1}}|a_{k_m+i}-a_{k_m}|\le\alpha_1 a_{k_m}^2\eta_{m+1}^*. \tag{5.20}$$

Using (A3), it follows from (5.12) that

$$\|\varepsilon_{k_m+i}\|\le W_{m+1}^{(0)}\overset{\triangle}{=}\max_{\theta\in D}\|f(\theta)\|+Z_{m+1}\quad\text{for}\quad 1\le i\le\eta_{m+1}. \tag{5.21}$$

By Lemma 5.1 and conditions (A3) and (A11), one has

$$\frac{1}{\sqrt{m^v}}\eta_{m+1}^*W_{m+1}^{(0)}\xrightarrow[m\to\infty]{}0\quad\text{a.s.}, \tag{5.22}$$

which yields

$$\sqrt{a_{k_{\sigma(n)}}}\eta_{\sigma(n)+1}(\theta^*)W_{\sigma(n)+1}^{(0)}\xrightarrow[n\to\infty]{}0\quad\text{a.s.} \tag{5.23}$$

By (5.16), (5.20), and (5.21), we have

$$\frac{a_{n+1}^{\delta'}}{\sqrt{a_{n+1}}}\left\|\sum_{j=k_{\sigma(n)}+1}^{n}a_j\psi_{n,j+1}\varepsilon_{j+1}\right\|\le\frac{a_{n+1}^{\delta'}}{\sqrt{a_{n+1}}}c_0a_{k_{\sigma(n)}}W_{\sigma(n)+1}^{(0)}\eta_{\sigma(n)+1}(\theta^*)$$

$$\le c_0a_{n+1}^{\delta'}\left(\frac{a_{k_{\sigma(n)}}}{a_{k_{\sigma(n)+1}}}\right)^{1/2}a_{k_{\sigma(n)}}^{1/2}\eta_{\sigma(n)+1}(\theta^*)W_{\sigma(n)+1}^{(0)}$$

$$\le c_0a_{n+1}^{\delta'}(1+\alpha_1a_{k_{\sigma(n)}}\eta_{\sigma(n)+1}(\theta^*))^{1/2}\sqrt{a_{k_{\sigma(n)}}}\eta_{\sigma(n)+1}(\theta^*)W_{\sigma(n)+1}^{(0)}\xrightarrow[n\to\infty]{}0\quad\text{a.s.},$$

which implies that the first term on the right-hand side of the second equality in (5.18) converges to zero as $n \to \infty$ a.s.

(ii) Using (A3), it follows from (1.3) and (2.2) that

$$\|\theta_{k_m+i} - \theta_{k_m}\| \leq \sum_{j=1}^{i} \|\theta_{k_m+j} - \theta_{k_m+j-1}\| \leq a_{k_m} W_{m+1} \quad \text{for } 1 \leq i \leq \eta_{m+1}, \tag{5.24}$$

where $W_{m+1} = \eta^*_{m+1} Z_{m+1}$.

By (A7), (5.20), and (5.24), one has

$$\begin{aligned} & a_{n+1}^{\delta'-1/2} \left\| \sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_{m+1}} a_{k_m+i-1} \psi_{n,k_m+i} (f(\theta_{k_m}) - f(\theta_{k_m+i-1})) \right\| \\ & \leq B_1 a_{n+1}^{\delta'} \sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_{m+1}} a_{k_m+i-1} \|\psi_{n,k_m+i}\| \frac{1}{\sqrt{a_{n+1}}} a_{k_m} W_{m+1} \\ & \leq B_1 a_{n+1}^{\delta'} \sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_{m+1}} a_{k_m+i-1} \|\psi_{n,k_m+i}\| \left(\frac{a_{k_m+i}}{a_{n+1}} \right)^{1/2} \\ & \qquad \cdot \max_{1 \leq i \leq \eta_{m=1}} \left(\frac{a_{k_m}}{a_{k_m+i}} \right)^{1/2} \sqrt{a_{k_m}} W_{m+1} \\ & \leq B_1 c_0 a_{n+1}^{\delta'} \sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_{m+1}} a_{k_m+i-1} \exp\left(-c \sum_{j=k_m+i}^{n} a_j \right) \exp\left(o(1) \sum_{j=k_m+i}^{n} a_j \right) \\ & \qquad \cdot (1 + \alpha_1 a_{k_m} \eta^*_{m+1})^{1/2} \sqrt{a_{k_m}} W_{m+1} \\ & \xrightarrow[n \to \infty]{} 0 \quad \text{a.s.,} \end{aligned} \tag{5.25}$$

since by (5.16)–(5.17),

$$\begin{aligned} & \sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_{m+1}} a_{k_m+i-1} \exp\left(-c \sum_{j=k_m+i}^{n} a_j \right) \exp\left(o(1) \sum_{j=k_m+i}^{n} a_j \right) \\ & \leq \sum_{i=0}^{n} a_i \exp\left(-c \sum_{j=i+1}^{n} a_j \right) \exp\left(o(1) \sum_{j=i+1}^{n} a_j \right) \;=\; O(1) \quad \text{a.s.,} \end{aligned}$$

and by (5.22)

$$(1 + \alpha_1 a_{k_m} \eta^*_{m+1})^{1/2} \sqrt{a_{k_m}} W_{m+1} \xrightarrow[m \to \infty]{} 0 \quad \text{a.s.}$$

Thus, the second term on the right-hand side of the second equality in (5.18) converges to zero a.s. as $n \to \infty$.

(iii) By condition (A9) one has

$$\|f_{k_m+i} - D_{m,i}\| \leq a_{k_m} Z^{(0)}_{m+1} \quad \text{a.s.} \tag{5.26}$$

for $1 \le i \le \eta_{m+1}$. Similar to (5.22), by Lemma 5.1 it follows that

$$\sqrt{a_{k_m}} Z^{(0)}_{m+1} \xrightarrow[m \to \infty]{} 0 \quad \text{a.s.} \tag{5.27}$$

By the same argument as in (5.25), it is seen that

$$\begin{aligned} & a_{n+1}^{\delta'-1/2} \left\| \sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_{m+1}} a_{k_m+i-1} \psi_{n,k_m+i} (f_{k_m+i} - D_{m,i}) \right\| \\ & \le a_{n+1}^{\delta'} \sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_{m+1}} a_{k_m+i-1} \|\psi_{n,k_m+i}\| \left(\frac{a_{k_m+i}}{a_{n+1}} \right)^{1/2} \max_{1 \le i \le \eta_{m+1}} \left(\frac{a_{k_m}}{a_{k_m+i}} \right)^{1/2} \\ & \qquad \cdot \sqrt{a_{k_m}} Z^{(0)}_{m+1} \\ & \le c_0 a_{n+1}^{\delta'} \sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_{m+1}} a_{k_m+i-1} \exp\left(-c \sum_{j=k_m+i}^{n} a_j \right) \exp\left(o(1) \sum_{j=k_m+i}^{n} a_j \right) \\ & \qquad \cdot (1 + \alpha_1 a_{k_m})^{1/2} \sqrt{a_{k_m}} Z^{(0)}_{m+1} \\ & \xrightarrow[n \to \infty]{} 0 \quad \text{a.s.,} \end{aligned} \tag{5.28}$$

which implies that the third term on the right-hand side of the second inequality in (5.18) converges to zero as $n \to \infty$ a.s.

(iv) Analogous to (5.25), by (5.20) it is derived that the fourth term on the right-hand side of the second inequality in (5.18) converges to zero as $n \to \infty$ a.s.

(v) By the definition (5.14), we have

$$\begin{aligned} \psi_{n,k_m+i} - \psi_{n,k_m} &= \sum_{j=1}^{i} a_{k_m+j-1} \psi_{n,k_m+j} M_1 \\ &= \psi_{n,k_m+i} \sum_{j=1}^{i} a_{k_m+j-1} \psi_{k_m+i-1,k_m+j} M_1. \end{aligned} \tag{5.29}$$

Similar to (5.25), we arrive at the conclusion that the fifth term on the right-hand side of the second inequality in (5.18) converges to zero as $n \to \infty$ a.s.

(vi) We now consider the convergence of the sixth term. First, by the definition (5.14),

$$\sum_{m=0}^{\tau(n)} \psi_{n,k_m} a_{k_m} w_{m+1} = \sum_{m=0}^{n+1} a_{k_m} \psi_{n,k_m} w_{m+1}. \tag{5.30}$$

Let $h(m) = m^\zeta \ \forall \ m \ge 1$, where ζ is some constant satisfying $0 < \zeta < \delta' \nu$. Then by condition (A11) and Lemma 2.1 we have

$$\frac{\|w_{m+1}\|}{h(m)} \le \frac{W^{(0)}_{m+1} \eta^*_{m+1}}{m^\zeta} \xrightarrow[m \to \infty]{} 0 \quad \text{a.s.} \tag{5.31}$$

By (5.16)–(5.17), it is derived that

$$
\begin{aligned}
(5.32)\qquad \frac{1}{a_{n+1}}\sum_{m=0}^{n+1} a_{k_m}^2\|\psi_{n,k_m}\|^2 &\le \frac{1}{a_{n+1}}\sum_{m=0}^{n+1} a_{k_m-1}a_{k_m}\|\psi_{n,k_m}\|^2\\
&\le \frac{1}{a_{n+1}}\sum_{i=0}^{n+1} a_{i-1}\|\psi_{n,i}\|^2 a_i\\
&\le c_0^2\sum_{i=0}^{n+1}\exp\left(-2c\sum_{j=i}^{n}a_j\right)a_{i-1}\exp\left(o(1)\sum_{j=i}^{n}a_j\right)=O(1),
\end{aligned}
$$

which, incorporating with (5.31) and Lemma 5.2, leads to

$$
(5.33)\qquad \frac{1}{\sqrt{a_{n+1}}}\sum_{m=0}^{n+1}\psi_{n,k_m}a_{k_m}w_{m+1}=O(1)+o(h(n)\log n).
$$

By (5.33), it follows from (5.30) that

$$
\begin{aligned}
a_{n+1}^{\delta'-1/2}\sum_{m=0}^{\tau(n)}a_{k_m}\psi_{n,k_m}w_{m+1} &= O(a_{n+1}^{\delta'})+o\left(a_{n+1}^{\delta'}h(n)\log n\right)\\
&= o\left(\frac{1}{n^{\delta'\nu-\zeta}}\log n\right)=o(1)\quad \text{a.s.},
\end{aligned}
$$

which implies that $a_{n+1}^{\delta'-1/2}\sum_{m=0}^{\sigma(n)-1}a_{k_m}\psi_{n,k_m}w_{m+1}\to 0$ a.s. as $n\to\infty$, since by (5.16), (5.20), and (5.22)–(5.23) we have

$$
\begin{aligned}
&\frac{a_{n+1}^{\delta}}{\sqrt{a_{n+1}}}\|\psi_{n,k_{\sigma(n)}}a_{k_{\sigma(n)}}w_{\sigma(n)+1}+\psi_{n,k_{\tau(n)}}a_{k_{\tau(n)}}w_{\tau(n)+1}\|\\
&\le c_0a_{n+1}^{\delta}\Big((1+\alpha_1a_{k_{\sigma(n)}}\eta^*_{\sigma(n)+1})^{1/2}\sqrt{a_{k_{\sigma(n)}}}\eta^*_{\sigma(n)+1}W^{(0)}_{\sigma(n)+1}\\
&\qquad+\left(\frac{a_n}{a_{n+1}}\right)^{1/2}\sqrt{a_{k_{\tau(n)}}}\eta^*_{\tau(n)+1}W^{(0)}_{\tau(n)+1}\Big)\xrightarrow[n\to\infty]{}0\quad\text{a.s.}
\end{aligned}
$$

(vii) Set

$$
V_{m+1}^{(0)}=\sum_{i=\eta_1^{(m+1)}(\theta_{k_m})+1}^{\eta_{m+1}}(D_{m,i}-f(\theta_{k_m}))I\{\eta_{m+1}>\eta_1^{(m+1)}(\theta_{k_m})\}\quad\forall\, m\ge 0.
$$

By (5.31), we get

$$
(5.34)\qquad \frac{V_m^{(0)}-\mathsf{E}[V_m^{(0)}|\mathcal{F}^{(m-1)}]}{h(m)}\xrightarrow[m\to\infty]{}0\quad\text{a.s.},
$$

which incorporating with (5.32) and Lemma 5.2 yields

$$
\begin{aligned}
(5.35)\qquad &\frac{1}{\sqrt{a_{n+1}}}\sum_{m=0}^{n+1}\psi_{n,k_m}a_{k_m}(V_{m+1}^{(0)}-\mathsf{E}[V_{m+1}^{(0)}|\mathcal{F}^{(m)}])\\
&=O(1)+o(h(n)\log n)=o(h(n)\log n)\quad\text{a.s.}
\end{aligned}
$$

By Hölder's inequality and condition (A10) for any $p_1>1$ we have

$$\begin{aligned}(5.36)\qquad &\|\mathsf{E}[V_{m+1}^{(0)}|\mathcal{F}^{(m)}]\|\\ &\leq \mathsf{E}[\eta_{m+1}^{*}W_{m+1}^{(0)}I\{\eta_{m+1}>\eta_1^{(m+1)}(\theta_{k_m})\}|\mathcal{F}^{(m)}]\\ &\leq \left(\mathsf{E}\left[(W_{m+1}^{(0)}\eta_{m+1}^{*})^{p_1}\right]\right)^{1/p_1}P\{\eta_{m+1}>\eta_1^{(m+1)}(\theta_{k_m})|\mathcal{F}^{(m)}\}^{1-1/p_1}\\ &\leq \alpha_0^{1-1/p_1}\left(\mathsf{E}\left[(W_{m+1}^{(0)}\eta_{m+1}^{*})^{p_1}\right]\right)^{1/p_1}a_{k_m}^{\gamma_1(1-1/p_1)},\end{aligned}$$

which leads to

$$\begin{aligned}(5.37)\qquad &\frac{1}{a_{n+1}^{1/2-\delta'}}\left\|\sum_{m=0}^{n+1}\psi_{n,k_m}a_{k_m}\mathsf{E}[V_{m+1}^{(0)}|\mathcal{F}^{(m)}]\right\|\\ &\leq \alpha_0^{1-1/p_1}\sup_m\left\{\left(\mathsf{E}\left[(W_{m+1}^{(0)}\eta_{m+1}^{*})^{p_1}\right]\right)^{1/p_1}\right\}\\ &\cdot\frac{1}{a_{n+1}^{1/2-\delta'}}\sum_{m=0}^{n+1}\|\psi_{n,k_m}\|a_{k_m-1}a_{k_m}^{\gamma_1(1-1/p_1)}\\ &\leq O(1)a_{n+1}^{\delta'}\sum_{i=-1}^{n+1}a_i\exp\left(-c\sum_{j=i+1}^{n}a_j\right)\left(\frac{a_{i+1}}{a_{n+1}}\right)^{1/2}a_{i+1}^{\gamma_1(1-1/p_1)-1/2}\\ &=o(1)\quad\text{a.s.,}\end{aligned}$$

provided that $\gamma_1(1-1/p_1)\geq 1/2$. Combining (5.35) and (5.37) gives

$$\frac{1}{a_{n-1}^{1/2-\delta'}}\left\|\sum_{m=0}^{n+1}\psi_{n,k_m}a_{k_m}V_{m+1}^{(0)}\right\|=o(1)+o(a_{n+1}^{\delta'}h(n)\log n)=o(1)\quad\text{a.s.,}$$

where $0<\zeta<\delta'\nu$.

(viii) In the same way as in (vii), one shows that the last term on the right-hand side of the second equality in (5.18) converges to zero a.s. □

Proof of Theorem 4.1(ii) for $L = 1$. By Theorem 4.1(i), there is an a.s. finite r.v. n_0^* such that no projection occurs after step n_0^*; i.e.,

$$(5.38)\qquad \theta_{n+1}=\theta_n-a_n(f(\theta_n)+\varepsilon_{n+1})\quad\forall\ n\geq n_0^*.$$

Then, for a deterministic integer n_0, on the event $\{n\geq n_0\geq n_0^*\}$, one has

$$\begin{aligned}(5.39)\qquad \frac{\theta_{n+1}-\theta^0}{a_{n+1}^{1/2-\delta'}}&=a_{n+1}^{\delta'-1/2}\psi_{n,n_0}(\theta_{n_0}-\theta^0)-a_{n+1}^{\delta'-1/2}\sum_{j=n_0}^{n}\psi_{n,j+1}a_j\varepsilon_{j+1}\\ &-a_{n+1}^{\delta'-1/2}\sum_{j=n_0}^{n}\psi_{n,j+1}a_j(f(\theta_j)-M_1(\theta_j-\theta^0)).\end{aligned}$$

Let $r\leq r_0$, where r_0 is given by condition (A5). Define

$$(5.40)\qquad \sigma^*=\begin{cases}0 & \text{if }\ \|\theta_{n_0}-\theta^0\|\geq r,\\ \inf\{j:\ \ j>n_0,\ \|\theta_j-\theta^0\|\geq r\} & \text{otherwise.}\end{cases}$$

By (5.16), it is easy to see that

$$\begin{aligned}(5.41)\qquad &\|a_{n+1}^{\delta'-1/2}\psi_{n,n_0}(\theta_{n_0}-\theta^0)\|\\ &\le c_0 a_{n_0}^{\delta'-1/2}\left(\frac{a_{n_0}}{a_{n+1}}\right)^{1/2-\delta'}\exp\left(-c\sum_{j=n_0}^{n}a_j\right)\|\theta_{n_0}-\theta^0\|\\ &\le O(1)\exp\left(-c\sum_{j=n_0}^{n}a_j\right)\exp\left(o(1)\sum_{j=n_0}^{n}a_j\right)\\ &=o(1)\quad\text{as}\quad n\to\infty.\end{aligned}$$

Let c_0^* be a constant such that

$$\sup_{j\ge1}\left\{\left(\frac{a_j}{a_{j+1}}\right)^{1/2-\delta'}\right\}\le c_0^*.$$

By condition (A5) and (5.16), we have

$$\begin{aligned}(5.42)\quad & a_{n+1}^{\delta'-1/2}\left\|\sum_{j=n_0}^{n}\psi_{n,j+1}a_j(f(\theta_j)-M_1(\theta_j-\theta^0))I\{\sigma^*>n+1,\ n_0^*\le n_0\}\right\|\\ &\le c_0c_1a_{n+1}^{\delta'-1/2}\sum_{j=n_0}^{n}\exp\left(-c\sum_{s=j+1}^{n}a_s\right)a_j\|\theta_j-\theta^0\|^2I\{\sigma^*>j,\ n_0^*\le n_0\}\\ &\le c_0c_1r\sum_{j=n_0}^{n}\exp\left(-c\sum_{s=j+1}^{n}a_s\right)a_j\left(\frac{a_j}{a_{n+1}}\right)^{1/2-\delta'}\frac{\|\theta_j-\theta^0\|I\{\sigma^*>j,\ n_0^*\le n_0\}}{a_j^{1/2-\delta'}}\\ &\le c_0^*c_0c_1r\sum_{j=n_0}^{n}\exp\left(-\frac{c}{2}\sum_{s=j+1}^{n}a_s\right)a_j\frac{\|\theta_j-\theta^0\|I\{\sigma^*>j,\ n_0^*\le n_0\}}{a_j^{1/2-\delta'}}\end{aligned}$$

if n_0 is large enough.

By Lemma 5.3 and (5.41)–(5.42), it follows from (5.39) that

$$\begin{aligned}&\|\theta_{n+1}-\theta^0\|I\{\sigma^*>n+1,\ n_0^*\le n_0\}a_{n+1}^{\delta'-1/2}\\ &\le o(1)+c_0^*c_0c_1r\sum_{j=n_0}^{n}\exp\left(-\frac{c}{2}\sum_{s=j+1}^{n}a_s\right)a_j\\ (5.43)\qquad &\cdot\|\theta_j-\theta^0\|^2I\{\sigma^*>j,\ n_0^*\le n_0\}a_j^{\delta'-1/2},\end{aligned}$$

which, incorporating the Bellman–Gronwall inequality, leads to

$$\begin{aligned}(5.44)\qquad &\|\theta_n-\theta^0\|I\{\sigma^*>n,\ n_0^*\le n_0\}a_n^{\delta'-1/2}\\ &\le o(1)+o(1)\sum_{j=n_0}^{n-1}\exp\left(-\frac{c}{2}\sum_{s=j+1}^{n}a_s\right)a_j\exp\left(c_0^*c_0c_1r\sum_{s=j+1}^{n-1}a_s\right)\end{aligned}$$

$$\leq o(1) + o(1) \sum_{j=n_0}^{n-1} a_j \exp\left(-\frac{c}{4} \sum_{s=j+1}^{n-1} a_s \right) \; = \; o(1),$$

where r is sufficiently small such that $c_0^* c_0 c_1 r \leq c/4$. From (5.44), it is readily seen that $\|\theta_n - \theta^0\| = o(a_n^{1/2-\delta'})$ as $n \to \infty$ a.s. □

Proof of Theorem 4.1(iii) for $L = 1$. In order to prove Theorem 4.1(iii), we need several lemmas. From (5.14), we have

$$\psi_{n,j} = \psi_{n-1,j} - a_n M_1 \psi_{n-1,j} = I - M_1 \sum_{i=j}^{n} a_i \psi_{i-1,j},$$

which gives

$$(5.45) \quad a_{j-1} \sum_{i=j}^{n} \psi_{i-1,j} = \sum_{i=j}^{n} a_i \psi_{i-1,j} + \sum_{i=j}^{n} (a_{j-1} - a_i) \psi_{i-1,j} = M_1^{-1} + G_{n,j},$$

where

$$(5.46) \qquad G_{n,j} \overset{\triangle}{=} -M_1^{-1} \psi_{n,j} + \sum_{i=j}^{n} (a_{j-1} - a_i) \psi_{i-1,j} \quad \forall\ n \geq j.$$

LEMMA 5.4. *Let the conditions of Theorem* 4.1(iii) *be satisfied. Then*

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{j=1}^{n-1} (M_1^{-1} + G_{n,j+1})(f(\theta_j) - M_1(\theta_j - \theta^0)) = 0 \quad a.s.$$

Proof. For any $\delta' \in (0,\ 1/2(1 - 1/(2\nu)))$, by condition (A4) and Theorem 4.1(ii) it follows that

$$\sum_{i=1}^{\infty} \frac{1}{\sqrt{i}} \|\theta_i - \theta^0\|^2 = \sum_{i=1}^{\infty} \frac{1}{\sqrt{i}} o(a_i^{1-2\delta'}) = \sum_{i=1}^{\infty} \frac{1}{\sqrt{i}} o\left(\frac{1}{i^{\nu(1-2\delta')}} \right) < \infty \ \text{ a.s.},$$

which gives

$$(5.47) \qquad \lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{j=1}^{n-1} \|\theta_j - \theta^0\|^2 = 0 \quad \text{a.s.}$$

via the Kronecker lemma (see, e.g., [11]).

It is shown in [3, Lemma 1] that for all $n \geq j \geq 1$, $G_{n,j}$ defined by (5.46) are bounded. From Theorem 4.1, condition (A5), and (5.47), the lemma follows easily. □

LEMMA 5.5. *Suppose that conditions* (A0)–(A11) *are satisfied. Then we have that* $1/\sqrt{n} \sum_{i=1}^{n} \varepsilon_i \to N(0,\ S_4)$ *in distribution as* $n \to \infty$*, where* $S_4 = \overline{\eta}(\theta^0)^{-1} S$.

Proof. As in (5.13), we decompose

$$
\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i
&= \frac{1}{\sqrt{n}}\sum_{j=k_{\sigma(n)}+1}^{n}\varepsilon_j + \frac{1}{\sqrt{n}}\sum_{m=0}^{\sigma(n)-1}\sum_{i=1}^{\eta_{m+1}}(f_{k_m+i}-D_{m,i})\\
&\quad+\frac{1}{\sqrt{n}}\sum_{m=0}^{\sigma(n)-1}\sum_{i=1}^{\eta_{m+1}}(f(\theta_{k_m})-f(\theta_{k_m+i-1}))\\
&\quad+\frac{1}{\sqrt{n}}\sum_{m=0}^{\sigma(n)-1}\sum_{i=\eta_1^{(m+1)}(\theta_{k_m})+1}^{\eta_{m+1}}(D_{m,i}-f(\theta_{k_m}))I\{\eta_{m+1}>\eta_1^{(m+1)}(\theta_{k_m})\}\\
&\quad-\frac{1}{\sqrt{n}}\sum_{m=0}^{\sigma(n)-1}\sum_{i=\eta_{m+1}+1}^{\eta_1^{(m+1)}(\theta_{k_m})}(D_{m,i}-f(\theta_{k_m}))I\{\eta_{m+1}<\eta_1^{(m+1)}(\theta_{k_m})\}\\
&\quad+\frac{1}{\sqrt{n}}\sum_{m=0}^{\sigma(n)-1}w_{m+1},
\end{aligned} \tag{5.48}
$$

where $\sigma(n)$ is defined by (5.15) and w_{m+1} is defined by (5.19).

Initially, we show that each of the first five terms on the right-hand side of (5.48) converges to zero a.s. as $n\to\infty$. Then we show that the last term converges to a $N(0,\ S_4)$ in distribution. Hence the desired result follows.

(i) Similar to (5.23), it is easy to show that the first term on the right-hand side of (5.48) converges to zero as $n\to\infty$ a.s.

(ii) By (5.26) it is derived that

$$
\begin{aligned}
\sum_{m=1}^{\infty}\frac{1}{\sqrt{m}}\mathsf{E}\left[\left\|\sum_{i=1}^{\eta_{m+1}}(f_{k_m+i}-D_{m,i})\right\|\,\middle|\mathcal{F}^{(m)}\right] &\le \sum_{m=1}^{\infty}\frac{1}{\sqrt{m}}a_{k_m}\mathsf{E}\left[\eta^*_{m+1}Z^*_{m+1}Z^{(0)}_{m+1}\right]\\
&\le \sum_{m=1}^{\infty}\frac{\overline{a}}{m^{\nu+1/2}}\mathsf{E}\left[\eta^*_{m+1}Z^*_{m+1}Z^{(0)}_{m+1}\right]<\infty \quad \text{a.s.},
\end{aligned}
$$

which implies

$$
\sum_{m=1}^{\infty}\frac{1}{\sqrt{m}}\sum_{i=1}^{\eta_{m+1}}(f_{k_m+i}-D_{m,i})<\infty \quad \text{a.s.} \tag{5.49}
$$

by the local convergence theorem of martingales (see, e.g., [6, 20, 37]). By the Kronecker lemma, it follows from (5.49) that

$$
\lim_{n\to\infty}\frac{1}{\sqrt{n}}\sum_{m=0}^{n}\sum_{i=1}^{\eta_{m+1}}(f_{k_m+i}-D_{m,i})=0 \quad \text{a.s.},
$$

which gives

$$
\lim_{n\to\infty}\frac{1}{\sqrt{n}}\sum_{m=0}^{\sigma(n)-1}\sum_{i=1}^{\eta_{m+1}}(f_{k_m+i}-D_{m,i})=0 \quad \text{a.s.}, \tag{5.50}
$$

since $\sigma(n) \le n \ \forall \ n \ge 1$.

(iii) By (5.24), similar to (5.50) we can prove that the third term on the right-hand side of (5.48) converges to zero as $n \to \infty$ a.s.

(iv) By (5.36) and condition (A4), we derive that

$$\mathsf{E}\left\| \sum_{m=1}^{\infty} \frac{1}{\sqrt{m}} \sum_{i=\eta_1^{(m+1)}(\theta_{k_m})+1}^{\eta_{m+1}} (D_{m,i} - f(\theta_{k_m})) I\{\eta_{m+1} > \eta_1^{(m+1)}(\theta_{k_m})\} \right\| \tag{5.51}$$
$$\le \sum_{m=1}^{\infty} \frac{1}{\sqrt{m}} \alpha_0^{1-1/p_1} \left(\mathsf{E}\left[(W_{m+1}^{(0)} \eta_{m+1}^*)^{p_1} \right] \right)^{1/p_1} \overline{a}^{\gamma_1(1-1/p_1)} m^{-\gamma_1(1-1/p_1)\nu} < \infty$$

if $\nu\gamma_1(1 - 1/p_1) > 1/2$. By (5.51), it is easy to prove that the fourth term on the right-hand side of (5.48) converges to zero as $n \to \infty$ a.s.

(v) Similar to (iv), we can prove that the fifth term on the right-hand side of (5.48) converges to zero a.s. as $n \to \infty$.

(vi) By the central limit theorem for martingales (see, e.g., [11] and [20]),

$$\frac{1}{\sqrt{n}} \sum_{m=0}^{n} w_{m+1} \xrightarrow[n \to \infty]{d} N(0, \ S). \tag{5.52}$$

We now show that

$$\frac{\sigma(n)}{n} \xrightarrow[n \to \infty]{} \frac{1}{\overline{\eta}(\theta^0)} \quad \text{a.s.} \tag{5.53}$$

One can decompose

$$\frac{1}{n}\sum_{i=1}^{n} \eta_i = \frac{1}{n}\sum_{i=1}^{n} (\eta_i - \eta_1^{(i)}(\theta_{k_{i-1}})) + \frac{1}{n}\sum_{i=1}^{n} (\eta_1^{(i)}(\theta_{k_{i-1}}) - \overline{\eta}(\theta_{k_{i-1}})) \tag{5.54}$$
$$+ \frac{1}{n}\sum_{i=1}^{n} (\overline{\eta}(\theta_{k_{i-1}}) - \overline{\eta}(\theta^0)) + \overline{\eta}(\theta^0).$$

By conditions (A9) and (A10), we have

$$\mathsf{E}\left[\sum_{i=1}^{\infty} \frac{1}{i} \|\eta_i - \eta_1^{(i)}(\theta_{k_{i-1}})\| \right] \le \sum_{i=1}^{\infty} \frac{1}{i} \mathsf{E}\left[\eta_i^* I\{\eta_i \ne \eta_1^{(i)}(\theta_{k_{i-1}})\} \right]$$
$$\le \sum_{i=1}^{\infty} \frac{1}{i} \sqrt{\mathsf{E}[(\eta_1^*)^2]} \sqrt{P\{\eta_i \ne \eta_1^{(i)}(\theta_{k_{i-1}})\}} \le O(1) \sum_{i=1}^{\infty} \frac{1}{i} i^{-\gamma_1\nu/2} < \infty,$$

which, combining with the Kronecker lemma, yields $\lim_{n\to\infty} 1/n \sum_{i=1}^{n} (\eta_i - \eta_1^{(i)}(\theta_{k_{i-1}})) < \infty$ a.s. By the local convergence theorem of martingales and the Kronecker lemma, it is derived that $\lim_{n\to\infty} 1/n \sum_{i=1}^{n} (\eta_1^{(i)}(\theta_{k_{i-1}}) - \overline{\eta}(\theta_{k_{i-1}})) = 0$ a.s. By Theorem 3.1, $\theta_n \to \theta^0$ a.s. as $n \to \infty$, which implies $1/n \sum_{i=1}^{n} (\overline{\eta}(\theta_{k_{i-1}}) - \overline{\eta}(\theta^0)) \to 0$ as $n \to \infty$ a.s. via the continuity of $\overline{\eta}(\theta)$ at θ^0. Thus it follows from (5.54) that

$$\frac{1}{n}\sum_{i=1}^{n} \eta_i \xrightarrow[n \to \infty]{} \overline{\eta}(\theta^0) \quad \text{a.s.} \tag{5.55}$$

From the definition (5.15), we have $n < \sum_{i=1}^{\sigma(n)+1} \eta_i \le n + \eta_{\sigma(n)+1}$, which gives

$$1 < \left(\frac{\sigma(n)}{n} + \frac{1}{n}\right) \frac{1}{\sigma(n)+1} \sum_{i=1}^{\sigma(n)+1} \eta_i \le 1 + \frac{\sigma(n)+1}{n} \frac{1}{\sigma(n)+1} \eta^*_{\sigma(n)+1}. \tag{5.56}$$

This yields (5.53) by (5.55) and Lemma 5.1.

To prove

$$\frac{1}{\sqrt{n}} \sum_{m=0}^{\sigma(n)-1} w_{m+1} \xrightarrow[n \to \infty]{} N(0,\ S_4), \tag{5.57}$$

we need a central limit theorem for stopped martingales. We note that if the Kolmogorov inequality is replaced by Doob's inequality (see, e.g., [11]), the proof of Theorem 9.4.1 in [11] goes through for the stopped martingales. Then, by (5.52) and (5.53), (5.57) follows. □

LEMMA 5.6. *If the conditions of Theorem* 4.1(iii) *are fulfilled, then we have that* $1/\sqrt{n}\sum_{j=1}^{n} G_{n,j}\varepsilon_j \to 0$ *in probability as* $n \to \infty$.

Proof. Similar to (5.48), we have

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_{j=1}^{n} G_{n,j}\varepsilon_j \\
&= \frac{1}{\sqrt{n}} \sum_{j=k_{\sigma(n)}+1}^{n} G_{n,j}\varepsilon_j + \frac{1}{\sqrt{n}} \sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_{m+1}} G_{n,k_m+i}(f_{k_m+i} - D_{m,i}) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_{m+1}} G_{n,k_m+i}(f(\theta_{k_m}) - f(\theta_{k_m+i-1})) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{m=0}^{\sigma(n)-1} \sum_{i=\eta_1^{(m+1)}(\theta_{k_m})+1}^{\eta_{m+1}} G_{n,k_m+i}(D_{m,i} - f(\theta_{k_m})) I\{\eta_{m+1} > \eta_1^{(m+1)}(\theta_{k_m})\} \\
&\quad - \frac{1}{\sqrt{n}} \sum_{m=0}^{\sigma(n)-1} \sum_{i=\eta_{m+1}+1}^{\eta_1^{(m+1)}(\theta_{k_m})} G_{n,k_m+i}(D_{m,i} - f(\theta_{k_m})) I\{\eta_{m+1} < \eta_1^{(m+1)}(\theta_{k_m})\} \\
&\quad + \frac{1}{\sqrt{n}} \sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_1^{(m+1)}(\theta_{k_m})} G_{n,k_m}(D_{m,i} - f(\theta_{k_m})) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_1^{(m+1)}(\theta_{k_m})} (G_{n,k_m+i} - G_{n,k_m})(D_{m,i} - f(\theta_{k_m})).
\end{aligned} \tag{5.58}$$

Since $G_{n,j}$ is bounded for $n \ge j \ge 1$, it follows from the proof of Lemma 5.5 that each of the first five terms on the right side of (5.58) converges to zero a.s. as $n \to \infty$. In what follows we prove that as $n \to \infty$ the sixth term converges to zero in probability, while the last term converges to zero a.s.

(i) By the definition (5.46), we get

$$\mathsf{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{m=0}^{\tau(n)} G_{n,k_m} w_{m+1}\right\|^2\right] \le \frac{1}{n}\mathsf{E}\left[\sum_{m=0}^{\tau(n)} \|G_{n,k_m}\|^2 \mathsf{E}[\|w_{m+1}\|^2 | \mathcal{F}^{(m)}]\right] \tag{5.59}$$

$$\le \mathsf{E}\left[(\eta_1^* W_1^{(0)})^2\right] \frac{1}{n}\sum_{i=0}^{n+1} \|G_{n,i}\|^2 \xrightarrow[n\,\to\,\infty]{} 0$$

via Lemma 1 of [3]. By (5.59), it is easy to derive that $1/\sqrt{n}\sum_{m=0}^{\sigma(n)-1} G_{n,k_m} w_{m+1} \to 0$ in probability as $n \to \infty$.

(ii) By (5.14) and (5.45), we have

$$\|G_{n,j} - G_{n,j-1}\| \tag{5.60}$$

$$= \left\|(a_{j-1} - a_{j-2})\sum_{s=j}^{n} \psi_{s-1,j} + a_{j-2}\sum_{s=j}^{n}(\psi_{s-1,j} - \psi_{s-1,j-1}) - a_{j-2}\psi_{j-2,j-1}\right\|$$

$$\le \alpha_1 a_{j-1} a_{j-2} \sum_{s=j}^{n} \|\psi_{s-1,j}\| + a_{j-2}\sum_{s=j}^{n} a_{j-1}\|M_1\| \; \|\psi_{s-1,j}\| + a_{j-2}$$

$$\le a_{j-1}\left((\alpha_1 + \|M_1\|)\sup_j\left\{\frac{a_{j-2}}{a_{j-1}}\right\} \; \sup_j\left\{a_{j-1}\sum_{s=j}^{n}\|\psi_{s-1,j}\|\right\} + \sup_j\left\{\frac{a_{j-2}}{a_{j-1}}\right\}\right)$$

$$\le c_3 a_{j-1} \qquad \forall\; n \ge j \ge 1,$$

where α_1 and c_3 are some constants. By (5.60) and (5.26), it follows that

$$\frac{1}{\sqrt{n}}\left\|\sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_1^{(m+1)}(\theta_{k_m})} (G_{n,k_m+i} - G_{n,k_m})(D_{m,i} - f(\theta_{k_m}))\right\|$$

$$\le \frac{1}{\sqrt{n}}\sum_{m=0}^{\sigma(n)-1} \sum_{i=1}^{\eta_1^{(m+1)}(\theta_{k_m})} \left\|\sum_{j=1}^{i}(G_{n,k_m+j} - G_{n,k_m+j-1})\right\| \cdot \|D_{m,i} - f(\theta_{k_m})\|$$

$$\le \frac{c_3}{\sqrt{n}}\sum_{m=0}^{\sigma(n)-1} a_{k_m}(\eta_{m+1}^*)^2 W_{m+1}^{(0)} \xrightarrow[n\,\to\,\infty]{} 0 \quad \text{a.s.}$$

via similar arguments as for (5.50). Thus, the proof of Lemma 5.6 is completed. □

Proof of Theorem 4.1(iii) for $L = 1$. By (5.45), from (5.38) it follows that

$$\sqrt{n}(\overline{\theta}_n - \theta^0) = o(1) + \frac{1}{\sqrt{n}}\sum_{i=n_0^*}^{n}(\theta_i - \theta^0) \tag{5.61}$$

$$= o(1) + \frac{1}{\sqrt{n}}\frac{1}{a_{n_0^*-1}}(M_1^{-1} + G_{n,n_0^*})(\theta_{n_0^*} - \theta^0)$$

$$-\frac{1}{\sqrt{n}}\sum_{j=n_0^*}^{n-1}(M_1^{-1} + G_{n,j+1})(f(\theta_j) - M_1(\theta_j - \theta^0))$$

$$-\frac{M_1^{-1}}{\sqrt{n}}\sum_{j=n_0^*}^{n-1}\varepsilon_{j+1}-\frac{1}{\sqrt{n}}\sum_{j=n_0^*}^{n-1}G_{n,j+1}\varepsilon_{j+1}.$$

Since the $G_{n,j}$ are bounded uniformly for $n \geq j \geq 1$, one has

$$\frac{1}{\sqrt{n}}\frac{1}{a_{n_0^*-1}}(M_1^{-1}+G_{n,n_0^*})(\theta_{n_0^*}-\theta^0)\xrightarrow[n\to\infty]{}0 \quad \text{a.s.} \tag{5.62}$$

By (5.62) and Lemmas 5.4–5.6, the desired result follows from (5.61). □

Proof of Theorem 4.1 for $L > 1$. Let $\widetilde{\varepsilon}_n = \beta_n - f(\widetilde{\theta}_{n-1})$ $\forall$ $n \geq 1$. The proof for the case $L = 1$ applies to the present setting if we replace θ_n, ε_n, f_n by $\widetilde{\theta}_n$, $\widetilde{\varepsilon}_n$, and β_n, respectively (cf. [40] and [41]). We mention only the key point for Theorem 4.1(iii). By the definitions (5.12) and (2.2),

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i=\frac{1}{\sqrt{n}}\sum_{i=0}^{n}\left(\frac{1}{L}\sum_{j=1}^{L}\beta_{iL+j}-f(\theta_{i-1})\right) \tag{5.63}$$

$$=\frac{1}{L\sqrt{n}}\sum_{i=1}^{(n+1)L}(\beta_i-f(\widetilde{\theta}_{i-1})) \;=\; \frac{1}{L\sqrt{n}}\sum_{i=1}^{(n+1)L}\widetilde{\varepsilon}_i.$$

As in the proof of Lemma 5.5, one can show that $1/\sqrt{n}\sum_{i=1}^{n}\widetilde{\varepsilon}_i \to N(0,\ S_4)$ in distribution as $n\to\infty$. Combined with (5.63), this gives $1/\sqrt{n}\sum_{i=1}^{n}\varepsilon_i \to N(0,\ S_4^*)$ in distribution as $n\to\infty$, where $S_4^* = L^{-1}S_4 = L^{-1}\overline{\eta}(\theta^0)^{-1}S$. The rest of the proof works the same way as for $L = 1$. □

REFERENCES

[1] S. Asmussen, *Applied Probability and Queues*, John Wiley, New York, 1987.

[2] K. B. Athreya and P. Ney, *A new approach to the limit theorems of recurrent Markov chains*, Trans. Amer. Math. Soc., 245 (1978), pp. 493–501.

[3] H. F. Chen, *Asymptotic efficient stochastic approximation*, Stochastics Stochastics Rep., 45 (1993), pp. 1–16.

[4] H. F. Chen, *Stochastic approximation and its new applications*, in Proceedings, 1994 Hong Kong International Workshop on New Directions in Control and Manufacturing, Hong Kong, 1994, pp. 2–12.

[5] H. F. Chen, *Continuous-time stochastic approximation: Convergence and asymptotic efficiency*, Stochastics Stochastics Rep., 51 (1994), pp. 111–132.

[6] H. F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*, Birkhäuser, Cambridge, MA, 1991.

[7] H. F. Chen and Y. M. Zhu, *Stochastic Approximation*, Shanghai Scientific and Technological Publishers, 1996.

[8] E. K. P. Chong and P. J. Ramadge, *Convergence of recursive optimization algorithms using infinitesimal perturbation analysis estimates*, Discrete Event Dynamic Systems: Theory and Applications, 1 (1992), pp. 339–372.

[9] E. K. P. Chong and P. J. Ramadge, *Optimization of queues using an infinitesimal perturbation analysis-based stochastic algorithm with general updates times*, SIAM J. Control Optim., 31 (1993), pp. 698–732.

[10] E. K. P. Chong and P. J. Ramadge, *Stochastic optimization of regenerative systems using infinitesimal perturbation analysis*, IEEE Trans. Automat. Control, 39 (1994), pp. 1400–1410.

[11] Y. S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingale*, 2nd ed., Springer-Verlag, Berlin, New York, 1988.

[12] P. Dupuis and H. J. Kushner, *Stochastic approximations via large deviations: Asymptotic properties*, SIAM J. Control Optim., 23 (1985), pp. 675–696.

[13] P. Dupuis and R. Simha, *On sampling controlled stochastic approximation*, IEEE Trans. Automat. Control, 36 (1991), pp. 915–924.

[14] M. C. Fu, *Convergence of a stochastic approximation algorithm for the GI/G/*1 *queue using infinitesimal perturbation analysis*, J. Optim. Theory Appl., 65 (1990), pp. 149–160.

[15] P. Glasserman, *Gradient Estimation via Perturbation Analysis*, Kluwer Academic, Boston, 1991.

[16] P. Glasserman, *Stationary waiting time derivatives*, QUESTA, 12 (1992), pp. 369–390.

[17] P. Glasserman, *Regenerative derivatives of regenerative sequences*, Adv. Appl. Probab., 25 (1993), pp. 116–139.

[18] P. Glasserman, J. Q. Hu, and S. G. Strickland, *Strongly consistent steady-state derivative estimates*, Probab. Engrg. Inform. Sci., 5 (1991), pp. 391–413.

[19] L. Guo, D. W. Huang, and E. J. Hannan, *On ARX (∞) approximation*, J. Multivariate Anal., 32 (1990), pp. 17–47.

[20] P. Hall and C. C. Heyde, *Martingale Limit Theory and Its Application*, Academic Press, New York, 1980.

[21] Y. C. Ho and X. R. Cao, *Perturbation Analysis of Discrete Event Dynamic Systems*, Kluwer Academic, Boston, 1991.

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

[23] H. J. Kushner and F. J. Vázquez-Abad, *Stochastic approximation methods for systems of interest over an infinite horizon*, SIAM J. Control Optim., 34 (1996), pp. 712–756.

[24] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer, New York, 1997.

[25] P. L'Ecuyer, *A unified view of the IPA, SF, and LR gradient estimation techniques*, Management Sci., 36 (1990), 1364–1383.

[26] P. L'Ecuyer and P. W. Glynn, *Stochastic optimization by simulation: Convergence proofs for the GI/G/*1 *queue in steady-state*, Management Sci., 40 (1994), pp. 1562–1578.

[27] P. L'Ecuyer, N. Giroux, and P. W. Glynn, *Stochastic optimization by simulation: Numerical experiments with the M/M/*1 *queue in steady-state*, Management Sci., 40 (1994), pp. 1245–1261.

[28] P. L'Ecuyer and G. Yin, *Budget-dependent convergence rate of stochastic approximation*, SIAM J. Optim., 8 (1998), pp. 217–247.

[29] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, New York, 1993.

[30] E. Nummelin, *General Irreducible Markov Chains and Non-negative Operators*, Cambridge Univ. Press, New York, 1984

[31] B. T. Polyak, *New method of stochastic approximation type procedure*, Automat. Remote Control, 51 (1990), pp. 937–946.

[32] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855.

[33] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Statist., 22 (1951), pp. 400–407.

[34] K. Sigman, *Notes on the stability of closed queueing networks*, J. Appl. Probab., 26 (1989), pp. 678–682; *Corrections*, J. Appl. Probab., 27 (1990), p. 735.

[35] K. Sigman, *The stability of open queueing networks*, Stochastic Process. Appl., 35 (1990), pp. 11–25.

[36] K. Sigman and R. W. Wolff, *A review of regenerative processes*, SIAM Rev., 35 (1993), pp. 269–288.

[37] W. F. Stout, *Almost Sure Convergence*, Academic Press, New York, 1974.

[38] R. Suri, *Perturbation analysis: The state of the art and research issues explained via the GI/G/*1 *queue*, Proc. IEEE, 77 (1989), pp. 114–137.

[39] R. Suri and Y. T. Leung, *Single run optimization of discrete event simulations—An empirical study using the M/M/*1 *queue*, IIE Trans., 21 (1989), pp. 35–49.

[40] Q. Y. Tang and H. F. Chen, *Convergence of perturbation analysis based optimization algorithm with fixed number of customers period*, Discrete Event Dynam. Systems: Theory Appl., 4 (1994), pp. 359–375.

[41] Q. Y. Tang, H. F. Chen, and Z. J. Han, *Convergence rates of perturbation-analysis-Robbins-Monro-single-run algorithms for single server queues*, IEEE Trans. Automat. Control, 42 (1997), pp. 1442–1447.

[42] R. Wolff, *Stochastic Modeling and the Theory of Queues*, Prentice–Hall, Englewood Cliffs, NJ, 1989.

[43] G. Yin, *On extensions of Polyak's averaging approach to stochastic approximation*, Stochastics Stochastics Rep., 36 (1991), pp. 245–264.

SIAM J. Control Optim.
Vol. 37, No. 6, pp. 1848–1864

A PASSIVE SYSTEM APPROACH TO FEEDBACK STABILIZATION OF NONLINEAR CONTROL STOCHASTIC SYSTEMS[*]

PATRICK FLORCHINGER[†]

Abstract. The purpose of this paper is to provide sufficient conditions for asymptotic stabilization in probability of nonlinear control stochastic differential systems by means of smooth state feedback laws. In particular, we prove that, as in the case of stochastic differential systems affine in the control, one can compute stabilizing feedback laws provided the unforced stochastic differential system is Lyapunov stable in probability and some rank conditions are fulfilled. Some well-known stabilization results for stochastic differential systems affine in the control proved in the literature are extended to nonaffine nonlinear control stochastic differential systems.

Key words. stability in probability, control stochastic differential equation, smooth state feedback law, passive stochastic system

AMS subject classifications. 60H10, 93C10, 93D05, 93D15, 93E15

PII. S0363012997317487

1. Introduction.

The aim of this paper is to state sufficient conditions for asymptotic stabilization in probability of some nonlinear control stochastic differential systems. In particular, we prove that stabilizing feedback laws can be computed provided the unforced stochastic differential system is stable in probability and some rank conditions are satisfied.

The stabilizability of nonlinear stochastic differential systems affine in the control has been studied in the past years by means of the stochastic Lyapunov machinery developed by Khasminskii [12]. Among several contributions to the subject, we wish to outline those which are more closely related to the work of the present paper.

The pioneering work of Jurdjevic and Quinn [10] on the stabilization of nonlinear deterministic systems affine in the control, the unforced dynamics of which are linear, has been extended to nonlinear stochastic differential systems in [4]. Following the method developed by Kalouptsidis and Tsinias [11] for deterministic systems, Boulanger has given in [1], by using output feedback laws, an extension of the stochastic version of Jurdjevic and Quinn's theorem to stochastic systems, the unforced dynamics of which are stable in probability. Furthermore, the stabilizing feedback laws computed in the latter work are smooth provided restrictive assumptions on the Lyapunov function associated to the unforced dynamics of the system are satisfied. In [5] and [7] the conditions proposed in [4] are weakened and more general sufficient conditions for asymptotic stability in probability of nonlinear stochastic differential systems are given, provided the unforced dynamics are Lyapunov stable in probability and some rank condition is satisfied.

The problem of computing stabilizing state feedback laws for nonlinear stochastic differential systems is much harder (see, for example, [5] or [3] and the references therein). The purpose of this paper is to address this question and provide an easy way to reach this goal in some specific cases.

The notion of "passivity" has been used in order to analyze the stability of some deterministic interconnected nonlinear systems (see [17], for example). In [8] and

[*] Received by the editors February 25, 1997; accepted for publication (in revised form) October 21, 1998; published electronically October 8, 1999.

http://www.siam.org/journals/sicon/37-6/31748.html

[†] 23 Allée des Oeillets, F 57160 Moulins les Metz, France (patrick.florchinger@wanadoo.fr).

[9], Hill and Moylan have combined the methodology of passive systems theory with the Lyapunov stability theory in order to prove stabilization results for nonlinear deterministic systems. Byrnes, Isidori, and Willems [2] have developed a framework for studying the stabilizability of minimum-phase nonlinear deterministic systems, and they have shown by means of techniques of passive systems theory that the stabilization results proved in [10], [11], and [15] for deterministic systems can be deduced from a basic stability property of passive systems affine in the control.

Noticing that passive systems theory has been developed by Willems [18] for nonaffine nonlinear deterministic systems, Lin [16] extends the results proved in [2] to a wider class of systems which are not necessarily affine in the control.

The purpose of this paper is to introduce the notion of "passivity" for nonlinear control stochastic differential systems by using a method close to that developed by Lin [16] in the deterministic case. With this aim, we show that the problem of asymptotic stabilization in probability of nonlinear stochastic differential systems can be solved by means of the stochastic passive systems theory developed in the following. Furthermore, we will see that certain stabilization results proposed in the literature by analytic methods can be recovered from the basic stabilizability property of passive stochastic systems. This shows that a number of stabilization schemes reduce to that of a passive stochastic system subject to pure gain feedback. Hence, the role played by the concept of passivity in order to compute stabilizing state feedback laws leads to the following question: when can a nonlinear stochastic differential system be rendered passive via state feedback law? This question will be solved in this paper in the particular case of interconnected stochastic differential systems.

This paper is divided into six sections and is organized as follows. In section 1, we recall some basic facts concerning the asymptotic stability in probability of the equilibrium solution of a stochastic differential equation proved by Khasminskii [12] and we state the stochastic version of La Salle's theorem proved by Kushner [13]. In section 2, we introduce the class of stochastic differential systems we are dealing with in this paper. In section 3, we introduce the notion of passivity for input/output nonlinear stochastic differential systems and we prove a first stabilization result for such systems and a nonlinear version of the Kalman–Yacubovitch–Popov lemma for stochastic differential systems affine in the control. In section 4, we use the methodology for passive stochastic systems developed in section 3 in order to prove two feedback stabilization results for nonlinear stochastic differential systems. In section 5, we study the particular case of interconnected stochastic differential systems which are feedback equivalent to passive stochastic differential systems. In section 6, we illustrate some results of the paper by design examples.

2. Stochastic stability.

The purpose of this section is to recall the main results concerning the asymptotic stability in probability of the equilibrium solution of a stochastic differential equation that we need in the following as well as the stochastic version of La Salle's theorem.

For a complete presentation of the stochastic stability theory and the stochastic Lyapunov machinery we refer the reader to the book of Khasminskii [12] for examples.

Let $(\Omega, \mathcal{F}, P)$ be a complete probability space and denote by $w = \{w_t \ ; \ t \geq 0\}$ a standard $\mathbb{R}^p$-valued Wiener process defined on this space.

Consider the stochastic process solution $x_t \in \mathbb{R}^n$ of the stochastic differential equation written in the sense of Itô,

$$x_t = x_0 + \int_0^t f(x_s)\, ds + \int_0^t g(x_s)\, dw_s \tag{2.1}$$

where

(1) x_0 is given in $\mathbb{R}^n$,

(2) f and g are Lipschitz functions mapping $\mathbb{R}^n$ into $\mathbb{R}^n$ and $\mathbb{R}^{n\times p}$, respectively, vanishing in the origin, and such that there exists a nonnegative constant K such that for any $x \in \mathbb{R}^n$,

$$|f(x)| + |g(x)| \leq K(1 + |x|). \tag{2.2}$$

If for any $s \geq 0$ and $x \in \mathbb{R}^n$, $x_t^{s,x}$, $t \geq s$, denotes the solution at time t of the stochastic differential equation (2.1) starting from the state x at time s, one can introduce the notion of stability in probability for the equilibrium solution of the stochastic differential equation (2.1) as follows.

Definition 2.1. *The equilibrium solution $x_t \equiv 0$ of the stochastic differential equation* (2.1) *is stable in probability if and only if for any $s \geq 0$ and $\epsilon > 0$,*

$$\lim_{x\to 0} P\left(\sup_{s\leq t} |x_t^{s,x}| > \epsilon\right) = 0.$$

Furthermore, the equilibrium solution of the stochastic differential equation (2.1) is (locally) asymptotically stable in probability if the following property is satisfied.

Definition 2.2. (1) *The equilibrium solution of the stochastic differential equation* (2.1) *is locally asymptotically stable in probability if and only if it is stable in probability and for any $s \geq 0$,*

$$\lim_{x\to 0} P\left(\lim_{t\to +\infty} |x_t^{s,x}| = 0\right) = 1.$$

(2) *The equilibrium solution of the stochastic differential equation* (2.1) *is asymptotically stable in probability if and only if it is stable in probability and for any $s \geq 0$,*

$$P\left(\lim_{t\to +\infty} |x_t^{s,x}| = 0\right) = 1$$

for any $x \in \mathbb{R}^n$.

Denote by L the infinitesimal generator of the stochastic process solution x_t of the stochastic differential equation (2.1). Then, by means of martingale theory arguments, one can prove the following version of the Lyapunov theorem.

Theorem 2.3. *Assume that there exists a Lyapunov function V defined on a bounded open neighborhood D of the origin in $\mathbb{R}^n$ (i.e., a function V in $C^2(D;\mathbb{R})$ which is positive definite) such that*

$$LV(x) \leq 0$$
$$(\textit{respectively},\ LV(x) < 0)$$

for any $x \in D\backslash\{0\}$. Then, the equilibrium solution $x_t \equiv 0$ of the stochastic differential equation (2.1) *is stable (respectively, locally asymptotically stable) in probability.*

Moreover, if $D = \mathbb{R}^n$ and the Lyapunov function V is proper, then the equilibrium solution $x_t \equiv 0$ of the stochastic differential equation (2.1) *is asymptotically stable in probability provided*

$$LV(x) < 0$$

for any $x \in \mathbb{R}^n \setminus \{0\}$.

For a detailed proof of Theorem 2.3 we refer the reader to Khasminskii [12].

The assumptions on the coefficients of the stochastic differential equation (2.1) can be relaxed in the following way provided the equilibrium solution of the system is asymptotically stable in probability (see [12, Theorem 4.1, p. 84]).

THEOREM 2.4. *Suppose that the functions f and g are locally Lipschitz and satisfy inequality* (2.2) *on any ball $B(0,R) = \{x \in \mathbb{R}^n \ / \ ||x|| \le R\}$. Suppose that there exists a proper Lyapunov function V defined on $\mathbb{R}^n$ and a positive constant c such that*

$$LV(x) \le cV(x)$$

for every x in $\mathbb{R}^n$. Then, the stochastic differential equation (2.1) *has a unique solution $(x_t)_{t\ge 0}$ with continuous trajectories.*

To conclude this section, recall the stochastic version of La Salle's theorem proved by Kushner [13] which gives the ω-limit set for a stable in probability equilibrium solution of the stochastic differential equation (2.1).

THEOREM 2.5. *Assume there exists a Lyapunov function V such that*

$$LV(x) \le 0$$

for any $x \in \mathbb{R}^n$. Then, the stochastic process solution x_t of the stochastic differential equation (2.1) *tends in probability to the largest invariant set whose support is contained in the locus $LV(x_t) = 0$ for any $t \ge 0$.*

REMARK 2.6. *If the equilibrium solution $x_t \equiv 0$ of the stochastic differential equation* (2.1) *is stable in probability, then the existence of a Lyapunov function V such that $LV(x) \le 0$ for any $x \in \mathbb{R}^n$ is given by the converse Lyapunov theorem proved by Kushner* [14].

3. Problem setting.

Denote by $(\Omega, \mathcal{F}, P)$ a complete probability space and by $w = \{w_t; t \ge 0\}$ a standard $\mathbb{R}^p$-valued Wiener process defined on this space.

Consider the stochastic differential system in $\mathbb{R}^n$ written in the sense of Itô,

$$x_t = x_0 + \int_0^t f(x_s, u)ds + \int_0^t g(x_s)dw_s \tag{3.1}$$

where

(1) x_0 is given in $\mathbb{R}^n$,

(2) u is an $\mathbb{R}^m$-valued measurable control law,

(3) f and g are Lipschitz functions defined in $C^\infty(\mathbb{R}^n \times \mathbb{R}^m; \mathbb{R}^n)$ and $C^\infty(\mathbb{R}^n; \mathbb{R}^{n\times p})$, respectively, vanishing in the origin, and such that there exists a nonnegative constant K such that for any $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$,

$$|f(x,u)| + |g(x)| \le K(1 + |x| + |u|). \tag{3.2}$$

Furthermore, assume that the unforced stochastic differential system deduced from (3.1), that is, the stochastic differential system

$$dx_t = f(x_t, 0)dt + g(x_t)dw_t, \tag{3.3}$$

is such that there exists a Lyapunov function V defined on $\mathbb{R}^n$ satisfying

$$L_0V(x) \le 0 \tag{3.4}$$

for every $x \in \mathbb{R}^n$ where L_0 is the infinitesimal generator of the stochastic differential equation (3.3).

Note that according to the stochastic Lyapunov Theorem (Theorem 2.3) the equilibrium solution $x_t \equiv 0$ of the stochastic differential system (3.3) is stable in probability.

REMARK 3.1. *If one assumes that the equilibrium solution $x_t \equiv 0$ of the stochastic differential system* (3.3) *is stable in probability, the existence of a Lyapunov function V satisfying* (3.4) *is given by the converse Lyapunov theorem proved by Kushner* [14].

The aim of this paper is to state sufficient conditions for the stochastic differential system (3.1) to be (locally) asymptotically stabilizable in probability by means of a smooth state feedback law.

4. Passive stochastic systems. Consider the input/output nonlinear stochastic differential system in $\mathbb{R}^n \times \mathbb{R}^m$ written in the sense of Itô,

$$\left\{\begin{array}{lcl} dx_t & = & f(x_t, u)dt + g(x_t)dw_t, \\ y_t & = & h(x_t, u), \end{array}\right. \tag{4.1}$$

where

(1) $w = \{w_t; t \geq 0\}$ is an $\mathbb{R}^p$-valued standard Wiener process defined on $(\Omega, \mathcal{F}, P)$,
(2) u is an $\mathbb{R}^m$-valued measurable control law,
(3) f, g, and h are smooth functions mapping $\mathbb{R}^n \times \mathbb{R}^m$, $\mathbb{R}^n$, and $\mathbb{R}^n \times \mathbb{R}^m$, respectively, in $\mathbb{R}^n$, $\mathbb{R}^{n\times p}$, and $\mathbb{R}^m$, respectively, vanishing in the origin, and satisfying a linear growth condition similar to (3.2).

First, we introduce the notion of "passivity" for the stochastic differential system (4.1) as follows.

DEFINITION 4.1. *The input/output stochastic differential system* (4.1) *is said to be passive if there exists a Lyapunov function V defined on $\mathbb{R}^n$, called the storage function, such that*

$$LV(x) \leq h(x, u)^\star u \tag{4.2}$$

for every $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$ where L is the infinitesimal generator of the stochastic process solution of the first equation in (4.1).

Then, necessary conditions for the input/output nonlinear stochastic differential system (4.1) to be passive can be obtained as follows.

PROPOSITION 4.2. *Necessary conditions for the stochastic differential system* (4.1) *to be passive with a storage function V are*

(1) $L_0V(x) \leq 0$ *for every* $x \in \mathbb{R}^n$,
(2) $\sum_{i=1}^n \frac{\partial f_i}{\partial u}(x, 0)\frac{\partial V}{\partial x_i}(x) = h(x, 0)^\star$ *for every* $x \in S_1$,
(3) $\sum_{i=1}^n \frac{\partial^2 f_i}{\partial u^2}(x, 0)\frac{\partial V}{\partial x_i}(x) \leq \frac{\partial h^\star}{\partial u}(x, 0) + \frac{\partial h}{\partial u}(x, 0)$ *for every* $x \in S_1$

where L_0 is, as in the previous section, the infinitesimal generator of the unforced stochastic differential system deduced from the first equation in (4.1) *and* $S_1 = \{x \in \mathbb{R}^n / L_0V(x) = 0\}$.

Proof of Proposition 4.2. Consider the auxiliary function F mapping $\mathbb{R}^n \times \mathbb{R}^m$ into $\mathbb{R}$ defined for every $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$ by

$$F(x, u) = LV(x) - h(x, u)^\star u.$$

Since the stochastic differential system (4.1) is passive, it is obvious, according to Definition 4.1, that $F(x, u) \leq 0$ for every $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$ and hence, one gets (2.1) by setting $u = 0$.

Conversely, noticing that for all $x \in S_1$, $F(x,0) = L_0V(x) = 0$, one can deduce that

$$F(x,u) \leq F(x,0)$$

for every $(x,u) \in S_1 \times \mathbb{R}^m$, which implies that the function F achieves its maximum at $u = 0$ on the set S_1.

Therefore, for any $x \in S_1$, it yields

$$\frac{\partial F}{\partial u}(x,0) = 0 \text{ and } \frac{\partial^2 F}{\partial u^2}(x,0) \leq 0$$

which give (2.2) and (3.1) by easy computations similar to those in the proof of Proposition 4.3 in [16]. □

REMARK 4.3. *If the stochastic differential system* (4.1) *is passive, property* (2.1) *in the previous proposition implies, according to the stochastic Lyapunov theorem* (Theorem 2.3)*, that the equilibrium solution $x_t \equiv 0$ of the unforced stochastic differential system* (3.3) *is stable in probability.*

If the stochastic differential system is affine in the control and the function h does not depend on the control, then one can prove that the system is passive provided it satisfies a nonlinear version of the Kalman–Yacubovitch–Popov (KYP) criterion.

DEFINITION 4.4. *The stochastic differential system*

$$\begin{cases} dx_t &= (f(x_t) + \bar{f}(x_t)u)dt + g(x_t)dw_t, \\ y_t &= h(x_t) \end{cases} \tag{4.3}$$

satisfies a KYP property if there exists a Lyapunov function V defined on $\mathbb{R}^n$ such that

$$L_0V(x) \leq 0$$

and

$$\nabla V(x)\bar{f}(x) = h(x)^\star$$

for every $x \in \mathbb{R}^n$.

Then, one can prove the following criterion which extends Proposition 2.12 in [2] (see also [8]) to stochastic differential systems.

PROPOSITION 4.5. *The stochastic differential system* (4.3) *is passive if and only if it satisfies a KYP property.*

Proof of Proposition 4.5. If the stochastic differential system (4.3) is passive, there exists a Lyapunov function V defined on $\mathbb{R}^n$ such that

$$LV(x) \leq h(x)^\star u$$

for every $(x,u) \in \mathbb{R}^n \times \mathbb{R}^m$.

But, since for every $(x,u) \in \mathbb{R}^n \times \mathbb{R}^m$,

$$LV(x) = L_0V(x) + \nabla V(x)\bar{f}(x)u \tag{4.4}$$

one can easily deduce from the above estimates that the KYP property is satisfied by the stochastic differential system (4.3).

Conversely, if the KYP property is satisfied, there exists a Lyapunov function V defined on $\mathbb{R}^n$ such that for every x in $\mathbb{R}^n$,

$$L_0 V(x) \leq 0$$

and

$$\nabla V(x) \bar{f}(x) = h(x)^\star .$$

Then, since for every $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$, (4.4) holds, one can deduce from the previous estimates that

$$LV(x) \leq h(x)^\star u \tag{4.5}$$

and hence, the stochastic differential system (4.3) is passive. □

Before stating a stabilization result for passive stochastic systems, we introduce the following notations.

Denote by $\bar{f}_i$, $1 \leq i \leq m$, the functions defined on $\mathbb{R}^n$ by

$$\bar{f}_i(x) = \frac{\partial f}{\partial u_i}(x, 0)$$

and, if V is a C^{2r} $(r \geq 1)$ Lyapunov function defined on $\mathbb{R}^n$, introduce the space

$$\Lambda = \operatorname{span}\left\{ \mathrm{ad}_{L_0}^k \bar{f}_i \ / \ 0 \leq k \leq n-1, 1 \leq i \leq m \right\}$$

and the sets Γ and S associated with Λ defined by

$$\Gamma = \left\{ x \in \mathbb{R}^n \ / \ L_0^k V(x) = 0 \ , k = 1, \ldots, r \right\}$$

and

$$S = \left\{ x \in \mathbb{R}^n \ / \ L_0^k Y V(x) = 0, \forall Y \in \Lambda, k = 0, \ldots, r-2 \right\},$$

where YV denotes the Lie derivative of V with respect to Y.

Then, as in the case of stochastic differential systems affine in the control, a non-affine passive stochastic differential system has the following stabilizability property by output feedback.

THEOREM 4.6. *Assume that the stochastic differential system* (4.1) *is passive with a C^{2r} $(r \geq 1)$ storage function V, and let $s : \mathbb{R}^n \to \mathbb{R}^m$ be any first/third sector function (i.e., $y^\star s(y) > 0$ $\forall y \neq 0$ and $s(0) = 0$). Then, the output feedback law*

$$u = -s(y) \tag{4.6}$$

renders the stochastic differential system (4.1) *locally asymptotically stable in probability provided $\Gamma \cap S = \{0\}$.*

Moreover, if the Lyapunov function V is proper, the equilibrium solution of the stochastic differential system (4.1) *is asymptotically stable in probability.*

Proof of Theorem 4.6. The proof of this theorem, which is an extension of Theorem 3.2 in [2] to stochastic differential systems, follows from passive inequality (4.2) with the feedback law (4.6) and the stochastic version of La Salle's theorem (Theorem 2.5).

Indeed, with the feedback law u given by (4.6) one can deduce from the passive inequality (4.2) that

$$LV(x) \leq -h(x, u)^\star s\left(h(x, u)\right) \tag{4.7}$$

for every $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$ and, since s is a first/third sector function, yields

$$LV(x) \leq 0$$

which implies according to the stochastic Lyapunov theorem (Theorem 2.3) that the equilibrium solution of the stochastic differential system (4.1) is stable in probability.

Furthermore, according to the stochastic version of La Salle's theorem (Theorem 2.5) the equilibrium solution of the stochastic differential system (4.1) tends in probability to the largest invariant set whose support is contained in the locus $LV(x_t) = 0$ for all $t \geq 0$.

But, if $LV(x_t) = 0$ for all $t \geq 0$, one can deduce from (4.7) and the fact that s is a first/third sector function that $y_t = h(x_t, u) = 0$ for all $t \geq 0$ which implies that $L_0V(x_t) = 0$ for all $t \geq 0$.

Therefore, by successive applications of Itô's formula to the stochastic process $LV(x_t)$, one gets that $L_0^kV(x_t) = 0$ $\forall t \geq 0$ and $k \in \{1, \ldots, r\}$ and $L_0^kYV(x_t) = 0$ $\forall t \geq 0$, $Y \in D$, and $k \in \{0, \ldots, r-2\}$ which implies, since $\Gamma \cap S = \{0\}$, that the stochastic process x_t tends in probability to 0.

This proves, by the stochastic version of La Salle's theorem, that the equilibrium solution of the stochastic differential equation (4.1) is locally asymptotically stable in probability.

Moreover, if the Lyapunov function V is proper, the equilibrium solution of the stochastic differential system (4.1) is asymptotically stable in probability. □

The following result, which is an easy consequence of the previous theorem, will enable us to recover well-known results of the stochastic stabilization theory.

COROLLARY 4.7. *Assume that the stochastic differential system* (4.1) *is passive with a proper* C^{2r} $(r \geq 1)$ *storage function* V. *If* $\Gamma \cap S = \{0\}$, *then, for each* $k > 0$, *the feedback law*

$$u = -ky$$

renders the stochastic differential system (4.1) *asymptotically stable in probability.*

REMARK 4.8. *The result of the previous corollary can be interpreted as a generalization of the results proved in* [1], [4], *and* [5] *on the stabilization of stochastic differential systems affine in the control*

$$dx_t = (f(x_t) + \bar{f}(x_t)u)dt + g(x_t)dw_t. \tag{4.8}$$

For example, if the stochastic differential system (4.8) *satisfies the hypotheses stated in* [4], *then Theorem* 3.2 *in* [4] *is a particular case of Corollary* 4.7.

Indeed, by taking the dummy output

$$y_t = \nabla V(x_t)\bar{f}(x_t),$$

one can prove easily that the hypotheses stated in [4] *imply those stated in the previous corollary and hence, the feedback law*

$$u(x) = -\nabla V(x)\bar{f}(x)$$

renders the stochastic differential system (4.8) *asymptotically stable in probability.*

5. The main results. We are now ready to state the main results of the paper on the stabilization of the class of systems introduced previously. The first result is given by the following statement.

Theorem 5.1. *If the stochastic differential system* (3.1) *is such that* $\Gamma \cap S = \{0\}$*, then it is locally asymptotically stabilizable in probability by a local smooth state feedback law* $u = \alpha(x)$*,* $\alpha(0) = 0$*, which can be solved uniquely from the equation*

$$u + \left(\frac{\partial V}{\partial x}(x) k(x,u) \right)^{\star} = 0, \tag{5.1}$$

where k *is the smooth function mapping* $\mathbb{R}^n \times \mathbb{R}^m$ *into* $\mathbb{R}^{n \times m}$ *defined by*

$$f(x,u) - f(x,0) = k(x,u)u.$$

Moreover, if V *is proper and equation* (5.1) *has a solution* $u = \alpha(x)$ *which is well-defined on* $\mathbb{R}^n$*, then the stochastic differential system* (3.1) *is asymptotically stabilizable in probability.*

Proof of Theorem 5.1. Choosing the output y defined by

$$y = h(x,u) = \left(\frac{\partial V}{\partial x}(x) k(x,u) \right)^{\star}, \tag{5.2}$$

one can prove easily that the input/output stochastic differential system (3.1)–(5.2) is passive.

Indeed, for every $(x,u) \in \mathbb{R}^n \times \mathbb{R}^m$,

$$\begin{aligned} LV(x) &= L_0 V(x) + \left(\frac{\partial V}{\partial x}(x) k(x,u) \right) u \\ &\leq h(x,u)^{\star} u. \end{aligned}$$

Therefore, since $\Gamma \cap S = \{0\}$ one can deduce from Theorem 4.6 that the output feedback control law

$$u = -y = -\left(\frac{\partial V}{\partial x}(x) k(x,u) \right)^{\star}$$

renders the stochastic differential system (3.1)–(5.2) locally asymptotically stable in probability.

Hence, the state feedback law $u = \alpha(x)$ renders the equilibrium solution $x_t \equiv 0$ of the stochastic differential equation (3.1) locally asymptotically stable in probability provided that there exists a C^{∞} function α, locally defined on a neighborhood of the origin, such that (5.1) is satisfied.

This is indeed the case since the function H defined on $\mathbb{R}^n \times \mathbb{R}^m$ by

$$H(x,u) = u + \left(\frac{\partial V}{\partial x}(x) k(x,u) \right)^{\star}$$

satisfies the assumptions of the implicit function theorem.

Thus, the equilibrium solution of the stochastic differential equation (3.1) is locally asymptotically stable in probability and, if V is proper and equation (5.1) has a solution defined on all $\mathbb{R}^n$, then the result also holds globally.

This completes the proof of Theorem 5.1. □

Remark 5.2. *In the case of nonlinear stochastic differential systems affine in the control,* (5.1) *reduces to*

$$u = -\left(\sum_{i=1}^{n} \frac{\partial f_i}{\partial u}(x,0) \frac{\partial V}{\partial x_i}(x) \right)^{\star}$$

which implies that it always has a global solution on $\mathbb{R}^n$*. Hence, Theorem* 5.1 *covers the stabilization result proved in* [4]*.*

The second result concerns stochastic differential systems in $\mathbb{R}^n$ written in the sense of Itô in the form

$$x_t = x_0 + \int_0^t \left(f(x_s) + \bar{f}(x_s)u + \sum_{i=1}^m u_i R_i(x_s)u \right) ds + \int_0^t g(x_s)dw_s \tag{5.3}$$

where

(1) x_0 is given in $\mathbb{R}^n$,

(2) u is an $\mathbb{R}^m$-valued measurable control law,

(3) f and g are smooth Lipschitz functions defined in $C^\infty(\mathbb{R}^n;\mathbb{R}^n)$ and $C^\infty(\mathbb{R}^n;\mathbb{R}^{n\times p})$, respectively, vanishing in the origin, and satisfying a linear growth condition similar to (3.2),

(4) $\bar{f}$ and R_i, $1 \leq i \leq m$, are smooth Lipschitz functions defined in $C^\infty(\mathbb{R}^n;\mathbb{R}^{n\times m})$, vanishing in the origin, and satisfying a linear growth condition as in (3.2).

Furthermore, assume that the unforced stochastic differential system deduced from (5.3), that is, the stochastic differential system

$$dx_t = f(x_t)dt + g(x_t)dw_t, \tag{5.4}$$

is such that there exists a Lyapunov function V defined on $\mathbb{R}^n$ satisfying

$$L_0V(x) \leq 0 \tag{5.5}$$

for every $x \in \mathbb{R}^n$, where L_0 is the infinitesimal generator of the stochastic differential equation (5.4).

Note that according to the stochastic Lyapunov theorem (Theorem 2.3) the equilibrium solution $x_t \equiv 0$ of the stochastic differential system (5.4) is stable in probability.

The drift term of the stochastic differential system (5.3) can be regarded as "second degree approximation" of the drift term of the stochastic differential system (3.1).

We are now ready to state the second result of the section.

THEOREM 5.3. *Assume that the stochastic differential system* (5.3) *is such that* $\Gamma \cap S = \{0\}$ *and that for every* $x \in \mathbb{R}^n$*, the matrix* $\Delta(x)$ *defined by*

$$\Delta(x) = Id_{m\times m} + \begin{pmatrix} \nabla V(x)R_1(x) \\ .. \\ .. \\ .. \\ \nabla V(x)R_m(x) \end{pmatrix}^\star$$

is invertible. Then, the stochastic differential system (5.3) *is locally asymptotically stabilizable in probability by the smooth state feedback law u defined by*

$$u(x) = -\left(\Delta(x)\right)^{-1}\left(\nabla V(x)\bar{f}(x)\right)^\star . \tag{5.6}$$

Moreover, if in addition the Lyapunov function V is proper, then the state feedback law u given by (5.6) *renders the stochastic differential system* (5.3) *asymptotically stable in probability.*

Proof of Theorem 5.3. Choosing the output y defined by

$$y = \left(\nabla V(x)\bar{f}(x)\right)^{\star} + \begin{pmatrix} \nabla V(x)R_1(x) \\ .. \\ .. \\ .. \\ \nabla V(x)R_m(x) \end{pmatrix}^{\star} u, \tag{5.7}$$

one can prove easily that the input/output stochastic differential system (5.3)–(5.7) is passive.

Indeed, a straightforward computation gives

$$\begin{aligned} LV(x) &= L_0V(x) + \nabla V(x)\bar{f}(x)u + \sum_{i=1}^{m} u_i \left(\nabla V(x)R_i(x)\right) u \\ &\leq \left(\nabla V(x)\bar{f}(x) + u^{\star} \begin{pmatrix} \nabla V(x)R_1(x) \\ .. \\ .. \\ .. \\ \nabla V(x)R_m(x) \end{pmatrix} \right) u \\ &\leq y^{\star}u. \end{aligned}$$

Therefore, since $\Gamma \cap S = \{0\}$ one can deduce from Theorem 4.6 that the output feedback control law $u = -y$ renders the stochastic differential system (5.3)–(5.7) locally asymptotically stable in probability.

Hence, the state feedback law u given by (5.6) renders the equilibrium solution $x_t \equiv 0$ of the stochastic differential system (5.3) locally asymptotically stable in probability.

If the Lyapunov function V is proper, the global stabilization result of the theorem is obvious.

This completes the proof of Theorem 5.3. □

REMARK 5.4. *Since $\nabla V(0) = 0$, one has $\Delta(0) = Id_{m \times m}$ and hence, the matrix $\Delta(x)$ is always invertible on a neighborhood of the origin in $\mathbb{R}^n$. Therefore, the state feedback law u given by* (5.6) *is locally well-defined and smooth on a neighborhood of the origin.*

In the particular case of a single input, system (5.3) can be written as

$$x_t = x_0 + \int_0^t \left(f(x_s) + \bar{f}(x_s)u + r(x_s)u^2\right) ds + \int_0^t g(x_s)dw_s \tag{5.8}$$

and Theorem 5.3 leads to the following result.

COROLLARY 5.5. *Assume that the stochastic differential system* (5.8) *is such that $\Gamma \cap S = \{0\}$. Then, the smooth state feedback law u defined by*

$$u(x) = -\frac{1}{1 + \nabla V(x)r(x)} \nabla V(x)\bar{f}(x) \tag{5.9}$$

renders the stochastic differential system (5.8) *locally asymptotically stable in probability.*

Moreover, if the Lyapunov function V is proper and $1 + \nabla V(x)r(x) \neq 0$ for every $x \in \mathbb{R}^n$, then the equilibrium solution of the stochastic differential system (5.8) *is asymptotically stabilized in probability by the state feedback law* (5.9).

Remark 5.6. *In the case of stochastic differential systems affine in the control, the matrix $\Delta(x)$ is invertible on all $\mathbb{R}^n$ since $R_i(x) \equiv 0$ for every $i \in \{1, \dots, m\}$. Therefore, the stabilization result proved in* [4] *and* [1] *can be deduced from the result proved in Theorem* 5.3.

6. Interconnected stochastic differential systems.

Consider the input/output interconnected nonlinear stochastic differential system in $\mathbb{R}^q \times \mathbb{R}^n \times \mathbb{R}^m$ written in the sense of Itô,

$$d\xi_t = (\tilde{f}(\xi_t) + \tilde{f}_0(\xi_t, y_t)y_t)dt + \tilde{g}(\xi_t)d\tilde{w}_t, \tag{6.1}$$

$$dx_t = (f(x_t) + \bar{f}(x_t)u)dt + g(x_t)dw_t, \tag{6.2}$$

$$y_t = h(x_t), \tag{6.3}$$

where

(1) $w = \{w_t, t \geq 0\}$ and $\tilde{w} = \{\tilde{w}_t, t \geq 0\}$ are independent standard Wiener processes defined on $(\Omega, \mathcal{F}, P)$ with values in $\mathbb{R}^p$ and $\mathbb{R}^r$, respectively,

(2) u is an $\mathbb{R}^m$-valued measurable control law,

(3) $\tilde{f}$ and $\tilde{g}$ are smooth Lipschitz functions mapping $\mathbb{R}^q$ into $\mathbb{R}^q$ and $\mathbb{R}^{q\times r}$, respectively, vanishing in the origin, and satisfying a linear growth condition similar to (3.2),

(4) $\tilde{f}_0$ is a smooth Lipschitz function mapping $\mathbb{R}^q \times \mathbb{R}^m$ into $\mathbb{R}^{q\times m}$,

(5) f, $\bar{f}$ and g are smooth Lipschitz functions mapping $\mathbb{R}^n$ into $\mathbb{R}^n$, $\mathbb{R}^{n\times m}$, and $\mathbb{R}^{n\times p}$, respectively, vanishing in the origin, and satisfying a linear growth condition similar to (3.2),

(6) h is a smooth function mapping $\mathbb{R}^n$ into $\mathbb{R}^m$.

First, we discuss conditions under which the stochastic differential system (6.1)–(6.3) is feedback equivalent to a passive stochastic differential system.

Proposition 6.1. *Suppose that $\xi_t \equiv 0$ is an asymptotically stable in probability equilibrium solution of the stochastic differential system*

$$d\xi_t = \tilde{f}(\xi_t)dt + \tilde{g}(\xi_t)d\tilde{w}_t, \tag{6.4}$$

for which we know a C^{2r} Lyapunov function U such that

$$\tilde{L}U(\xi) < 0$$

for every ξ in $\mathbb{R}^q$ (here, $\tilde{L}$ denotes the infinitesimal generator of the stochastic process solution of (6.4)*). Suppose that the stochastic differential system* (6.2)*–*(6.3) *is passive with a C^{2r} storage function. Then, the stochastic differential system* (6.1)*–*(6.3) *is feedback equivalent to a passive system with a C^{2r} storage function.*

Proof of Proposition 6.1. Setting $\zeta_t = (\xi_t, x_t)$, the stochastic differential system (6.1)–(6.3) has the form

$$d\zeta_t = (F(\zeta_t) + \bar{F}(\zeta_t)u)dt + G(\zeta_t)d\begin{pmatrix} \tilde{w}_t \\ \\ w_t \end{pmatrix}, \tag{6.5}$$

$$y_t = H(\zeta_t), \tag{6.6}$$

where

$$F(\zeta) = \begin{pmatrix} \tilde{f}(\xi) + \tilde{f}_0(\xi, h(x))h(x) \\ \\ f(x) \end{pmatrix},$$

$$\bar{F}(\zeta) = \begin{pmatrix} 0 \\ \\ \bar{f}(x) \end{pmatrix},$$

$$G(\zeta) = \begin{pmatrix} \tilde{g}(\xi) \\ \\ g(x) \end{pmatrix},$$

and

$$H(\zeta) = h(x).$$

In the following, we prove by means of Proposition 4.5 that the stochastic differential system (6.5)–(6.6) is rendered passive by the feedback law

$$u = -\left(\nabla U(\xi)\tilde{f}_0(\xi, h(x))\right)^{\star} + v. \tag{6.7}$$

Indeed, denoting by W the C^{2r} composite Lyapunov function defined on $\mathbb{R}^q \times \mathbb{R}^n$ by

$$W(\zeta) = U(\xi) + V(x)$$

yields

$$\begin{aligned} \nabla W(\zeta)\bar{F}(\zeta) &= \nabla V(x)\bar{f}(x) \\ &= h(x)^{\star} \\ &= H(\zeta)^{\star}. \end{aligned}$$

Furthermore, denoting by $\mathcal{L}_0$ the infinitesimal generator of the unforced stochastic differential system deduced from (6.5) when u is given by (6.7), one has

$$\begin{aligned} \mathcal{L}_0 W(\zeta) &= \tilde{L}U(\xi) + \nabla U(\xi)\tilde{f}_0(\xi, h(x))h(x) + LV(x) + \nabla V(x)\bar{f}(x)u \\ &= \tilde{L}U(\xi) + LV(x) \\ &\leq 0. \end{aligned}$$

Then, the stochastic differential system deduced from (6.5)–(6.6) when u is given by (6.7) satisfies a KYP property and, according to Proposition 4.5, is passive. □

Now, we apply some of the above results to the problem of asymptotic stabilization in probability of interconnected nonlinear stochastic differential systems.

THEOREM 6.2. *Suppose that the equilibrium solution $\xi_t \equiv 0$ of the stochastic differential system*

$$d\xi_t = \tilde{f}(\xi_t)dt + \tilde{g}(\xi_t)d\tilde{w}_t$$

is asymptotically stable in probability and that we know a C^{2r} Lyapunov function U such that

$$\tilde{L}U(\xi) < 0$$

for every ξ in $\mathbb{R}^q$. Suppose that the stochastic differential system (6.2)–(6.3) *is passive with a C^{2r} storage function and $\Gamma \cap S = \{0\}$. Then, the stochastic differential system* (6.1)–(6.2) *is asymptotically stabilizable in probability by means of a smooth state feedback law.*

Proof of Theorem 6.2. By Proposition 6.1 the stochastic differential system (6.1)–(6.3) is feedback equivalent to a passive stochastic differential system.

Then, setting $v = -s(y)$ where s is a first/third sector function in this stochastic differential system, one gets for every $(\xi, x) \in \mathbb{R}^q \times \mathbb{R}^n$,

$$\mathcal{L}W(\xi, x) \leq -h(x)s(h(x)),$$

where $\mathcal{L}$ is the infinitesimal generator of the closed-loop system.

Therefore, the equilibrium solution $(\xi_t, x_t) \equiv (0,0)$ of the closed-loop system is stable in probability, and according to the stochastic version of La Salle's theorem (Theorem 2.5) it tends in probability to the largest invariant set whose support is contained in the locus $\mathcal{L}W(\xi_t, x_t) = 0 \; \forall t \geq 0$.

Arguing as in the proof of Theorem 4.6, the latter condition implies that $\tilde{L}U(\xi_t) = 0$ and $LV(x_t) = 0 \; \forall t \geq 0$.

Then, since $\xi_t \equiv 0$ is asymptotically stable in probability, the stochastic process ξ_t tends in probability to 0 and, since $\Gamma \cap S = \{0\}$, successive applications of Itô's formula to the process $LV(x_t)$ implies that the stochastic process x_t tends in probability to 0.

This proves that the equilibrium solution $(\xi_t, x_t) \equiv (0,0)$ of the closed-loop system deduced from the equivalent passive system to (6.1)–(6.3) when v is an output feedback law given by a first/third sector function is asymptotically stable in probability. □

Moreover, if (6.2) is linear, one can deduce as a corollary of the previous theorem, the following result originally proven in [6].

Consider the stochastic process $(\xi_t, x_t) \in \mathbb{R}^q \times \mathbb{R}^n$ solution of the interconnected stochastic differential system written in the sense of Itô,

$$d\xi_t = (\tilde{f}(\xi_t) + \tilde{f}_0(\xi_t, x_t)Dx_t)dt + \tilde{g}(\xi_t)d\tilde{w}_t, \tag{6.8}$$

$$dx_t = (Ax_t + Bu)dt + \sum_{k=1}^{p} C_k x_t dw_t^k, \tag{6.9}$$

where

1. $w = \{w_t, t \geq 0\}$ and $\tilde{w} = \{\tilde{w}_t, t \geq 0\}$ are independent standard Wiener processes defined on $(\Omega, \mathcal{F}, P)$ with values in $\mathbb{R}^p$ and $\mathbb{R}^r$, respectively,
2. u is an $\mathbb{R}^m$-valued measurable control law,
3. $\tilde{f}$ and $\tilde{g}$ are smooth Lipschitz functions mapping $\mathbb{R}^q$ into $\mathbb{R}^q$ and $\mathbb{R}^{q \times r}$, respectively, vanishing in the origin, and satisfying a linear growth condition similar to (3.2),
4. $\tilde{f}_0$ is a smooth Lipschitz function mapping $\mathbb{R}^q \times \mathbb{R}^n$ into $\mathbb{R}^{q \times m}$,
5. D is a matrix in $\mathcal{M}_{m \times n}(\mathbb{R})$,
6. A, B, and C_k, $1 \leq k \leq p$, are matrices in $\mathcal{M}_{n \times n}(\mathbb{R})$, $\mathcal{M}_{n \times m}(\mathbb{R})$, and $\mathcal{M}_{n \times n}(\mathbb{R})$, respectively.

Then, one can prove the following stabilization result for the stochastic differential system (6.8)–(6.9).

COROLLARY 6.3. *Assume that $\xi_t \equiv 0$ is an asymptotically stable in probability equilibrium solution of the stochastic differential system*

$$d\xi_t = \tilde{f}(\xi_t)dt + \tilde{g}(\xi_t)d\tilde{w}_t,$$

for which we know a C^{2r} Lyapunov function U such that

$$\tilde{L}U(\xi) < 0$$

for every ξ in $\mathbb{R}^q$. Suppose that there exists a matrix K in $\mathcal{M}_{m\times n}(\mathbb{R})$ and a symmetric and positive definite matrix P in $\mathcal{M}_{n\times n}(\mathbb{R})$ such that

$$(A+BK)^\star P + P(A+BK) + \sum_{k=1}^{p} C_k^\star P C_k < 0$$

and

$$D = 2B^\star P.$$

Then, the stochastic differential system (6.8)–(6.9) *is asymptotically stabilizable in probability by means of a smooth state feedback law.*

Proof. Choosing the output y defined by

$$y_t = Dx_t, \tag{6.10}$$

one can prove easily that the input/output stochastic differential system (6.9)–(6.10) is passive with the storage function V defined on $\mathbb{R}^n$ by $V(x) = \langle Px, x\rangle$ and satisfies $\Gamma \cap S = \{0\}$.

The conclusion follows immediately by application of Theorem 6.2. □

7. Examples. *Example* 1. Consider the stochastic differential system defined on $\mathbb{R}^2$ by

$$\begin{cases} dx_1 &= \left(-\frac{1}{2}x_1 + x_2 + x_2^2u^3\right)dt + x_1 dw_t^1, \\ dx_2 &= \left(-x_1 - \frac{1}{2}x_2 + x_1x_2u\right)dt + x_2 dw_t^2, \end{cases} \tag{7.1}$$

where x_0 is given in $\mathbb{R}^2$.

Then, denoting by V the Lyapunov function defined on $\mathbb{R}^2$ by

$$V(x) = \frac{1}{2}\left(x_1^2 + x_2^2\right),$$

one gets by routine calculations that $L_0V(x) = 0$ for every $x \in \mathbb{R}^2$.

Furthermore,

$$\nabla V(x)\bar{f}(x) = x_1x_2^2,$$

$$\mathrm{ad}_{L_0}\bar{f}V(x) = \frac{3}{2}x_1x_2^2 - 2x_1^2x_2 + x_2^3,$$

and

$$L_0\left(\mathrm{ad}_{L_0}\bar{f}V\right)(x) = -\frac{25}{4}x_1x_2^2 - 6x_1^2x_2 + 2x_1^3 + 3x_2^2.$$

Then, one has $S = \{0\}$ and since $\Gamma = \mathbb{R}^2$ one gets $\Gamma \cap S = \{0\}$.

Hence, by Theorem 5.1, there exists a smooth state feedback law which renders the stochastic differential system (7.1) asymptotically stable in probability.

Example 2. Consider the stochastic differential system with two inputs defined on $\mathbb{R}^3$ by

$$(7.2)\begin{cases} dx_1 = \left(-\frac{1}{2}x_1 + -2x_2\left(x_2^2+x_3^4\right) + \ln\left(x_1^2+1\right)u_1 + \cos(x_2+x_3)u_1u_2\right)dt \\ \qquad + \left(x_1^3u_2^2\right)dt + x_1dw_t, \\ dx_2 = \left(x_1 - 2x_3^3 + x_2u_2 + \sin x_1u_1u_2 - 2x_3^3u_1^2\right)dt, \\ dx_3 = \left(x_2 + x_3u_2 + x_2u_1^2\right)dt, \end{cases}$$

where x_0 is given in $\mathbb{R}^3$.

Then, denoting by V the Lyapunov functional defined on $\mathbb{R}^3$ by

$$V(x) = \frac{1}{2}\left(x_1^2 + \left(x_2^2+x_3^4\right)^2\right),$$

one can prove by routine calculations that the hypothesis of Theorem 5.3 is satisfied and so the stochastic differential system (7.2) can be stabilized by the state feedback law u given by

$$u(x) = -\left(\Delta(x)\right)^{-1}\begin{pmatrix} x_1\ln\left(x_1^2+1\right) \\ \left(2x_2^2+4x_3^4\right)\left(x_2^2+x_3^4\right)\end{pmatrix}.$$

Example 3. Consider the stochastic differential system defined on $\mathbb{R}^2$ by

$$(7.3)\qquad\begin{cases} dx_1 &= \left(-x_1^3 - \frac{3}{2}x_1x_2^2 + x_1\mathrm{e}^{x_2}u^2\right)dt + x_1x_2dw_t^1, \\ dx_2 &= \left(-\frac{3}{2}x_2^3 + x_2^2u\right)dt + x_2^2dw_t^2, \end{cases}$$

where x_0 is given in $\mathbb{R}^2$.

Then, denoting by V the Lyapunov functional defined on $\mathbb{R}^2$ by

$$V(x) = \frac{1}{4}\left(x_1^4 + x_2^4\right),$$

one gets by straightforward computations that $L_0V(x) = -x_1^6$ for every $x \in \mathbb{R}^2$.

Moreover, since for every $x \in \mathbb{R}^2$,

$$\nabla V(x)\bar{f}(x) = x_2^5,$$

one has $\Gamma \cap S = \{0\}$ and, according to Corollary 5.5, the smooth state feedback law u defined by

$$u(x) = -\frac{x_2^5}{1+x_1^4\mathrm{e}^{x_2}}$$

is a global stabilizer for the stochastic differential system (7.3).

REFERENCES

[1] C. Boulanger, *Output feedback stabilization of stochastic systems*, submitted.

[2] C. I. Byrnes, A. Isidori, and J. C. Willems, *Passivity, feedback equivalence and global stabilization of minimum phase nonlinear system*, IEEE Trans. Automat. Control, 36 (1991), pp. 1228–1240.

[3] H. Deng and M. Krstic, *Stabilization of stochastic nonlinear systems driven by noise of unknown covariance*, submitted.

[4] P. Florchinger, *A stochastic version of Jurdjevic–Quinn theorem*, Stochastic Anal. App., 12 (1994), pp. 473–480.

[5] P. Florchinger, *Lyapunov-like techniques for stochastic stability*, SIAM J. Control Optim., 33 (1995), pp. 1151–1169.

[6] P. Florchinger, *Global stabilization of composite stochastic systems*, Comput. Math. Appl., 33 (1997), pp. 127–135.

[7] P. Florchinger, *Global stabilization of affine stochastic systems*, Stochastic Anal. Appl., 15 (1997), pp. 493–502.

[8] D. Hill and P. Moylan, *The stability of nonlinear dissipative systems*, IEEE Trans. Automat.Control, 21 (1976), pp. 708–711.

[9] D. Hill and P. Moylan, *Stability results for nonlinear feedback systems*, Automatica J. IFAC, 13 (1977), pp. 377–382.

[10] V. Jurdjevic and J. P. Quinn, *Controllability and stability*, J. Differential Equations, 28 (1978), pp. 381–389.

[11] N. Kalouptsidis and J. Tsinias, *Stability improvement of nonlinear systems by feedback*, IEEE Trans. Automat. Control, 29 (1984), pp. 364–367.

[12] R. Z. Khasminskii, *Stochastic Stability of Differential Equations*, Sijthoff & Noordhoff, Alphen aan den Rijn, the Netherlands, 1980.

[13] H. J. Kushner, *Stochastic Stability*, in Stability of Stochastic Dynamical Systems, R. Curtain, ed., Lecture Notes in Math. 294, Springer-Verlag, Berlin, Heidelberg, New York, 1972.

[14] H. Kushner, *Converse theorems for stochastic Liapunov functions*, SIAM J. Control, 5 (1967), pp. 228–233.

[15] K. K. Lee and A. Arapostathis, *Remarks on smooth feedback stabilization of nonlinear systems*, Systems Control Lett., 10 (1988), pp. 41–44.

[16] W. Lin, *Feedback stabilization of general nonlinear control systems: a passive system approach*, Systems Control Lett., 25 (1995), pp. 41–52.

[17] M. Vidyasagar, *New passivity-type criteria for large-scale interconnected systems*, IEEE Trans. Automat. Control, 24 (1979), pp. 575–579.

[18] J. C. Willems, *Dissipative dynamic systems*, Arch. Rational Mech. Anal., 45 (1972), pp. 321–393.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 6, pp. 1865–1873

SOLUTION OF MIMO $\mathcal{H}_2/\ell_1$ PROBLEM WITHOUT ZERO INTERPOLATION*

MURTI V. SALAPAKA[†], M. KHAMMASH[†], AND M. DAHLEH[‡]

Abstract. In this paper we present a methodology to obtain converging lower and upper bounds to a multiple objective problem where an $\mathcal{H}_2$ performance objective is minimized subject to an ℓ_1 constraint. This methodology gives a computationally efficient synthesis procedure by avoiding many of the problems that are present in methods that employ zero interpolation techniques to characterize achievable closed loop maps.

Key words. robust control, ℓ_1 optimization, discrete time

AMS subject classifications. 49N05, 49N10, 49N15, 49N35, 93C55

PII. S0363012998332573

1. Notation.

$(X, \|\| \cdot \|\|)$	The set X endowed with the norm $\| \cdot \|$.
R	The real number system.
R^n	The n-dimensional Euclidean space.
$\hat{x}(\lambda)$	The λ transform of a right sided real sequence $x = (x(k))_{k=0}^{\infty}$ defined as $\hat{x}(\lambda) := \sum_{k=0}^{\infty} x(k)\lambda^k$.
ℓ	The vector space of sequences.
$\ell^{m\times n}$	The vector space of matrix sequences of size $m \times n$.
ℓ_1	The Banach space of right sided absolutely summable real sequences with the norm given by $\|x\|_1 := \sum_{k=0}^{\infty} \|x(k)\|$.
$\ell_1^{m\times n}$	The Banach space of matrix valued right sided real sequences with the norm $\|x\|_1 := \max_{1\leq i\leq m} \sum_{j=1}^{n} \|x_{ij}\|_1$, where $x \in \ell_1^{m\times n}$ is the matrix (x_{ij}) and each x_{ij} is in ℓ_1.
c_0	The subspace of ℓ_∞ with elements x that satisfy $\lim_{k\to\infty} x(k) = 0$.
ℓ_2	The Banach space of right sided square summable sequences with the norm given by $\|x\|_2 := (\sum_{k=0}^{\infty} x(k)^2)^{1/2}$.
$\mathcal{H}_2$	The isometric isomorphic image of ℓ_2 under the λ transform $\hat{x}(\lambda)$ with the norm given by $\|\hat{x}(\lambda)\|_2 = \|x\|_2$.
P_n	The truncation operator on the space of sequences; $P_n(x(0)x(1)\ldots) = (x(0)\ x(1)\ldots x(n)0\ 0\ldots)$.
X^*	The dual space of the Banach space X. $\langle x, x^*\rangle$ denotes the value of the bounded linear functional x^* at $x \in X$.
$W(X^*, X)$	The weak star topology on X^* induced by X.
$\mathcal{D}$	The closed unit disc in the complex plane.
A'	The transpose of the matrix A.

[*] Received by the editors January 2, 1998; accepted for publication (in revised form) December 23, 1998; published electronically October 8, 1999. This research was supported by National Science Foundation grants ECS-9733802, ECS-9457485, ECS-9632820, and ECS-9457485; AFOSR grant F49620-97-1-0168; and a SPRIG grant by Iowa State University.

http://www.siam.org/journals/sicon/37-6/33257.html

[†] Electrical and Computer Engineering Department, Iowa State University, Ames, IA 50011 (murti@iastate.edu, khammash@iastate.edu).

[‡] Mechanical and Environmental Engineering Department, University of California at Santa Barbara, Santa Barbara, CA 93106 (dahleh@engineering.ucab.edu).

2. Introduction. Consider Figure 3.1, where a generalized linear time-invariant plant G is depicted with a controller K connected in a feedback arrangement. Also shown are exogenous inputs w_1 and w_2 and regulated variables z_1 and z_2. The controller determines the control effort based on the measured variable y. A number of control objectives can be cast into the setup shown in Figure 3.1. The controller's objective is to stabilize the system and enhance its performance with respect to external inputs. The nature of the exogenous inputs and the particular characteristics of the regulated variables determine the appropriate systems' "measures" that must be used to quantify performance. The $\mathcal{H}_2$, $\mathcal{H}_\infty$, and ℓ_1 measures are frequently used as objectives for control synthesis. The $\mathcal{H}_2$ measure of a system is the variance of the regulated output when the exogenous input is modeled as white noise, whereas the $\mathcal{H}_\infty$ measure is the maximum energy (the ℓ_2 norm) of the regulated variable when the exogenous signal is any signal with unit energy. Another measure of a system is the ℓ_1 measure, which is the maximum magnitude of the regulated variable when the exogenous input is allowed to be any signal with unity maximum magnitude. The design of controllers which minimize these measures have been extensively studied [1].

It is known that performance with respect to a measure (usually the ℓ_1, $\mathcal{H}_2$, or the $\mathcal{H}_\infty$ norm of the closed loop) is not a guarantee of good performance with respect to some other measure. For example, in Figure 3.1 a control design objective may be stated in terms of the ℓ_1 performance between w_1 and z_1 and the minimization of the variance of z_2 with w_2 as white noise. A standard ℓ_1 solution or a standard $\mathcal{H}_2$ solution might fail to address such multiobjective concerns. Motivated by such issues researchers have focused their attention on multiobjective problems that incorporate two or more different measures in their problem definition.

An important class of problems that falls under this category is the one that incorporates time domain objectives and the $\mathcal{H}_2$ objective. In [5] it was shown that the single-input single-output problem of minimizing the ℓ_1 norm of the closed loop subject to an $\mathcal{H}_2$ constraint can be solved via finite-dimensional convex programming. In a related result it was shown in [6] that problems that incorporate the ℓ_1 norm and the $\mathcal{H}_2$ norm of the various transfer functions in a closed loop map of a multiple-input–multiple-output (MIMO) system can be formulated and solved via finite-dimensional quadratic programming. In [2] a method based on positive cones was used to minimize an $\mathcal{H}_2$ measure of the closed loop map subject to an ℓ_1 constraint.

Most approaches that incorporate the ℓ_1 objective characterize the achievability of a closed loop map through a stabilizing controller by using zero interpolation conditions on the closed loop map [1]. Computation of the zeros and the zero directions can be done by finding the nullspaces of certain Toeplitz-like matrices. Once the optimal closed loop map is determined, the task of determining the controller still remains. The closed loop map needs to satisfy the zero interpolation conditions exactly to guarantee that the correct cancellations take place while solving for the controller. However, numerical errors are always present and there exists a need to determine which poles and zeros cancel. These difficulties exist even for the pure MIMO ℓ_1 problem when zero interpolation methods are employed. However, recently in [3] it was shown that converging upper and lower bounds can be determined to the ℓ_1 problem by solving an auxiliary problem that does not require zero interpolation and thus avoids the above mentioned problems.

In this paper we formulate an auxiliary problem to the one given in [6]. We show that converging upper and lower bounds can be computed without zero interpolation

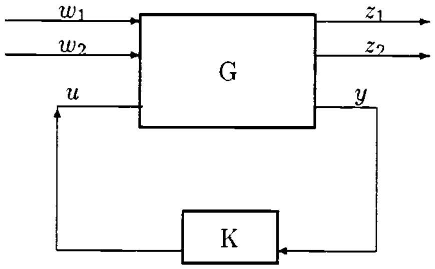


FIG. 3.1. *Closed loop system.*

for the most general MIMO case. This provides an attractive method for solving the problem.

The paper is organized as follows. In section 3 we present the preliminary material. In section 4 we formulate the multiobjective control problem and an auxiliary problem which regularizes it. In section 5 we solve the problem through lower and upper bound approximations. Section 6 contains conclusions and future directions.

3. Preliminaries. In this section we present a brief summary of mathematical and system results that will be utilized later in the paper.

3.1. System preliminaries. Consider the system in Figure 3.1, where $w := (w_1\ w_2)'$ is the exogenous disturbance, $z := (z_1\ z_2)'$ is the regulated output, u is the control input, and y is the measured output. In feedback control design the objective is to design a controller K such that with $u = Ky$ the resulting closed loop map Φ_{zw} from w to z is stable (see Figure 3.1) and satisfies certain performance criteria. In [7] a parametrization of all closed loop maps that are achievable via stabilizing controllers was derived. A good treatment of the issues involved is presented in [1]. Following the notation used in [1] we denote by n_u, n_w, n_z, and n_y the number of control inputs, exogenous inputs, regulated outputs, and measured outputs, respectively, of the plant G. We represent by Θ the set of impulse responses of closed loop maps of the plant G that are achievable through stabilizing controllers. $H \in \ell_1^{n_z \times n_3}$, $U \in \ell_1^{n_z \times n_u}$, and $V \in \ell_1^{n_y \times n_w}$ characterize the Youla parametrization of the plant [7]. The following theorem follows from the Youla parametrization.

THEOREM 3.1. $\Theta = \{\Phi \in \ell_1^{n_z \times n_w}$*: there exists a* $Q \in \ell_1^{n_u \times n_y}$ *with* $\hat{\Phi} = \hat{H} - \hat{U}\hat{Q}\hat{V}\}$*, where* $\hat{f}$ *denotes the* λ *transform* (*see* [1]) *of* f.

If Φ is in Θ we say that Φ is an *achievable* closed loop map. We assume throughout the paper that $\hat{U}$ has normal rank n_u and $\hat{V}$ has normal rank n_y. There is no loss of generality in making this assumption [1].

3.2. Mathematical preliminaries. In this subsection we summarize the mathematical results that are relevant to the paper. An exhaustive treatment of the subject matter of this subsection is given in [4]. The reader may skip this part of the paper and refer to this subsection whenever required.

DEFINITION 3.2 (convex sets). *A subset* Ω *of a vector space* X *is said to be convex if for any two elements* c_1 *and* c_2 *in* Ω *and for a real number* λ *with* $0 < \lambda < 1$ *the element* $\lambda c_1 + (1 - \lambda)c_2 \in \Omega$.

LEMMA 3.3. *Let* Ω *be a convex subset of a Banach space* X *and* $f : \Omega \rightarrow R$ *be strictly convex. If* f *achieves its minimum on* Ω*, then the minimizer is unique.*

THEOREM 3.4 (Banach–Alaoglu). *Let $(X, \|\cdot\|_x)$ be a normed vector space with X^* as its dual. The set*

$$B^* := \{x^* \in X^* : \|x^*\| \leq M\} \tag{3.1}$$

is compact in the weak-star topology for any $M \in R$.

LEMMA 3.5. *Suppose ϕ_k is a sequence in $\ell_2 \phi \in \ell_2$ and $\phi_k(t) \to \phi_0(t)$ for all t. Suppose also that $\|\phi_k\|_2 \nearrow \|\phi_0\|_2$. Then $\|\phi_k - \phi_0\|_2 \to 0$.*

Proof. Given $\epsilon > 0$ choose n such that

$$\|(I - P_n)\phi_0\|_2^2 \leq \min\left\{\frac{\epsilon}{8}, \left(\frac{\epsilon}{8(\|\phi_0\|_2 + 1)}\right)^2\right\}, \tag{3.2}$$

where P_n is the truncation operator. As $\phi_k(t) \to \phi_0(t)$ we can choose K_2 such that

$$k > K_2 \Rightarrow \|P_n(\phi_k - \phi_0)\|_2^2 \leq \frac{\epsilon}{4}. \tag{3.3}$$

We know that $\|P_n(\phi_k)\|_2 \to \|P_n(\phi_0)\|_2$ as $k \to \infty$. From the above and the fact that $\|\phi_k\|_2 \to \|\phi_0\|_2$ it follows that we can choose K_3 such that

$$k > K_3 \Rightarrow |\, \|(I - P_n)\phi_k\|_2^2 - \|(I - P_n)\phi_0\|_2^2 \,| \leq \frac{\epsilon}{4}. \tag{3.4}$$

Let $K \geq \max\{K_2, K_3\}$; then $k > K$ implies

$$\begin{aligned}
\|\phi_k - \phi_0\|_2^2 &= \|P_n(\phi_k - \phi_0)\|_2^2 + \|(I - P_n)(\phi_k - \phi_0)\|_2^2 \\
&\leq \frac{\epsilon}{4} + \|(I - P_n)(\phi_k)\|_2^2 + \|(I - P_n)(\phi_0)\|_2^2 + 2\sum_{t=n+1}^{\infty} |\phi_k(t)|\,|\phi_0(t)| \\
&\leq \frac{\epsilon}{4} + 2\|(I - P_n)(\phi_0)\|_2^2 + \frac{\epsilon}{4} + 2\sum_{t=n+1}^{\infty} |\phi_k(t)|\,|\phi_0(t)| \\
&\leq \frac{\epsilon}{4} + 2\frac{\epsilon}{8} + \frac{\epsilon}{4} + 2\|(I - P_n)\phi_k\|_2\,\|(I - P_n)\phi_0\|_2 \\
&\leq \frac{\epsilon}{4} + 2\frac{\epsilon}{8} + \frac{\epsilon}{4} + 2\|\phi_0\|_2 \frac{\epsilon}{8(\|\phi_0\|_2 + 1)} \\
&\leq \epsilon. \qquad \square
\end{aligned}$$

4. Problem statement. Let H, U, and V in the Youla parametrization be partitioned into submatrices according to the equation

$$H - U * Q * V = \begin{pmatrix} H^{11} & H^{12} \\ H^{21} & H^{22} \end{pmatrix} - \begin{pmatrix} U^1 \\ U^2 \end{pmatrix} * Q * (V^1 \;\; V^2),$$

where $Q \in \ell_2^{n_u \times n_y}$. The problem statement is as follows: *Given a plant G and a positive real number γ solve the problem*

$$\inf_{Q \in \ell_1^{n_u \times n_y}} \|H^{22} - U^2 * Q * V^2\|_2^2$$

subject to

$$\|H^{11} - U^1 * Q * V^1\|_1 \leq \gamma.$$

We denote by μ the optimal value obtained from the above problem.

Now we define an auxiliary problem that is intimately related to the one defined above. The auxiliary problem statement is: *Given a plant G and positive real numbers α and γ solve the problem*

$$\begin{aligned} &\inf_{Q\in\ell_1^{n_u\times n_y}} \|H^{22}-U^2*Q*V^2\|_2^2 \\ \textit{subject to}\quad & \\ &\|H^{11}-U^1*Q*V^1\|_1\leq\gamma, \\ &\|Q\|_1\leq\alpha. \end{aligned} \tag{4.1}$$

The optimal value obtained from the above problem is denoted by ν.

Note that in the problem statement of μ the allowable Youla parameter Q, which is in $\ell_1^{n_u\times n_y}$, needs to satisfy $\|H^{11}-U^1*Q*V^1\|_1\leq\gamma$. Therefore, it follows that $\|U^1*Q*V^1\|_1=\|H^{11}-U^1*Q*V^1-H^{11}\|_1\leq\|H^{11}-U^1*Q*V^1\|_1+\|H^{11}\|_1\leq \|H^{11}\|_1+\gamma$. Suppose $\hat{U}^1$ has more rows than columns and $\hat{V}^1$ has more columns than rows and both have full normal rank. Thus the left inverse of $\hat{U}^1$ exists (given by $(\hat{U}^1)^{-1}$) and the right inverse of $\hat{V}^1$ exists (given by $(\hat{V}^1)^{-r}$). Further, suppose that $\hat{U}^1$ and $\hat{V}^1$ have no zeros on the unit circle. Then it can be shown (see Lemma 4.2 and the discussion below) that there exists a β (which depends only on $(\hat{U}^1)^{-1}$ and $(\hat{V}^1)^{-r}$) such that $\|Q\|_1\leq\beta$. Thus if in the auxiliary problem we choose $\alpha\geq\beta$, then the constraint $\|Q\|_1\leq\alpha$ is redundant in the problem statement of ν and we get $\mu=\nu$. The extra constraint in the problem statement of ν is useful because it regularizes the problem (as will be seen). The following lemma is useful in estimating β.

LEMMA 4.1 (see [1]). *Let* $\operatorname{int}(\mathcal{D})$ *denote the interior of the unit disc in the complex plane. Given a function $\hat{f}(\,.\,)$ of the complex variable λ analytic in* $\operatorname{int}(\mathcal{D})$, *then* $\frac{d^kf}{d\lambda}|_{\lambda_0}=0$ *for $k=0,1,\ldots,(\sigma-1)$ and $\lambda_0\in\operatorname{int}(\mathcal{D})$ if and only if $\hat{f}(\lambda)=(\lambda-\lambda_0)^\sigma\hat{g}(\lambda)$, where $g(\,.\,)$ is analytic in* $\operatorname{int}(\mathcal{D})$.

LEMMA 4.2. *Let ϕ be an element of ℓ_1 such that $\|\phi\|_1\leq\gamma$ for some $\gamma>0$. Let $\hat{\phi}(\lambda)$ be the λ transform of ϕ. Suppose, $\hat{\phi}(\lambda)$ has a zero at $\lambda=a$ with $|a|<1$. If $\hat{\phi}(\lambda)=(\lambda-a)\hat{\psi}(\lambda)$, then $\|\hat{\psi}(\lambda)\|_1\leq\frac{\gamma}{1-|a|}$.*

Proof. As $\|\phi\|_1\leq\gamma$ it follows that $\|(\lambda-a)\hat{\psi}(\lambda)\|_1\leq\gamma$. This implies that $\sum_{t=-\infty}^{\infty}|\psi(t-1)-a\psi(t)|\leq\gamma$. This is true only if $\sum_{t=-\infty}^{\infty}(|\psi(t-1)|-|a|\,|\psi(t)|)\leq\gamma$, which implies that $\|\psi\|_1(1-|a|)\leq\gamma$. Therefore, $\|\psi\|_1\leq\frac{\gamma}{1-|a|}$. □

In the discussion above we have obtained an upper bound on the one norm of $R:=U^1*Q*V^1$ for any $Q\in\ell_1^{n_u\times n_y}$, which satisfies $\|H^{11}-U^1*Q*V^1\|_1\leq\gamma$. As U^1 and V^1 are left and right invertible it follows that $\hat{Q}=(\hat{U}^1)^{-l}\hat{R}(\hat{V}^1)^{-r}$. As Q is in $\ell_1^{n_u\times n_y}$ it is true that $\hat{R}$ interpolates the unstable poles of $(\hat{U}^1)^{-l}$ and $(\hat{V}^1)^{-r}$ none of which are on the unit circle by assumption. Using Lemma 4.2 one can obtain an upper bound on the one norm of Q that depends only on the upper bound of the one norm of R, $(U^1)^{-l}$, and $(V^1)^{-r}$.

The following lemma is a result on the uniqueness of the solution to (4.1).

LEMMA 4.3. *Let $Q^0\in\ell_1^{n_u\times n_y}$ be a solution to* (4.1). *Let $\Phi^0=H-U*Q^0*V$ with $\Phi^{22,o}=H^{22}-U^2*Q^0*V^2$ and $\Phi^{11,o}=H^{11}-U^1*Q^0*V^1$. Then $\Phi^{22,o}$ is unique. Furthermore, if $\hat{U}^2$ and $\hat{V}^2$ have full normal column and row ranks, respectively, then Q^0 is unique.*

Proof. Note that the problem statement of ν given by (4.1) can be recast as

$$\nu=\inf\{\|\Phi^{22}\|_2^2:\Phi^{22}\in A_{al}\}, \tag{4.2}$$

where A_{al} is the following set:

$$\{\Phi^{22} : \text{there exists } Q \in \ell_1^{n_u \times n_y} \text{ with } \Phi^{22} = H^{22} - U^2 * Q * V^2, \\ \|H^{11} - U^1 * Q * V^1\|_1 \leq \gamma, \text{ and } \|Q\|_1 \leq \alpha\}.$$

Its clear that A_{al} is a convex set. It is also true that $\|\,.\,\|_2^2$ is a strictly convex function. It follows from Lemma 3.3 that the minimizer of (4.2) given by $\Phi^{22,o}$, if it exists, is unique. If $\hat{U}^2$ and $\hat{V}^2$ have full column and row ranks, then it follows that

$$\hat{Q}^0 = (\hat{U}^2)^{-l} \hat{\Phi}^{22,o} (\hat{V}^2)^{-r},$$

where $(\hat{U}^2)^{-l}$ and $(\hat{V}^2)^{-r}$ represent the left and the right inverses of $\hat{U}^2$ and $\hat{V}^2$, respectively. Thus $\hat{Q}^0$ is unique. This proves the lemma. □

5. Converging lower and upper bounds.

5.1. Converging lower bounds.

5.1. Converging lower bounds. Let ν_n be defined by

$$(5.1) \qquad \begin{aligned} &\inf_{Q \in \ell_1^{n_u \times n_y}} \|P_n(H^{22} - U^2 * Q * V^2)\|_2^2 \\ \text{subject to} \quad & \\ &\|P_n(H^{11} - V^1 * Q * V^1)\|_1 \leq \gamma, \\ &\|Q\|_1 \leq \alpha. \end{aligned}$$

It is clear that only the parameters of $Q(0), \ldots, Q(n)$ enter into the optimization problem and therefore (5.1) is a finite-dimensional quadratic programming problem. Once optimal $Q(0), \ldots, Q(n)$ are found, $Q = \{Q(0), \ldots, Q(n), 0, \ldots\}$ will be a finite impulse response (FIR) optimal solution to (5.1).

THEOREM 5.1. *Suppose the constraint set in problem* (4.1) *is nonempty. Then problem* (4.1) *always has an optimal solution* $Q^0 \in \ell_1^{n_u \times n_y}$. *Furthermore,*

$$\nu_n \nearrow \nu.$$

Also, if $\Phi^{22,o} := H^{22} - U^2 * Q^0 * V^2$ *and* $\Phi^{22,n} := H^{22} - U^2 * Q^n * V^2$, *where* Q^n *is a solution to* (5.1), *then there exists a subsequence* $\{\Phi^{22,n_m}\}$ *of the sequence* $\{\Phi^{22,n}\}$ *such that*

$$\|\Phi^{22,n_m} - \Phi^{22,o}\|_2 \to 0 \text{ as } m \to \infty.$$

If $\hat{U}^2$ *and* $\hat{V}^2$ *have full normal column and row ranks, respectively, then* Q^0 *is unique and*

$$\|\Phi^{22,n} - \Phi^{22,o}\|_2 \to 0 \text{ as } n \to \infty.$$

Proof. We know that for any $Q \in \ell_1^{n_u \times n_y}$, $\|P_n(H^{11} - U^1 * Q * V^1)\|_1 \leq \|P_{n+1}(H^{11} - U^1 * Q * V^1)\|_1$, and $\|P_n(H^{22} - U^2 * Q * V^2)\|_2^2 \leq \|P_{n+1}(H^{22} - U^2 * Q * V^2)\|_2^2$. Therefore, $\nu_n \leq \nu_{n+1}$ for all $n = 1, 2, \ldots$. Thus $\{\nu_n\}$ forms an increasing sequence. Similarly it can be shown that for all n, $\nu_n \leq \nu$.

For $n = 1, 2, \ldots$, let $\{Q^n\} \in \ell_1^{n_u \times n_y}$ be FIR solutions of (5.1). As the sequence $\{Q^n\}$ is uniformly bounded by α in $\ell_1^{n_u \times n_y}$ it follows from the Banach–Alaoglu theorem that there exists a subsequence $\{Q^{n_m}\}$ of $\{Q^n\}$ and $Q^0 \in \ell_1^{n_u \times n_y}$ such that

$Q_{ij}^{n_m} \to Q_{ij}^0$ in the $W(c_0^*, c_0)$ topology. This implies that $Q^{n_m}(t) \to Q^0(t)$ for all $t = 0, 1, \dots$. Therefore, for all n, $P_n(U * Q^{n_m} * V)$ converges to $P_n(U * Q^0 * V)$ as $m \to \infty$. Now for any $n > 0$ and for any $n_m > n$, $\|P_n(H^{11} - U^1 * Q^{n_m} * V^1)\|_1 \leq \gamma$. This implies that $\|P_n(H^{11} - U^1 * Q^0 * V^1)\|_1 \leq \gamma$. Since n is arbitrary, we have

$$\|H^{11} - U^1 * Q^0 * V^1\|_1 \leq \gamma.$$

Similarly for any $n > 0$ and for any $n_m > n$, $\|P_n(H^{22} - U^2 * Q^{n_m} * V^2)\|_2^2 \leq \nu$. Again, this implies that $\|P_n(H^{22} - U^2 * Q^0 * V^2)\|_2^2 \leq \nu$. Since n is arbitrary, it follows that

$$\|H^{22} - U^2 * Q^0 * V^2\|_2^2 \leq \nu.$$

It follows that Q^0 is an optimal solution for (4.1).

To prove that $\nu_n \nearrow \nu$, we note that

$$\|P_n(H^{22} - U^2 * Q^{n_m} * V^2)\|_2^2 \leq \|P_{n_m}(H^{22} - U^2 * Q^{n_m} * V^2)\|_2^2 = \nu_{n_m} \qquad \forall n > 0,\ \forall n_m > n.$$

Taking the limit as m goes to infinity we have

$$\|P_n(H^{22} - U^2 * Q^0 * V^2)\|_2^2 \leq \lim_{m \to \infty} \nu_{n_m} \ \forall n > 0.$$

It follows that

$$\|H^{22} - U^2 * Q^0 * V^2\|_2^2 \leq \lim_{m \to \infty} \nu_{n_m}.$$

Thus we have shown that $\lim_{m \to \infty} \nu_{n_m} = \nu$. Since ν_n is a monotonically increasing sequence, it follows that $\nu_n \nearrow \nu$.

It is clear from Lemma 4.3 that $\Phi^{22,o} := H^{22} - U^2 * Q^0 * V^2$ is unique. If $\Phi^{22,n} := P_n(H^{22} - U^2 * Q^n * V^2)$, then from the discussion above it follows that $\nu_{n_m} = \|\phi^{22,n_m}\|_2^2$ converges to $\nu = \|\Phi^{22,o}\|_2^2$. Also, $\Phi^{22,n_m}(t)$ converges to $\Phi^{22,o}(t)$. It follows from Lemma 3.5 that

$$\|\Phi^{22,n_m} - \Phi^{22,o}\|_2 \to 0 \text{ as } m \to \infty.$$

From Lemma 4.3 we also have that if $\hat{U}^2$ and $\hat{V}^2$ have full normal column and row ranks, respectively, then Q^0 is unique. From the uniqueness of Q^0 it follows that the original sequence $\{\Phi^{22,n}\}$ converges to $\Phi^{22,o}$ in the two norm. This proves the theorem. □

5.2. Converging upper bounds. Let $\nu^n(\gamma)$ be defined by

$$\begin{aligned} &\inf_{Q \in \ell_1^{n_u \times n_y}} \|H^{22} - U^2 * Q * V^2\|_2^2 \\ &\text{subject to} \\ &\|H^{11} - U^1 * Q * V^1\|_1 \leq \gamma, \\ &\|Q\|_1 \leq \alpha, \\ &Q(k) = 0 \text{ if } k > n. \end{aligned} \tag{5.2}$$

We will assume that γ, which characterizes the ℓ_1 constraint level, is in the interior of the domain of the function ν. The following theorem shows that $\{\nu^n(\gamma)\}$ defines a sequence of upper bounds to $\nu(\gamma)$ which converge to $\nu(\gamma)$.

THEOREM 5.2. *For all n, $\nu^n(\gamma) \geq \nu^{n+1}(\gamma) \geq \nu(\gamma)$. Also*

$$\nu^n(\gamma) \searrow \nu(\gamma).$$

Proof. It is clear that $\nu^n(\gamma) \geq \nu^{n+1}(\gamma)$ because any $Q \in \ell_1^{n_u \times n_y}$ that satisfies the constraints in the problem definition of $\nu^n(\gamma)$ will satisfy the constraints in the problem definition of $\nu^{n+1}(\gamma)$. For the same reason we also have $\nu^n(\gamma) \geq \nu(\gamma)$ for all relevant n.

Thus $\{\nu^n(\gamma)\}$ is a decreasing sequence of real numbers bounded below by $\nu(\gamma)$. It can be shown that $\nu(\gamma)$ is a continuous function of γ (see Theorem 6.5 in [5]).

Given $\epsilon > 0$ choose $\delta > 0$ such that

$$\nu(\gamma - \delta) - \nu(\gamma) < \frac{\epsilon}{2}. \tag{5.3}$$

Such a δ exists from the continuity of $\nu(\gamma)$ in γ. Let $Q^{\gamma-\delta}$ be a solution to the problem $\nu(\gamma - \delta)$, which is guaranteed to exist from Theorem 5.1. Let M be large enough so that $m \geq M$ implies that

$$| \|H^{22} - U^2 * P_m(Q^{\gamma-\delta}) * V^2\|_2^2 - \|H^{22} - U^2 * Q^{\gamma-\delta} * V^2\|_2^2| < \frac{\epsilon}{2} \tag{5.4}$$

and

$$| \|H^{11} - U^1 * P_m(Q^{\gamma-\delta}) * V^1\|_1 - \|H^{11} - U^1 * Q^{\gamma-\delta} * V^1\|_1| < \frac{\delta}{2}. \tag{5.5}$$

As $Q^{\gamma-\delta}$ is a solution to the problem $\nu(\gamma - \delta)$ it is also true that

$$\|H^{22} - U^2 * Q^{\gamma-\delta} * V^2\|_2^2 = \nu(\gamma - \delta),$$
$$\|H^{11} - U^1 * Q^{\gamma-\delta} * V^1\|_1 \leq \gamma - \delta,$$

and

$$\|Q^{\gamma-\delta}\|_1 \leq \alpha.$$

From the above and (5.4), (5.5) it follows that for all $m \geq M$,

$$\|H^{22} - U^2 * P_m(Q^{\gamma-\delta}) * V^2\|_2^2 - \nu(\gamma - \delta) \leq \frac{\epsilon}{2}, \tag{5.6}$$

$$\|H^{11} - U^1 * P_m(Q^{\gamma-\delta}) * V^1\|_1 \leq \gamma, \tag{5.7}$$

and

$$\|P_m(Q^{\gamma-\delta})\|_1 \leq \alpha. \tag{5.8}$$

From (5.3) and the above it follows that for all $m \geq M$, $P_m(Q^{\gamma-\delta})$ satisfies all the constraints of problem $\nu^m(\gamma)$ and

$$\|H^{22} - U^2 * P_m(Q^{\gamma-\delta}) * V^2\|_2^2 - \frac{\epsilon}{2} - \nu(\gamma) \leq \frac{\epsilon}{2}.$$

Thus for all $m \geq M$ it follows that

$$\nu^m(\gamma) - \nu(\gamma) \leq \epsilon.$$

This proves the theorem. □

6. Conclusions. In this paper, we have formulated a problem that incorporates the $\mathcal{H}_2$ performance measure and the ℓ_1 measure. It is shown that converging upper and lower bounds can be obtained via finite-dimensional convex programming problems. This methodology avoids many of the problems of the zero interpolation based methods previously employed.

Ongoing research has indicated that the method developed here can be generalized to solve multiple-objective problems that involve the $\mathcal{H}_2$ measure and various time domain measures (including the ℓ_1 norm). Future research involves implementation of the method developed.

REFERENCES

[1] M. A. Dahleh and I. J. Diaz-Bobillo, *Control of Uncertain Systems: A Linear Programming Approach*, Prentice–Hall, Englewood Cliffs, NJ, 1995.

[2] N. Elia and M. A. Dahleh, *Controller design with multiple objectives*, IEEE Trans. Automat. Control, 42 (1997), pp. 596–613.

[3] M. Khammash, *Solution of the ℓ_1 MIMO control problem without zero interpolation*, in Proceedings of the IEEE Conference on Decision and Control, Kobe, Japan, IEEE, Piscataway, NJ, 1996, pp. 4040–4045.

[4] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley and Sons, Inc., New York, 1969.

[5] M. V. Salapaka, M. Dahleh, and P. Voulgaris, *Mixed objective control synthesis: Optimal $\ell_1/\mathcal{H}_2$ control*, SIAM J. Control Optim., 35 (1997), pp. 1672–1689.

[6] M. V. Salapaka, M. Dahleh, and P. Voulgaris, *MIMO optimal control design: The interplay of the $\mathcal{H}_2$ and the ℓ_1 norms*, IEEE Trans. Automat. Control, 43 (1998), pp. 1374–1388.

[7] D. C. Youla, H. A. Jabr, and J. J. Bongiorno, *Modern Wiener–Hopf design of optimal controllers—Part 2: The multivariable case*, IEEE Trans. Automat. Control, 21 (1976), pp. 319–338.

SIAM J. Control Optim.
Vol. 37, No. 6, pp. 1874–1896

ON THE NULL ASYMPTOTIC STABILIZATION OF THE TWO-DIMENSIONAL INCOMPRESSIBLE EULER EQUATIONS IN A SIMPLY CONNECTED DOMAIN*

JEAN-MICHEL CORON†

Abstract. We study the asymptotic stabilization of the origin for the two-dimensional (2-D) Euler equation of incompressible inviscid fluid in a bounded domain. We assume that the controls act on an arbitrarily small nonempty open subset of the boundary. We prove the null global asymptotic stabilizability by means of explicit feedback laws if the domain is connected and simply connected.

Key words. inviscid fluid, stabilization, nonlinear control

AMS subject classifications. 76C99, 93B52, 93C20, 93D15

PII. S036301299834140X

1. Introduction.

In previous papers [2, 3] we have considered the problem of the controllability of the two-dimensional (2-D) Euler equation of incompressible inviscid fluid in a bounded domain. In particular, we have proved that, if the controls act on an arbitrarily small open subset of the boundary which meets every connected component of this boundary, then the 2-D Euler equation is exactly controllable. This result has been extended recently by Glass to the three-dimensional (3-D) Euler equation in [9, 10].

For linear control systems, the exact controllability implies the asymptotic stabilizability by means of feedback laws. This is well known for linear control systems of finite dimension and, by Slemrod [22], J.-L. Lions [17], Lasiecka–Triggiani [16], and Komornik [15], it also holds in infinite dimension in very general cases. But, as pointed out by Sussmann in [24], Sontag–Sussman in [23], and Brockett in [1], this is no longer true for *nonlinear* control systems, even of finite dimension. Let us also notice that, as in the counterexample of [1], the linearized control system of the Euler equation around the origin is not controllable.

Therefore it is natural to ask what is the situation for the asymptotic stabilizability of the origin for the 2-D Euler equation of incompressible inviscid fluid in a bounded domain when the controls act on an arbitrarily small open subset of the boundary which meets any connected component of this boundary. In this paper we prove the null global asymptotic stabilizability by means of feedback laws if the domain is simply connected.

Our paper is organized as follows.

- In section 2, we give explicit feedback laws which globally asymptotically stabilize the origin and state our main results.
- In sections 3 and 4, we give the proofs of our main results.

2. Explicit stabilizing feedbacks.

Let Ω be a nonempty bounded connected and simply connected subset of $\mathbb{R}^2$ of class C^∞ and let γ be a nonempty open subset of the boundary $\partial\Omega$ of Ω. This set γ is the location of the control. Let y be the velocity

*Received by the editors July 6, 1998; accepted for publication (in revised form) December 21, 1998; published electronically October 8, 1999. This research was partially supported by the GdR “Automatique.”

http://www.siam.org/journals/sicon/37-6/34140.html

†Analyse Numérique et EDP, Université de Paris-Sud, Bâtiment 425, 91405 Orsay Cedex, France (Jean-Michel.Coron@math.u-psud.fr).

field of the inviscid fluid contained in Ω. We assume that the fluid is incompressible, so that

$$\text{div } y = 0. \tag{2.1}$$

Since Ω is simply connected, y is completely characterized by $\omega := \text{curl } y$ and $y \cdot n$ on $\partial\Omega$, where n denotes the unit outward normal to $\partial\Omega$. For the problem of controllability, one does not really need to specify the control and the state: one considers the "Euler control system" as an underdetermined system by requiring $y \cdot n = 0$ on $\partial\Omega \setminus \gamma$ instead of $y \cdot n = 0$ on $\partial\Omega$ as for the uncontrolled usual Euler equation. For the stabilization problem, one needs to specify more precisely the control and the state. In this paper the state is ω. For the control there are at least two natural possibilities:

(a) The control is $y \cdot n$ on γ and the time derivative $\partial\omega/\partial t$ of the vorticity at the points of γ where $y \cdot n < 0$, i.e., at the points where the fluid enters into the domain Ω.

(b) The control is $y \cdot n$ on γ and the vorticity ω at the points where $y \cdot n < 0$.

Let us point out that, by (2.1), in both cases $y \cdot n$ has to satisfy $\int_{\partial\Omega} y \cdot n = 0$.

2.1. Case where one controls the time derivative of the vorticity of the incoming flow.

In this subsection we concentrate on case (a); for case (b), see subsection 2.2.

Let us give our stabilizing feedback laws. Let $g \in C^\infty(\partial\Omega)$ be such that

$$\text{Support } g \subset \gamma, \tag{2.2}$$

$$\gamma_+ := \{g > 0\} \text{ and } \gamma_- := \{g < 0\} \text{ are connected}, \tag{2.3}$$

$$g \neq 0, \tag{2.4}$$

$$\overline{\gamma_+} \cap \overline{\gamma_-} = \emptyset, \tag{2.5}$$

$$\int_{\partial\Omega} g = 0. \tag{2.6}$$

For any compact set K of $\mathbb{R}^q$ and any $f \in C^0(K; \mathbb{R}^m)$, we denote

$$|f|_{0,K} = \text{Max } \{|f(x)|\,; x \in K\}.$$

For simplicity, we write $|f|_0$ instead of $|f|_{0,\overline{\Omega}}$. Our stabilizing feedback laws are

$$y \cdot n = M\, |\omega|_0\, g \text{ on } \gamma,$$

$$\frac{\partial\omega}{\partial t} = -M\, |\omega|_0\, \omega \text{ on } \gamma_- \text{ if } |\omega|_0 \neq 0,$$

where $M > 0$ is large enough. With these feedback laws, a function $\omega : I \times \overline{\Omega} \to \mathbb{R}$, where I is an interval, is a solution of the closed loop system Σ if

$$\frac{\partial\omega}{\partial t} + \text{ div } (\omega y) = 0 \text{ in } \overset{\circ}{I} \times \Omega, \tag{2.7}$$

$$\text{div } y = 0 \text{ in } \overset{\circ}{I} \times \Omega, \tag{2.8}$$

$$\text{curl } y = \omega \text{ in } \overset{\circ}{I} \times \Omega, \tag{2.9}$$

$$y(t) \cdot n = M\, |\omega(t)|_0\, g \text{ on } \partial\Omega\, \forall t \in I, \tag{2.10}$$

$$\frac{\partial\omega}{\partial t} = -M\, |\omega(t)|_0\, \omega \text{ on } \{t;\, \omega(t) \neq 0\} \times \gamma_-, \tag{2.11}$$

where, for $t \in \Omega$, $\omega(t) : \overline{\Omega} \to \mathbb{R}$ and $y(t) : \overline{\Omega} \to \mathbb{R}^2$ are defined by requiring $\omega(t)(x) = \omega(t,x), y(t)(x) = y(t,x), \forall x \in \overline{\Omega}$. More precisely, the definition of a solution of system Σ is as follows.

Definition 2.1. *Let I be an interval. A function $\omega : I \to C^0(\overline{\Omega})$ is a solution of system Σ if*

(i) $\omega \in C^0(I; C^0(\overline{\Omega}))(\cong C^0(I \times \overline{\Omega}))$,

(ii) *for $y \in C^0(I \times \overline{\Omega}; \mathbb{R}^2)$ defined by requiring* (2.8) *and* (2.9) *in the sense of distributions and* (2.10), *one has* (2.7) *in the sense of distributions,*

(iii) *in the sense of distributions on the open manifold $\{t \in \overset{\circ}{I};\ \omega(t) \neq 0\} \times \gamma_-$ one has $\partial\omega/\partial t = -M\,|\omega(t)|_0\,\omega$.*

Our first theorem says that, for M large enough, the Cauchy problem for system Σ has at least one solution defined on $[0,+\infty)$ for any initial data in $C^0(\overline{\Omega})$. More precisely one has the following.

Theorem 2.2. *There exists $M_0 > 0$ such that, for any $M \geqslant M_0$, the following two properties hold:*

(i) *For any $\omega_0 \in C^0(\overline{\Omega})$, there exists a solution of system Σ defined on $[0,+\infty)$ such that $\omega(0) = \omega_0$.*

(ii) *Any maximal solution of system Σ defined at time* 0 *is defined on* $[0,+\infty)$ *(at least).*

Remark 2.3. (a) *In this theorem, property* (i) *is in fact implied by property* (ii) *and Zorn's lemma. We state* (i) *in order to emphasize the existence of a solution to the Cauchy problem for system Σ.* (b) *We do not know if the solution to the Cauchy problem is unique for positive time. (For negative time, one does not have uniqueness since there are solutions ω of system Σ defined on $[0,+\infty)$ such that $\omega(0) \neq 0$ and $\omega(T) = 0$ for $T \in [0,+\infty)$ large enough.) But let us emphasize that, already for control systems in finite dimension, one considers feedback laws which are merely continuous; with these feedback laws, the Cauchy problem for the closed loop system may have many solutions. It turns out that this lack of uniqueness is not a real problem. Indeed, in finite dimension at least, if a point is asymptotically stable for a continuous vector field, then there exists, as in the case of regular vector fields, a (smooth) strict Lyapunov function. This result is due to Kurzweil* [13]. *It is tempting to conjecture that a similar result holds in infinite dimension under reasonable assumptions. The existence of this Lyapunov function ensures some robustness to perturbations. It is precisely this robustness which makes the interest of feedback laws compared to open loop controls. We will see that, for our feedback laws, there exists also a strict Lyapunov—see Proposition* 3.6 *below—and therefore our feedback laws provide some kind of robustness.*

Our next theorem shows that, at least for M large enough, our feedback laws globally and strongly asymptotically stabilize the origin in $C^0(\overline{\Omega})$ for system Σ.

Theorem 2.4. *There exists a positive constant $M_1 \geqslant M_0$ such that, for any $\varepsilon \in (0,1]$, any $M \geqslant M_1/\varepsilon$, and any maximal solution ω of system Σ defined at time* 0,

$$|\omega(t)|_0 \leqslant \text{Min } \left\{|\omega(0)|_0, \frac{\varepsilon}{t}\right\}\ \forall t > 0. \tag{2.12}$$

Remark 2.5. *Due to the term $|\omega(t)|_0$ appearing in* (2.10) *and in* (2.11) *our feedback laws do not depend only on the value of ω on γ. Let us point out that there is no asymptotically stabilizing feedback law depending only on the value of ω on γ such that the origin is asymptotically stable for the closed loop system. In fact, given a nonempty open subset Ω_0 of Ω, there is no feedback law which does not depend on the values of ω on Ω_0. This phenomenon is due to the existence of "phantom vortices":*

there are smooth stationary solutions $\bar{y} : \overline{\Omega} \to \mathbb{R}^2$ *of the* 2*-D Euler equations such that* Support $\bar{y} \subset \Omega_0$ *and* $\bar{\omega} :=$ curl $\bar{y} \neq 0$*; see, e.g.,* [20]*. Then* $\omega(t) = \bar{\omega}$ *is a solution of the closed loop system if the feedback law does not depend on the values of* ω *on* Ω_0 *and vanishes for* $\omega = 0$*.*

Remark 2.6. *Let us emphasize that* (2.12) *implies that*

$$|\omega(t)|_0 \leqslant \varepsilon \quad \forall t \geqslant 1, \tag{2.13}$$

for any maximal solution ω *of system* Σ *defined at time 0 (whatever* $\omega(0)$ *is). It would be interesting to know if one could have a similar result for the* 2*-D Navier–Stokes equations of viscous incompressible flows, that is, if given* $\varepsilon > 0$*, does there exist a feedback law such that* (2.13) *holds for any solution of the closed loop Navier–Stokes control system? Note that* $y = 0$ *on* γ *is a feedback which leads to asymptotic stabilization of the null solution of the Navier–Stokes control system. But this feedback does not have the required property. For recent results on the controllability of the Navier–Stokes control system, see the papers by Imanuvilov and Fursikov* [6, 7, 8] *and the paper by Imanuvilov* [12] *as well as* [4, 5]*.*

2.2. Case where one controls the vorticity of the incoming flow. In this subsection one does no longer control the time-derivative of the vorticity of the incoming flow but the vorticity of the incoming flow itself. Therefore the control is $y \cdot n$ on γ with the constraint $\int_{\partial\Omega} y \cdot n = 0$ (as above) and the vorticity ω at the points where $y \cdot n < 0$. Of course in this new situation one cannot take the state ω in $C^0(\overline{\Omega})$: if the state ω is in $C^0(\overline{\Omega})$, then it will determine part of the control, namely, the vorticity of the incoming flow. It is therefore natural to consider the state ω as being in $L^\infty(\Omega)$.

For any measurable subset B of $\mathbb{R}^q$ and any $f \in L^\infty(B; \mathbb{R}^m)$, we denote by $|f|_{\infty,B}$ the essential supremum of f on B. For simplicity, we write $|f|_\infty$ instead of $|f|_{\infty,\overline{\Omega}}$. Our stabilizing feedback laws are

$$y \cdot n = M\, |\omega|_\infty\, g \text{ on } \gamma, \tag{2.14}$$

$$\omega = 0 \text{ on } \gamma_- \text{ if } |\omega|_\infty \neq 0, \tag{2.15}$$

where, again, $M > 0$ is large enough. With these feedback laws, a function $\omega : I \times \overline{\Omega} \to \mathbb{R}$, where I is an interval, is a solution of the closed loop system that we call Σ_1, if one has (2.7), (2.8), (2.9) and

$$y(t) \cdot n = M\, |\omega(t)|_\infty\, g \text{ on } \partial\Omega \text{ for almost every } t \in I, \tag{2.16}$$

$$\omega = 0 \text{ on } \{t;\, \omega(t) \neq 0\} \times \gamma_-. \tag{2.17}$$

Since $\omega(t)$ is only in $L^\infty(\Omega)$, the meaning of (2.17) has to be specified. As usual, (2.17) has to be understood in a "weak sense," which is obtained by multiplying (2.7) by suitable smooth test functions, integrating on $I \times \Omega$, and performing integration by parts. More precisely, the definition of a solution of system Σ_1 is as follows.

Definition 2.7. *Let* I *be an interval. A function* $\omega : I \to L^\infty(\Omega)$ *is a solution of system* Σ_1 *if*

(i) $\omega \in C^0(I; H^{-1}(\Omega))$,

(ii) $\omega \in L^\infty_{loc}(I; L^\infty(\Omega)) \cong L^\infty_{loc}(I \times \overline{\Omega})$,

(iii) *for any* $\varphi \in C^1(I \times \overline{\Omega})$ *with compact support such that*

$$\text{Support } \varphi \subset \left(\overset{\circ}{I} \times \Omega\right) \cup \left(\{t \in \overset{\circ}{I};\, |\omega(t)|_\infty > 0\} \times \gamma_-\right), \tag{2.18}$$

one has

$$\int_{I\times\Omega}\left(\omega\frac{\partial\varphi}{\partial t}+\omega(y\cdot\nabla)\varphi\right)=0, \tag{2.19}$$

where $y\in L^\infty\left(I;C^0\left(\overline{\Omega}\right)\right)$ *is defined by requiring* (2.16) *and, in the sense of distributions on* $\overset{\circ}{I}\times\Omega$, (2.8) *and* (2.9).

Our first theorem says that, for M large enough, the Cauchy problem for system Σ_1 has at least one solution defined on $[0,+\infty)$ for any initial data in $L^\infty(\Omega)$. More precisely, one has the following.

THEOREM 2.8. *There exists* $M_2\geqslant 0$ *such that, for any* $M\geqslant M_2$, *the following two properties hold.*

(i) *For any* $\omega_0\in L^\infty(\Omega)$, *there exists a solution of system* Σ_1 *defined on* $[0,+\infty)$ *such that* $\omega(0)=\omega_0$.

(ii) *Any maximal solution of system* Σ_1 *defined at time* 0 *is defined on* $[0,+\infty)$ *(at least).*

Our next theorem, which is analogous to Theorem 2.4, tells us that, at least for M large enough, feedbacks laws (2.14) and (2.15) globally and strongly asymptotically stabilize the origin in $L^\infty(\Omega)$ for system Σ_1.

THEOREM 2.9. *There exists a positive constant* $M_3>0$ *such that, for any* $\epsilon\in(0,1]$, *any* $M\geqslant M_3/\varepsilon$, *and any maximal solution of system* Σ_1 *defined at time* 0, *one has*

$$|\omega(t)|_\infty\leqslant \text{ Min }\left\{|\omega(0)|_\infty,\frac{\varepsilon}{t}\right\}\ \forall t>0.$$

The proof of Theorem 2.9 is very similar to the proof of Theorem 2.4 and therefore is omitted. The end of this paper is organized as follows.

- In section 3, we prove Theorems 2.2 and 2.4.
- In section 4, we prove Theorem 2.8.

3. Proof of Theorems 2.2 and 2.4.

3.1. Proof of Theorem 2.2. For a compact subset K and a function $y\in C^0(K;\mathbb{R}^2)$, we let

$$q_K(y):=|y|_0+\sup\{|y(x)-y(x')|/r(|x-x'|);(x,x')\in K^2,x\neq x'\},$$

where

$$r(s)=s+s\ln(1/s)\ \forall s\in(0,1),\ \text{and } r(s)=s\ \forall s\geqslant 1. \tag{3.1}$$

For simplicity, we write q instead of $q_{\overline{\Omega}}$. For technical reasons, it is useful to extend $y\in C^0(\overline{\Omega})$ outside $\overline{\Omega}$. Let $R>0$ be such that

$$\overline{\Omega}\subset B_{R/2}:=\{x\in\mathbb{R}^2;|x|<R/2\}.$$

Let $B_R:=\{x\in\mathbb{R}^2;|x|<R\}$. Let $\mathcal{P}:C^0(\overline{\Omega};\mathbb{R}^2)\to C^0(\overline{B_R};\mathbb{R}^2)$ be a continuous linear map such that

$$\mathcal{P}(y)(x)=y(x)\ \forall x\in\overline{\Omega}\ \forall y\in C^0(\overline{\Omega};\mathbb{R}^2), \tag{3.2}$$

$$\text{Support }\mathcal{P}(y)\subset\overline{B_{R/2}}\ \forall y\in C^0(\overline{\Omega};\mathbb{R}^2), \tag{3.3}$$

and such that, for some $C_0>0$,

$$q_{\overline{B_R}}(\mathcal{P}(y))\leqslant C_0q(y)\,\forall y\in C^0(\overline{\Omega};\mathbb{R}^2). \tag{3.4}$$

Let us recall the following important theorem, due to Wolibner [25] and Yudovich [26] (see also [14, Lemma 2.6]).

THEOREM 3.1 (see Wolibner [25] and Yudovich [26]). *Let T be a positive real number and let $y \in L^\infty\left((0,T);C^0\left(\overline{B_R};\mathbb{R}^2\right)\right)$ be such that*

(3.5)

$$y(t,x) := y(t)(x) = 0 \text{ for almost everywhere (a.e.) } (t,x) \in (0,T) \times \left(B_R \setminus B_{R/2}\right),$$

and, for some constant $K \in (0,+\infty)$,

$$q_{\overline{B_R}}(y(t)) \leqslant K \text{ for a.e. } t \in (0,T). \tag{3.6}$$

Then there exists one and only one map $\Phi^y \in C^0([0,T]\times[0,T]\times\overline{B_R};\overline{B_R})$, $(t,s,x) \to \Phi^y(t,s,x)$ such that

$$\Phi^y(t,s,x) = x + \int_s^t y(t',\Phi^y(t',s,x))dt' \,\forall (t,s,x) \in [0,T]\times[0,T]\times\overline{B_R}.$$

Moreover there exist two constants $C_1 = C_1(K,R,T) > 0$ and $\delta = \delta(K,R,T) > 0$, depending on K,R,T, such that, for any $(x,x') \in \overline{B_R}^2$, for any $(t,t',s,s') \in [0,T]^4$, and for any $y \in L^\infty\left((0,T);C^0\left(\overline{B_R};\mathbb{R}^2\right)\right)$ satisfying (3.5) *and* (3.6),

$$|\Phi^y(t',s',x') - \Phi^y(t,s,x)| \leqslant C_1(|s'-s|^\delta + |t'-t|^\delta + |x'-x|^\delta). \tag{3.7}$$

Our proof of Theorem 2.2 is divided in two parts.

- We first prove the existence of a solution to the Cauchy problem for small positive time.
- Then we prove that any maximal solution to Σ defined at time 0 is defined on $[0,+\infty)$.

So let us first start with the proof of the following proposition.

PROPOSITION 3.2. *There exists $M_0 > 0$ such that, for any $M \geqslant M_0$ and for any $\omega_0 \in C^0(\overline{\Omega})$, there exists $T > 0$ and a solution of system Σ defined on $[0,T]$ such that $\omega(0) = \omega_0$.*

Of course if $\omega_0 = 0$ one can take arbitrary $T > 0$ and choose $\omega = 0$. Therefore we may assume that $\omega_0 \neq 0$. Then there exists a point x^0 in Ω such that

$$\omega_0(x^0) \neq 0. \tag{3.8}$$

Let $M > 0$. Let $C_2 > 0$ (depending on M) be such that

$$|y|_0 \leqslant C_2 \tag{3.9}$$

for any $y \in C^0(\overline{\Omega};\mathbb{R}^2)$ such that

$$|\text{curl } y|_0 \leqslant |\omega_0|_0, \text{ div } y = 0,$$
$$|y \cdot n| \leqslant M|\omega_0|_0|g|_{0,\partial\Omega} \text{ on } \partial\Omega.$$

Let $\rho > 0$ be such that

$$B(x^0,\rho) := \{x \in \mathbb{R}^2; |x - x^0| < \rho\} \subset \Omega \tag{3.10}$$

and let

$$T = \rho/C_2. \tag{3.11}$$

Let us denote by $|\ |_{H^{-1}(\Omega)}$ one of the usual norm of the Sobolev space $H^{-1}(\Omega)$. Let $C_3 > 0$ be such that, for any $f \in L^\infty(\Omega; \mathbb{R}^2)$,

$$|\text{div } f|_{H^{-1}(\Omega)} \leqslant C_3 \, |f|_{L^\infty(\Omega)} \, . \tag{3.12}$$

Let also $C_4 > 0$ be such that, for any divergence free $f \in C^0(\overline{\Omega}; \mathbb{R}^2)$ with a bounded curl,

$$|f|_0 \leqslant C_4 \left(| \text{ curl } f|_{L^\infty(\Omega)} + |f \cdot n|_{0,\partial\Omega} \right) . \tag{3.13}$$

We are going to construct a solution $\omega \in C^0([0,T] \times \overline{\Omega})$ of system Σ satisfying the initial condition $\omega(0) = \omega_0$ as a fixed point of a map $F : X \to X$, where X is the set of functions $\omega \in C^0([0,T] \times \overline{\Omega})$ such that

$$\omega(0) = \omega_0, \tag{3.14}$$

$$t \in [0,T] \to |\omega(t)|_0 \text{ is nonincreasing,} \tag{3.15}$$

$$\left| \frac{\partial \omega}{\partial t} \right|_{L^\infty((0,T);H^{-1}(\Omega))} \leqslant C_3 C_4 |\omega_0|_0^2 \left(M |g|_{0,\partial\Omega} + 1 \right) . \tag{3.16}$$

Note that X is a closed convex subset of $C^0([0,T] \times \overline{\Omega})$ equipped with the sup-norm $|\ |_{0,[0,T]\times\overline{\Omega}}$. Let us define F. For $\omega \in X$, let us define $\tilde{y}_\omega \in C^0([0,T] \times \overline{\Omega}; \mathbb{R}^2)$ by requiring

$$\text{div } \tilde{y}_\omega = 0 \text{ in } (0,T) \times \Omega, \tag{3.17}$$

$$\text{curl } \tilde{y}_\omega = \omega \text{ in } (0,T) \times \Omega, \tag{3.18}$$

$$\tilde{y}_\omega(t,\cdot) \cdot n = M \text{ Max } \{ |\omega(t)|_0, |\omega_0(x^0)| \} g \text{ on } \partial\Omega \ \forall t \in [0,T]. \tag{3.19}$$

Note that, by (3.15), (3.14), (3.17), (3.18), (3.19), and a theorem due to Wolibner [25] (see also [14, Lemma 1.4]), there exists a constant C_5 such that

$$q(\tilde{y}_\omega(t,\cdot)) \leqslant C_5 \, \forall t \in [0,T] \ \forall \omega \in X. \tag{3.20}$$

Let $y_\omega \in C^0([0,T] \times \overline{B_R}; \mathbb{R}^2)$ be defined by

$$y_\omega(t,\cdot) = \mathcal{P}(\tilde{y}_\omega(t,\cdot)) \ \forall t \in [0,T]. \tag{3.21}$$

By (3.4) and (3.20),

$$q_{\overline{B_R}}(y_\omega(t,\cdot) \leqslant C_0 C_5 \ \ \forall \omega \in X. \tag{3.22}$$

In particular, by the Wolibner–Yudovich theorem (Theorem 3.1), there exists a flow Φ^{y_ω} associated with y_ω. For any interval I containing 0 such that $0 = \text{Min} I$ and for any $y \in L^\infty_{\text{loc}} \left(I; C^0 \left(\overline{B_R}; \mathbb{R}^2 \right) \right)$ satisfying (3.5) and such that $q_{\overline{B_R}}(y) \in L^\infty_{\text{loc}}(I)$, let us define $s_y : I \times \overline{\Omega} \to I$ by

$$s_y(t,x) = \text{ Max } \{ t' \in [0,t]; \ \Phi^y(t',t,x) \in \overline{\gamma_-} \}, \tag{3.23}$$

with the convention Max $\emptyset = 0$. Let us also define $a_y : I \times \overline{\Omega} \to \overline{\Omega}$ by

$$a_y(t,x) = \Phi^y(s_y(t,x),t,x). \tag{3.24}$$

With these notations, we can now define our map $F(\omega) : [0,T] \times \overline{\Omega} \to \mathbb{R}$ by

$$F(\omega)(t,x) = \omega_0 \left(a_{y_\omega}(t,x) \right) \exp \left(-M \int_0^{s_{y_\omega}(t,x)} |\omega(t')|_0 \, dt' \right) . \tag{3.25}$$

It follows from our construction of F (recall also (3.17)) that, in the sense of distributions,

$$\frac{\partial F(\omega)}{\partial t} + \text{ div } (F(\omega)y_\omega) = 0 \text{ in } (0,T) \times \Omega. \tag{3.26}$$

Indeed, let $\bar{t} \in (0,T)$ and $\bar{x} \in \Omega$. Then, for (t,x) close enough to $(\bar{t},\bar{x})$, one has

$$s_{y_\omega}(t,\Phi^{y_\omega}(t,\bar{t},x)) = s_{y_\omega}(\bar{t},x),$$
$$a_{y_\omega}(t,\Phi^{y_\omega}(t,\bar{t},x)) = a_{y_\omega}(\bar{t},x),$$

which imply that

$$F(\omega)(t,\Phi^{y_\omega}(t,\bar{t},x)) = F(\omega)(\bar{t},x). \tag{3.27}$$

But (3.27), together with (3.2), (3.17), (3.21), and standard smoothing procedures, gives (3.26) in the sense of distributions.

Let us now check that if $\omega \in X$ is a fixed point of F, then ω is a solution of system Σ which, by (3.14), satisfies the Cauchy initial data $\omega(0) = \omega_0$. Indeed, let ω be a fixed point of F and let $y = \tilde{y}_\omega$. Then, from (3.17) and (3.18), we get (2.8) and (2.9). From (3.26), we get (2.7). Let us also point out that, by (3.23),

$$s_{y_\omega}(t,x) = t \;\; \forall t \in [0,T] \, \forall x \in \gamma_-. \tag{3.28}$$

From (3.24), (3.25), and (3.28), one gets

$$\omega(t,x) = F(\omega)(t,x) = \omega_0(x) \exp\left(-M \int_0^t |\omega(t')|_0 \, dt'\right) \; \forall (t,x) \in [0,T] \times \gamma_-,$$

which implies (2.11). It remains only to verify that (2.10) holds. By (3.19), it suffices to check that

$$\left|\omega_0(x^0)\right| \leqslant |\omega(t)|_0 \;\; \forall t \in [0,T]. \tag{3.29}$$

From the definition of C_2, (3.2), (3.14), (3.15), (3.17), (3.18), (3.19), and (3.21), one gets that $|y_\omega(t)|_0 \leqslant C_2$ for any $t \in [0,T]$, which, with (3.10) and (3.11), gives

$$\Phi^{y_\omega}(t,0,x^0) \in \Omega \;\; \forall t \in [0,T].$$

Therefore $s_{y_\omega}(t,\Phi^{y_\omega}(t,0,x_0)) = 0$ for any $t \in [0,T]$, which implies that

$$F(\omega)(t,\Phi^{y_\omega}(t,0,x^0)) = \omega_0(x^0) \, \forall t \in [0,T]. \tag{3.30}$$

From (3.30), one has

$$(|\omega(t)|_0 =) \, |F(\omega(t))|_0 \geqslant \left|\omega_0(x^0)\right| \; \forall t \in [0,T].$$

Therefore (3.29) holds and the fixed point ω of F is indeed a solution of system Σ.

By the Leray–Schauder fixed point theorem, in order to prove the existence of a fixed point to F, it suffices to check that

$$F(X) \subset X, \tag{3.31}$$

$$F \text{ is continuous,} \tag{3.32}$$

$$F(X) \text{ is relatively compact in } C^0([0,T] \times \overline{\Omega}). \tag{3.33}$$

Let us first check (3.31). Let $\omega \in X$. It follows directly from (3.25) that

$$|F(\omega)|_{L^\infty(\overline{\Omega}\times[0,T])} \leqslant |\omega_0|_0 \,.$$

Clearly

$$F(\omega)(0) = \omega(0) = \omega_0. \tag{3.34}$$

Let us check that

$$t \in [0,T] \to |F(\omega(t))|_0 \text{ is nonincreasing.} \tag{3.35}$$

Let $0 \leqslant t_1 \leqslant t_2 \leqslant T$ and let $x \in \overline{\Omega}$. If $s_{y_\omega}(t_2,x) \leqslant t_1$, one has

$$F(\omega)(t_2,x) = F(\omega)\left(t_1, \Phi^{y_\omega}(t_1,t_2,x)\right).$$

If $s_{y_\omega}(t_2,x) > t_1$, one has

$$F(\omega)(t_2,x) = F(\omega)\left(a_{y_\omega}(t_2,x), t_1\right) \exp - \left(M \int_{t_1}^{s_{y_\omega}(t_2,x)} |\omega(t')|_0 dt' \right).$$

In both cases

$$|F(\omega)(t_2,x)| \leqslant |F(\omega)(t_1)|_0,$$

which shows (3.35). From (3.12), (3.13), (3.17), (3.18), (3.19), (3.26), (3.34), and (3.35), one gets that

$$\left| \frac{\partial F(\omega)}{\partial t} \right|_{L^\infty((0,T);H^{-1}(\Omega))} \leqslant C_3 C_4 |\omega_0|_0^2 \left(M|g|_{0,\partial\Omega} + 1 \right).$$

Therefore, in order to prove (3.31), it suffices to check that $F(\omega)$ is continuous on $[0,T] \times \overline{\Omega}$. From the continuity of Φ^{y_ω} and the definition (3.23) of s_{y_ω}, it is clear that

$$s_{y_\omega} \text{ is upper semicontinuous on } [0,T] \times \overline{\Omega}. \tag{3.36}$$

Since the continuity of s_{y_ω} implies the continuity of F, it remains only to check that

$$s_{y_\omega} \text{ is lower semicontinuous on } [0,T] \times \overline{\Omega}. \tag{3.37}$$

In order to prove this lower semicontinuity, let us assume that the following lemma, proved in Appendix B, holds.

Lemma 3.3. *There exists $M_0 > 0$ such that, for any $T > 0$ and for any $y \in L^\infty\left((0,T); C^0\left(\overline{B_R}; \mathbb{R}^2\right)\right)$ satisfying* (3.5), (3.6) *for some $K > 0$, and for some function $\alpha \in L^\infty((0,T);(0,+\infty))$,*

$$y(t,\cdot)\cdot n = \alpha(t) g \textit{ on } \partial\Omega \textit{ for a.e. } t \in (0,T), \tag{3.38}$$

$$M_0 |\text{curl } y(t)|_{L^\infty(\Omega)} \leqslant \alpha(t) \textit{ for a.e. } t \in (0,T), \tag{3.39}$$

$$\text{div } y = 0 \textit{ in } (0,T) \times \Omega \tag{3.40}$$

for any $(\tilde{t},\tilde{x}) \in (0,T] \times \overline{\gamma_-}$ and for any $\nu \in (0,\tilde{t})$, there exists $t \in (\tilde{t}-\nu,\tilde{t})$, such that

$$\Phi^y\left(t,\tilde{t},\tilde{x}\right) \notin \overline{\Omega}.$$

Let us also point out that, by the definition of γ_-, (3.2), (3.19), (3.21), and (3.22), for any $\tilde{x} \in \overline{\Omega}$, if for $0 \leqslant t' \leqslant t \leqslant T$ $\Phi^{y_\omega}(t',t,\tilde{x})$ is not in $\overline{\Omega}$, then

$$\exists t'' \in [t',t] \text{ such that } \Phi^{y_\omega}(t',t,\tilde{x}) \in \overline{\gamma_-}. \tag{3.41}$$

This is indeed clear if y_ω is smooth enough, for example, locally Lipschitz with respect to x. The case where y is not smooth follows from the main ingredient, due to Wolibner [25], to prove the uniqueness of Φ^y in Theorem 3.1. Let us briefly sketch the proof. Let $W \in C^\infty(\mathbb{R}^2)$ be such that $\{W = 0\} = \partial\Omega$, $W > 0$ in Ω, $W < 0$ in $\mathbb{R}^2 \setminus \overline{\Omega}$ and ∇W does not vanish on $\partial\Omega$. Let $w(s) = W(\Phi^{y_\omega}(s,t,\tilde{x}))$. If (3.41) does not hold one easily sees, using (3.2), (3.19), (3.21), and (3.22), that there exists $C_6 > 0$ such that

$$\frac{\mathrm{d}w}{\mathrm{d}s} \leqslant C_6 r(|w|) \text{ on } [t', t], \tag{3.42}$$

where r is defined in (3.1). Since $w(t') < 0 \leqslant w(t)$, there exists $t_1 \in (t', t]$ such that $w(t_1) = 0$. Then, using (3.42) and the fact that $\int_0^1 /r(s)ds = +\infty$, one gets that $w \geqslant 0$ on $[t', t_1]$, which is in contradiction with the fact that $w(t') < 0$. Hence we have (3.41). Using (3.2), (3.17), (3.18), (3.19), (3.21), and (3.22), one easily sees that $y = y_\omega$ satisfies the assumptions of Lemma 3.3 if $M \geqslant M_0$. From now on we assume that $M \geqslant M_0$ with M_0 as in Lemma 3.3. From the continuity of Φ^{y_ω}, the definition (3.23) of s_{y_ω} and Lemma 3.3, one easily sees that (3.37) holds. Hence we have (3.31). The continuity of F can be proved with the same type of arguments used to prove the continuity of $F(\omega)$. We omit the proof.

Let us now turn to the proof of (3.33). By Wolibner–Yudovich's theorem (Theorem 3.1), (3.22), (3.24), (3.25), and Ascoli's theorem, it suffices to check that

$$s_{y_\omega} \text{ is relatively compact in } C^0([0,T] \times \overline{\Omega}). \tag{3.43}$$

Let $(\omega_k;\, k \in \mathbb{N})$ be a sequence of functions in X. We want to prove the existence of a subsequence of the $(\omega_k;\, k \in \mathbb{N})$ converging in $C^0([0,T] \times \overline{\Omega})$. By Ascoli's theorem and Wolibner–Yudovich's theorem, (Theorem 3.1), the set $\{\Phi^{y_{\omega_k}};\, k \in \mathbb{N}\}$ is relatively compact in $C^0([0,T] \times [0,T] \times \overline{B_R}; \overline{B_R})$. Hence, without loss of generality, we may assume the existence of $\Phi \in C^0([0,T] \times [0,T] \times \overline{B_R}; \overline{B_R})$ such that the sequence $(\Phi^{y_{\omega_k}};\, k \in \mathbb{N})$ is converging to Φ in $C^0([0,T] \times [0,T] \times \overline{B_R}; \overline{B_R})$. Associated with Φ is the function $s : [0,T] \times \overline{\Omega} :\to \mathbb{R}$ defined by (see (3.23))

$$s(t,x) = \text{ Max } \{t' \in [0,t];\, \Phi(t',t,x) \in \overline{\gamma_-}\}.$$

(Let us recall the convention Max $\emptyset$ =0.) Let $(t_k;\, k \in \mathbb{N})$ be a sequence of real numbers in $[0,T]$ converging to some $\bar{t}$ as k goes to $+\infty$. Let $(x_k;\, k \in \mathbb{N})$ be a sequence of points in $\overline{\Omega}$ converging to some $\bar{x}$ as k goes to $+\infty$. It is clear that

$$s(\bar{t},\bar{x}) \geqslant \limsup_{k\to+\infty} s_{y_{\omega_k}}(t_k, x_k). \tag{3.44}$$

Therefore in order to prove (3.43) it suffices to check that

$$s(\bar{t},\bar{x}) \leqslant \liminf_{k\to+\infty} s_{y_{\omega_k}}(t_k, x_k). \tag{3.45}$$

Indeed, from (3.44) and (3.45), one gets that the sequence $(s_{y_{\omega_k}};\, k \in \mathbb{N})$ converges uniformly to s on $[0,T]\times\overline{\Omega}$ as $k \to +\infty$. Let us again point out that Φ has the following property: for any $x \in \overline{\Omega}$, if for $0 \leqslant t' \leqslant t \leqslant T$ $\Phi(t',t,x)$ is not in $\overline{\Omega}$, then there exists $t'' \in [t', t]$ such that $\Phi(t'',t,x) \in \overline{\gamma_-}$. Indeed, it follows from the fact that the $\Phi^{y_{\omega_k}}$ have this property (see above) and converge to Φ in $C^0([0,T] \times [0,T] \times \overline{B_R}; \overline{B_R})$ as k goes to $+\infty$. From this property and the convergence of the $\Phi^{y_{\omega_k}}$ to Φ in $C^0([0,T] \times [0,T] \times \overline{B_R}; \overline{B_R})$ one easily sees, as above, that (3.45) holds if Φ satisfies the following property:

$$\forall 0 < \nu < t \leqslant T\, \forall x \in \overline{\gamma_-},\ \Phi([t-\nu, t], t, x) \not\subset \overline{\Omega}. \tag{3.46}$$

Let us prove this property. Let $\theta \in C^\infty(\overline{\Omega})$ be defined by

$$\Delta\theta = 0 \text{ in } \overline{\Omega}, \tag{3.47}$$

$$\frac{\partial\theta}{\partial n} = g \text{ on } \partial\Omega. \tag{3.48}$$

Let us point out that the existence of θ follows from (2.6). Let us write

$$\tilde{y}_{\omega_k} = \alpha_k(t)\nabla\theta + \tilde{z}_k,$$

where $\alpha_k \in C^0([0,T])$ and $\tilde{z}_k \in C^0([0,T] \times \overline{\Omega}; \mathbb{R}^2)$ are defined by

$$\alpha_k(t) = M \text{ Max } \{|\omega_k(t)|_0, |\omega_0(x^0)|\} \, \forall t \in [0,T], \tag{3.49}$$

$$\text{div } \tilde{z}_k = 0 \text{ in } (0,T) \times \Omega, \tag{3.50}$$

$$\text{curl } \tilde{z}_k = \omega_k \text{ in } (0,T) \times \Omega, \tag{3.51}$$

$$\tilde{z}_k \cdot n = 0 \text{ on } [0,T] \times \partial\Omega. \tag{3.52}$$

Using (3.14) and (3.15) for $\omega = \omega_k$ and using (3.49), one gets

$$M|\omega_0(x^0)| \leqslant \alpha_k(t) \leqslant M|\omega_0|_0, \tag{3.53}$$

$$t \in [0,T] \to \alpha_k(t) \text{ is nonincreasing.} \tag{3.54}$$

From (3.53) and (3.54), we get the existence of $\alpha \in L^\infty(0,T)$ such that, extracting subsequences if necessary, one has

$$0 < M|\omega_0(x^0)| \leqslant \alpha(t) \text{ for a.e. } t \in (0,T), \tag{3.55}$$

$$\alpha_k(t) \to \alpha(t) \text{ as } k \to +\infty \text{ for a.e. } t \in (0,T). \tag{3.56}$$

Let us now turn to the sequence $(\tilde{z}_k; k \in \mathbb{N})$. Let us fix $r \in (2, +\infty)$. Using (3.14), (3.15), and (3.16) with $\omega = \omega_k$, using also (3.50), (3.51), and (3.52), one gets that

$$\tilde{z}_k \text{ is bounded in } C^0([0,T]; W^{1,r}(\Omega; \mathbb{R}^2)), \tag{3.57}$$

$$\frac{\partial\tilde{z}_k}{\partial t} \text{ is bounded in } L^\infty([0,T]; H^{-1}(\Omega; \mathbb{R}^2)). \tag{3.58}$$

Then, by a compactness lemma due to P.-L. Lions [18, Lemma C1, Appendix C] (take, with the notations of [18], $X = W^{1,r}(\Omega; \mathbb{R}^2)$ and $Y = H^{-1}(\Omega; \mathbb{R}^2)$) and by the Rellich–Kondrakov theorem, (3.57) and (3.58) imply that the sequence $(\tilde{z}_k; k \in \mathbb{N})$ is relatively compact in $C^0([0,T] \times \overline{\Omega}; \mathbb{R}^2)$. Hence, extracting subsequences if necessary, we may assume the existence of $\tilde{z} \in C^0([0,T] \times \overline{\Omega}; \mathbb{R}^2)$ such that $\tilde{z}_k$ tends to $\tilde{z}$ in $C^0([0,T] \times \overline{\Omega}; \mathbb{R}^2)$ as k tends to $+\infty$. Let $y : (0,T) \times B_R \to \mathbb{R}^2$ be defined by

$$y(t,\cdot) = \alpha(t)\mathcal{P}(\nabla\theta) + \mathcal{P}(\tilde{z}(t,\cdot)) \, \forall t \in (0,T). \tag{3.59}$$

Let us check that, if $M \geqslant M_0$, y satisfies the assumptions of Lemma 3.3. By (3.3) and (3.59), one has (3.5). It follows easily from (3.22) that

$$q_{\overline{B_R}}(y_\omega(t,\cdot) \leqslant C_0 C_5 \text{ for a.e. } t \in (0,T),$$

which implies (3.5). Passing to the limit in (3.52), one gets $z \cdot n = 0$, which with (3.59) gives (3.38). Finally, if $M \geqslant M_0$, (3.39) follows easily from the fact that

$$|\text{curl } y(t)|_{L^\infty(\Omega)} \leqslant \liminf_{k\to+\infty} |\text{curl } y_k(t)|_{L^\infty(\Omega)}.$$

Hence y satisfies the assumptions of Lemma 3.3, and, by this lemma, we have (3.46) if one has

$$\Phi = \Phi^{y}. \tag{3.60}$$

But, by the definition of $\Phi^{y_{\omega_k}}$,

$$\Phi^{y_{\omega_k}}(t,s,x) = x + \int_s^t y_{\omega_k}(t', \Phi^{y_{\omega_k}}(t',s,x))dt'. \tag{3.61}$$

Letting $k \to +\infty$ in (3.61), we get, by the dominated convergence theorem,

$$\Phi(t,s,x) = x + \int_s^t y(t', \Phi^{y_{\omega_k}}(t',s,x))dt',$$

for any $(t,s,x) \in [0,T] \times [0,T] \times \overline{B_R}$, which implies (3.60). This ends the proof of Proposition 3.2.

Let us now prove the next proposition.

PROPOSITION 3.4. *For any $M \geqslant M_0$, any maximal solution of system Σ defined at time* 0 *is defined on* $[0,+\infty)$ *(at least).*

Let us recall that $\omega : I \to C^0(\overline{\Omega})$ is a *maximal* solution of system Σ if, for any interval J containing I but not equal to I, there exists no solution of system Σ which is equal to ω on the interval I. The existence of a maximal solution follows, as usual, from Zorn's lemma; hence Theorem 2.2 is a corollary of Propositions 3.2 and 3.4.

Until the end of this paper, for any $\omega \in L^\infty(\Omega)$, one defines $\tilde{y}_\omega \in C^0(\overline{\Omega})$ and $y_\omega \in C^0(\overline{B_R})$ by requiring

$$\begin{gathered} \text{div } \tilde{y}_\omega = 0 \text{ and curl } \tilde{y}_\omega = \omega \text{ in } \Omega, \\ \tilde{y}_\omega \cdot n = M|\omega|_{L^\infty(\Omega)} g \text{ on } \partial\Omega, \\ y_\omega = \mathcal{P}(\tilde{y}_\omega). \end{gathered}$$

Of course if, for some interval I, ω is a map from I into $L^\infty(\Omega)$, the above conditions specified at any time in I, give maps $\tilde{y}_\omega : I \to C^0(\overline{\Omega})$ and $y_\omega : I \to C^0(\overline{B_R})$. With these notations, let us start the proof of Proposition 3.4 by the following simple observation.

LEMMA 3.5. *Let $T > 0$ and let $\omega \in C^0([0,T]; C^0(\overline{\Omega}))$ be a solution of system Σ. Then, for any $t \in [0,T]$ and for any $x \in \overline{\Omega}$,*

$$\omega(t,x) = \omega(0, a_{y_\omega}(t,x)) \exp\left(-M \int_0^{s_{y_\omega}(t,x)} |\omega(t')|_0 dt'\right). \tag{3.62}$$

In particular,

$$t \to |\omega(t)|_0 \textit{ is nonincreasing.} \tag{3.63}$$

Indeed, for $t \in (0,T]$ and $x \in \overline{\Omega}$, let $\omega^* : [0,t] \to \mathbb{R}$ be defined by $\omega^*(t') = \omega(t', \Phi^{y_\omega}(t',t,x))$. Using (2.7), (2.8), and standard smoothing procedures, one gets that, in the sense of distributions,

$$\frac{\mathrm{d}\omega^*}{\mathrm{d}t'} = 0 \text{ on } (s_{y_\omega}(t,x), t).$$

In particular,

$$\omega(s_{y_\omega}(t,x), a_{y_\omega}(t,x)) = \omega^*(s_{y_\omega}(t,x)) = \omega^*(t) = \omega(t,x). \tag{3.64}$$

If $s_{y_\omega}(t,x)=0$, this gives (3.62). Let us study the case where $0<s_{y_\omega}(t,x)$. It follows directly from (2.11) that, in the sense of distributions,

$$\frac{\partial\omega}{\partial t'}(t',a)=-M|\omega(t')|_0\omega(t',a) \text{ on } \{t'\in(0,T)\}\,\forall a\in\gamma_-.$$

In particular,

$$\omega^*(s_{y_\omega}(t,x))=\omega(0,a_{y_\omega}(t,x))\exp\left(-M\int_0^{s_{y_\omega}(t,x)}|\omega(t')|_0dt'\right),$$

which, with (3.64), gives again (3.62). Property (3.63) follows from (3.62) (see the proof of (3.35)) or note that, if $0\leqslant t_1\leqslant T$, $t\in[0,T-t_1]\to\omega(t+t_1)$ is a solution of system Σ and apply, for $t=t_2-t_1\in[0,T-t_1]$, (3.62) to this solution.

Let $\omega\in C^0(I\times\overline{\Omega})\cong C^0(I;C^0(\overline{\Omega}))$ be a maximal solution to Σ such that I is an interval containing 0. Let $T=\operatorname{Sup} I$. We want to prove that $T=+\infty$. Let us assume that $T<+\infty$. From Proposition 3.2, it follows that $T>0$ and $T\notin I$. Therefore, in order to get a contradiction with the maximal property of ω, it suffices to check that

$$\omega(t) \text{ converges in } C^0(\overline{\Omega}) \text{ as } t\to T^-. \tag{3.65}$$

Indeed, if (3.65) holds, then $\bar\omega : I\cup\{T\}\to C^0(\overline{\Omega})$ defined by $\bar\omega=\omega$ on I and $\bar\omega(T)=\lim_{t\to T}\omega(t)$ is also a solution to system Σ.

If $|\omega(t)|_0\to 0$ as $t\to T^-$, (3.65) holds. Therefore, by (3.63), we may assume that, for some $\eta>0$,

$$|\omega(t)|_0\geqslant\eta\ \forall t\in[0,T). \tag{3.66}$$

As above, using, in particular, Lemma 3.3 and (3.66), one gets that s_{y_ω} is continuous on $[0,T]\times\overline{\Omega}$. Therefore $F(\omega)$ defined by (3.25) is continuous on $[0,T]\times\overline{\Omega}$. But, by Lemma 3.5, this function is equal to ω on $[0,T)\times\overline{\Omega}$. This proves (3.65) and therefore ends the proof of Theorem 2.2.

3.2. Proof of Theorem 2.4. Let $V : C^0(\overline{\Omega})\to[0,+\infty)$ be defined by

$$V(\omega)=|\omega\exp(-\theta)|_0,$$

where $\theta\in C^\infty(\overline{\Omega})$ satisfies (3.47) and (3.48). Theorem 2.4 is an easy consequence of the following proposition.

PROPOSITION 3.6. *There exists $M_4\geqslant M_0$ and $\mu>0$ such that, for any $M\geqslant M_4$ and any solution $\omega:[0,+\infty)\to C^0(\overline{\Omega})$ of system Σ, one has, for any $t\in[0,+\infty)$,*

$$[-\infty,0]\ni\dot V(t):=\frac{d}{dt^+}V(\omega(t))\leqslant-\mu MV^2(\omega(t)), \tag{3.67}$$

where $d/dt^+V(\omega(t)):=\lim_{\varepsilon\to0^+}(V(\omega(t+\varepsilon))-V(\omega(t)))/\varepsilon$.

Let us check that this proposition indeed implies Theorem 2.4. Let $\omega:[0,+\infty)\to C^0(\overline{\Omega})$ be a solution of system Σ with $\omega(0)\neq0$ and $M\geqslant M_4$. Integrating (3.67), one gets

$$V(\omega(t))\leqslant\frac{V(\omega(0))}{1+\mu MtV(\omega(0))}\,\forall t\geqslant0.$$

In particular,

$$V(\omega(t))\leqslant\frac{1}{\mu Mt}\,\forall t>0. \tag{3.68}$$

But

$$|\omega(t)|_0 \leqslant |\exp(\theta)|_0 V(\omega(t))\, \forall t \geqslant 0,$$

which, with (3.68), gives

$$|\omega(t)|_0 \leqslant \frac{|\exp(\theta)|_0}{\mu M t}\, \forall t > 0.$$

Since, by (3.63), $|\omega(t)|_0 \leqslant |\omega(0)|_0$ for any $t \geqslant 0$, we get (2.12) if one takes $M \geqslant |\exp(\theta)|_0/(\mu\varepsilon)$. Therefore Theorem 2.4 holds with

$$M_1 = \text{ Max } \{M_4, |\exp(\theta)|_0/\mu\}.$$

Let us now turn to the proof of Proposition 3.6. Clearly, since system Σ is autonomous, it suffices to check that (3.67) holds for $t = 0$. Let us also assume that the following lemma, which is proved in Appendix A, holds.

LEMMA 3.7. *For any x in $\overline{\Omega}$,*

$$\nabla\theta(x) \neq 0. \tag{3.69}$$

From this lemma and standard elliptic estimates, we get the existence of $\mu_0 \in (0, 1]$ such that

$$\nabla\theta \cdot (\nabla\theta + z) \geqslant \mu_0 \text{ on } \overline{\Omega} \tag{3.70}$$

for any $z \in C^0(\overline{\Omega}; \mathbb{R}^2)$ such that $\operatorname{div} z = 0$, $|\operatorname{curl} z|_0 \leqslant \mu_0$, and $z \cdot n = 0$ on $\partial\Omega$. For $t \geqslant 0$, let $x(t) \in \overline{\Omega}$ be such that

$$V(\omega(t)) = |\omega(t, x(t))| \exp(-\theta(x(t))).$$

For simplicity, we assume that $\omega(t, x(t)) \geqslant 0$; the case where $\omega(t, x(t)) < 0$ can be treated in a similar way. We have

$$V(\omega(t)) - V(\omega(0)) \leqslant \kappa(t), \tag{3.71}$$

with

(3.72)
$$\begin{aligned}\kappa(t) &= \omega(t, x(t)) \exp(-\theta(x(t))) - \omega(0, a_{y_\omega}(t, x(t))) \exp(-\theta(a_{y_\omega}(t, x(t))))\\ &= V(\omega(t)) \left(1 - \exp\left(\theta(x(t)) - \theta(a_{y_\omega}(t, x(t))) + M \int_0^{s_{y_\omega}(t,x(t))} |\omega(t')|_0 dt'\right)\right).\end{aligned}$$

We choose $M_4 = \text{ Max } \{1/\mu_0, M_0\}$ and take any $M \geqslant M_4$. Let us again decompose y_ω in the following way:

$$y_\omega = M|\omega(t)|_0(\nabla\theta + z), \tag{3.73}$$

with $\operatorname{div} z = 0$, $\operatorname{curl} z = \omega/(M|\omega(t)|_0)$ ($= 0$ if $\omega(t) = 0$), and $z \cdot n = 0$ on $\partial\Omega$. For any x in $\overline{\Omega}$ and for any s in $[s_{y_\omega}(t, x), t]$, one has

$$\frac{\partial}{\partial s}\left(\theta\left(\Phi^{y_\omega}(s, t, x)\right)\right) = \nabla\theta(\Phi^{y_\omega}(s, t, x)) y_\omega\left(s, \Phi^{y_\omega}(s, t, x)\right),$$

which, with (3.70), (3.73), gives

$$\frac{\partial}{\partial s}\left(\theta\left(\Phi^{y_\omega}(s, t, x)\right)\right) \geqslant \mu_0 M |\omega(t)|_0.$$

In particular,

$$\theta(x(t)) - \theta\left(a_{y_\omega}(t, x(t))\right) \geqslant \mu_0 M|\omega(t)|_0(t - s_{y_\omega}(t, x(t))),$$

which, with (3.63) and (3.72), implies that

$$\begin{aligned}\kappa(t) &\leqslant MV(\omega(t))|\omega(t)|_0\left(-\mu_0\left(t - s_{y_\omega}\left(t, x\left(t\right)\right)\right) - s_{y_\omega}\left(t, x\left(t\right)\right)\right)\\ &\leqslant -\mu_0 MV(\omega(t))|\omega(t)|_0 t\\ &\leqslant -\mu_0 M \text{ Min } \{\exp(\theta(x));\ x \in \overline{\Omega}\} V(\omega(t))^2 t\end{aligned}$$

Hence, one gets Proposition 3.6 by taking $\mu = \mu_0$ Min $\{\exp(\theta(x));\ x \in \overline{\Omega}\}$.

4. Proof of Theorem 2.8. Let us first prove (i) of Theorem 2.8. We are going to deduce (i) of Theorem 2.8 from (the proof of) Theorem 2.2. Roughly speaking the idea is that, if we replace (2.11) with

$$\frac{\partial \omega}{\partial t} = -kM\left|\omega(t)\right|_0 \omega \text{ on } \{t;\ \omega(t) \neq 0\} \times \gamma_-,$$

then, as $k \to +\infty$, the solutions of the Cauchy problem for system Σ converge to a solution of the Cauchy problem for system Σ_1.

Reversing time in the proof of Lemma 3.3, one easily gets the following.

LEMMA 4.1. *There exists $M_5 \geqslant M_0$ such that, for any $T > 0$ and for any $y \in L^\infty\left((0,T); C^0\left(\overline{B_R}; \mathbb{R}^2\right)\right)$ satisfying* (3.5), (3.6) *for some $K > 0$, and for some function $\alpha \in L^\infty\left((0,T); (0, +\infty)\right)$,*

$$y(t,\cdot)\cdot n = \alpha(t) g \textit{ on } \partial\Omega \textit{ for a.e. } t \in (0,T), \tag{4.1}$$

$$M_5|\text{ curl } y(t)|_{L^\infty(\Omega)} \leqslant \alpha(t) \textit{ for a.e. } t \in (0,T), \tag{4.2}$$

$$\text{div y} = 0 \textit{ in } (0,T)\times\Omega \tag{4.3}$$

for any $(\tilde{t}, \tilde{x}) \in [0,T) \times \overline{\gamma_+}$ and for any $\nu \in (0, T - \tilde{t})$, there exists $t \in (\tilde{t}, \tilde{t} + \nu)$, such that

$$\Phi^y\left(t, \tilde{t}, \tilde{x}\right) \notin \overline{\Omega}.$$

We choose $M_2 =$ Max $\{M_0, M_5\}$, where M_0 is defined in the proof of Lemma 3.3; see (B.3). Let $M \geqslant M_2$. Let $\omega_0 \in L^\infty(\Omega)$. There exists a sequence $(\omega_{0,k};\ k \in \mathbb{N}^*)$ of functions in $C^0(\overline{\Omega})$ such that

$$|\omega_{0,k}(x)| \leqslant |\omega_0|_{\infty,\Omega\cap B(x,1/k)}\ \forall k \in \mathbb{N}^*\ \forall x \in \overline{\Omega}, \tag{4.4}$$

$$\omega_{0,k}(x) \underset{k\to+\infty}{\longrightarrow} \omega_0(x) \text{ for a.e. } x \in \Omega. \tag{4.5}$$

By (the proof of) Theorem 2.2, there exists a solution $\omega^k \in C^0([0,\infty); C^0(\overline{\Omega}))$ of system Σ, with (2.11) replaced by

$$\frac{\partial \omega_k}{\partial t} = -kM\left|\omega_k(t)\right|_0 \omega_k \text{ on } \{t;\ \omega_k(t) \neq 0\} \times \gamma_-, \tag{4.6}$$

such that $\omega_k(0) = \omega_{0,k}$. Let $\tilde{y}_k = \tilde{y}_{\omega_k} \in C^0([0,+\infty) \times \overline{\Omega}; \mathbb{R}^2)$ and let $y_k = y_{\omega_k} \in C^0([0,+\infty) \times \overline{B_R}; \mathbb{R}^2)$. Let $\alpha_k \in C^0([0,+\infty))$ and $\tilde{z}_k \in C^0([0,+\infty) \times \overline{\Omega}; \mathbb{R}^2)$ be defined by

$$\alpha_k(t) = M|\omega_k(t)|_0\ \forall t \in [0,+\infty), \tag{4.7}$$

$$\text{div } \tilde{z}_k = 0, \quad \text{curl } \tilde{z}_k = \omega_k \text{ in } (0,+\infty)\times\Omega,$$

$$\tilde{z}_k \cdot n = 0 \text{ on } [0,+\infty) \times \partial\Omega.$$

One has $\tilde{y}_k = \alpha_k \nabla\theta + \tilde{z}_k$. As in the proof of (3.43), extracting subsequences if necessary, one gets the existence of ω in $L^\infty((0,+\infty); L^\infty(\Omega)) \cap C^0([0,+\infty); H^{-1}(\Omega))$, of a nonincreasing function α in $L^\infty((0,+\infty),[0,+\infty))$ and of $\tilde{z}$ in $C^0([0,+\infty)\times\overline{\Omega})$ with $t \to q(\tilde{z}(t,\cdot)) \in L^\infty_{\mathrm{loc}}(0,+\infty)$ such that, for any $T \in [0,+\infty)$, one has, as $k \to +\infty$

$$\omega_k \to \omega \text{ in } C^0([0,T]; H^{-1}(\Omega)), \tag{4.8}$$

$$\omega_k \rightharpoonup \omega \text{ in } \sigma(L^1((0,+\infty)\times\Omega), L^\infty((0,+\infty)\times\Omega)), \tag{4.9}$$

$$\tilde{z}_k \to \tilde{z} \text{ in } C^0([0,T]\times\overline{\Omega};\mathbb{R}^2), \tag{4.10}$$

$$\alpha_k(t) \to \alpha(t) \text{ for a.e. } t \in (0,+\infty). \tag{4.11}$$

Let $\tilde{y}(t) = \alpha(t)\nabla\theta + \tilde{z}$. One has (2.8) and (2.9) for $y = \tilde{y}$ and $I = [0,T]$. Let $y \in L^\infty([0,+\infty); C^0(\overline{B_R};\mathbb{R}^2))$ be defined by $y(t,\cdot) = \mathcal{P}(\tilde{y}(t,\cdot))$. Let $\varphi \in C^1((0,+\infty)\times\overline{\Omega})$ with compact support such that

$$\text{Support } \varphi \subset ((0,+\infty)\times\Omega) \cup (\{t \in (0,+\infty)\,;\, \alpha(t) > 0\} \times \gamma_-)\,. \tag{4.12}$$

Let us check that (2.19) holds. Since α_k, $k \in \mathbb{N}^*$, and α are nonincreasing, one gets from (4.11) and (4.12) that, for some $k_0 \in \mathbb{N}^*$, one has, for any $k \geqslant k_0$,

$$\text{Support } \varphi \subset ((0,+\infty)\times\Omega) \times (\{t \in (0,+\infty);\, |\omega_k(t)|_0 > 0\} \times \gamma_-).$$

Hence, for any $k \geqslant k_0$,

$$\int_{(0,+\infty)\times\Omega} \left(\omega_k \frac{\partial\varphi}{\partial t} + \omega(\tilde{y}_k\cdot\nabla)\varphi\right) = \int_{(0,+\infty)\times\Omega} \alpha_k\omega_k g\varphi. \tag{4.13}$$

Let $t_0 > 0$ be such that

$$\alpha(t_0) > 0, \tag{4.14}$$

$$\text{Support } \varphi \subset [0,t_0] \times (\Omega \cup \gamma_-). \tag{4.15}$$

Since α_k, $k \in \mathbb{N}^*$, and α are nonincreasing, it follows again from (4.12) that there exists a positive integer $k_1 \geqslant k_0$ such that, for any $k \geqslant k_1$ and any $t \in [0,t_0]$, $\alpha_k(t) \geqslant \alpha(t_0)/2$. Hence, by (4.4) and (4.6), one has, for any $(t,x) \in [0,t_0]\times\gamma_-$,

$$|\omega_k(t,x)| \leqslant |\omega|_\infty \exp\left(-\frac{k\alpha(t_0)t}{2}\right). \tag{4.16}$$

Hence, letting k go to $+\infty$ in (4.13), and using (4.14), (4.15), and (4.16), one gets (2.19).

It remains to check that

$$M|\omega(t)|_\infty = \alpha(t) \text{ for a.e. } t \in (0,+\infty). \tag{4.17}$$

By (4.7), (4.8), and (4.11),

$$M|\omega(t)|_\infty \leqslant \alpha(t) \text{ for a.e. } t \in (0,+\infty).$$

Hence, in order to prove (4.17), it suffices to check that, if for some $0 < t_1 < t_2$, for some $\chi > 0$, and for some $k_2 > 0$,

$$\alpha_k(t_2) \geqslant \chi\,\forall k \geqslant k_2, \tag{4.18}$$

then

$$M|\omega(t_1)|_\infty \geqslant \chi. \tag{4.19}$$

Let $x_k \in \overline{\Omega}$ be such that

$$M|\omega_k(t_2, x_k)| = \alpha_k(t_2). \tag{4.20}$$

Still extracting subsequences if necessary, we may assume that, for some $x_\infty \in \overline{\Omega}$,

$$x_k \to x_\infty \text{ as } k \to \infty. \tag{4.21}$$

Let $\bar{x}_\infty = \Phi^y(0, t_2, x_\infty)$. Let us assume for the moment that the following lemma holds.

Lemma 4.2. *There exists $\rho > 0$ and $k_3 \geqslant k_2$ such that, for any $k \geqslant k_3$ and for any $t \in [0, t_1]$,*

$$\Phi^{y_k}(t, 0, \bar{x}) \in \overline{\Omega}\, \forall \bar{x} \in \overline{\Omega} \cap B(\bar{x}_\infty, \rho). \tag{4.22}$$

Letting $k \to +\infty$ in (4.22), one gets

$$\Phi^y(t_1, 0, \overline{\Omega} \cap B(\bar{x}_\infty, \rho)) \subset \overline{\Omega}. \tag{4.23}$$

From Lemma 3.3 and (4.22), it follows that, for any $\bar{x} \in \overline{\Omega} \cap B(\bar{x}_\infty, \rho)$ and any $k \geqslant k_3$,

$$s_{\omega_k}(t_1, \Phi^{y_k}(t_1, 0, \bar{x})) = 0,$$

so that, by (3.62),

$$\omega_k(t_1, \Phi^{y_k}(t_1, 0, \bar{x})) = \omega_0(\bar{x}). \tag{4.24}$$

Let us point out that (4.24) implies that

$$\omega(t_1, x) = \omega_0(\Phi^y(0, t_1, x)) \text{ for a.e. } x \in \Phi^y(t_1, 0, \overline{\Omega} \cap B(\bar{x}_\infty, \rho)). \tag{4.25}$$

Indeed, let us first notice that, since $\Phi^y(t, 0, \cdot)$ is a homeomorphism of $\overline{B_R}$—its inverse is $\Phi^y(0, t, \cdot)$—it follows from (4.23) and the invariance of domain theorem (see, e.g., [19, Theorem 3.3.2]) that

$$\Phi^y(t, 0, \Omega \cap B(\bar{x}_\infty, \rho)) \subset \Omega\, \forall t \in [0, t_1]. \tag{4.26}$$

Let $\psi \in C^\infty(\overline{\Omega})$, the support of which is included in the open subset $\Phi^y(t_1, 0, \Omega \cap B(\bar{x}_\infty, \rho)) \subset \Omega$. Then, by (4.8),

$$< \omega(t_1, x), \psi >_{H^{-1}(\Omega), H_0^1(\Omega)} = \lim_{k \to +\infty} I_k, \tag{4.27}$$

with

$$I_k = \int_{\Phi^y(t_1, 0, \Omega \cap B(\bar{x}_\infty, \rho))} \omega_k(t_1, x)\psi(x) dx.$$

Since div $\tilde{y}_k = 0$, one gets, using also (4.22) and (4.24), the existence of $k_4 \geqslant k_3$ such that, for any $k \geqslant k_4$,

$$\begin{aligned} I_k &= \int_{\Phi^{y_k}(t, 0, \Omega \cap B(\bar{x}_\infty, \rho))} \omega_k(t_1, x)\psi(x) dx \\ &= \int_{\Omega \cap B(\bar{x}_\infty, \rho)} \omega_k(t_1, \Phi^{y_k}(t_1, 0, \bar{x})\psi(\Phi^{y_k}(t_1, 0, \bar{x})) d\bar{x} \\ &= \int_{\Omega \cap B(\bar{x}_\infty, \rho)} \omega_0(\bar{x})\psi(\Phi^{y_k}(t_1, 0, \bar{x})) d\bar{x}. \end{aligned}$$

Note that, by (4.26) and the fact that div $y = 0$ in $(0, T) \times \Omega$, $\Phi^y(t_1, 0, \cdot)$ restricted to $\Omega \cap B(\bar{x}_\infty, \rho)$ preserves the Lebesgue measure. Hence

$$\begin{aligned}\lim_{k\to+\infty} I_k &= \int_{\Omega\cap B(\bar{x}_\infty,\rho)} \omega_0(\bar{x})\psi(\Phi^y(t_1, 0, \bar{x}))d\bar{x} \\ &= \int_{\Phi^y(t_1,0,\Omega\cap B(\bar{x}_\infty,\rho))} \omega_0(\Phi^y(0, t_1, x))\psi(x)dx,\end{aligned}$$

which, with (4.27), implies (4.25). From Lemma 3.5, with kM instead of M in (3.62), (4.18), and (4.20), one has

$$\begin{aligned}(4.28)\quad \chi \leqslant M|\omega_k(t_2, x_k)| \leqslant M|\omega_{0,k}(\Phi^{y_k}(s_{y_k}(t_2, x_k), t_2, x_k)| \\ \times (\exp(-kM\chi s_{y_k}(t_2, x_k)))\,.\end{aligned}$$

In particular, by (4.4),

$$s_{y_k}(t_2, x_k) \to 0 \text{ as } k \to +\infty,$$

which, with (4.4), (4.18), (4.20), and (4.28), implies that

$$(4.29)\qquad M|\omega_0|_{\infty,\Omega\cap B(\bar{x}_\infty,\rho)} \geqslant \chi.$$

This inequality, together with (4.25) and the fact that $\Phi^y(t_1, 0, \cdot)$ restricted to $\Omega \cap B(\bar{x}_\infty, \rho)$ preserves the Lebesgue measure, implies that

$$(4.30)\qquad M|\omega(t_1)|_{\infty,\Phi^y(t_1,0,\Omega\cap B(\bar{x}_\infty,\rho))} \geqslant \chi,$$

which gives (4.19).

It remains to prove Lemma 4.2. Let us first point out that

$$(4.31)\qquad \Phi^y(t, t_2, x_\infty) \in \overline{\Omega}\, \forall t \in [0, t_2].$$

Indeed, if (4.31) does not hold, there exists $t_3 \in (0, t_2]$ such that $\Phi^y(t_3, t_2, x_\infty) \notin \overline{\Omega}$. Then, for some $k_5 > 0$,

$$s_{y_k}(t_2, x_k) \geqslant t_3/2\, \forall k \geqslant k_5,$$

which, with (4.4) and (4.6), implies that

$$(4.32)\qquad \chi \leqslant M|\omega_k(t_2, x_k)| \leqslant M|\omega_0|_\infty \exp(-kM\chi t_3/2)\,.$$

Letting k go to $+\infty$ in (4.32), we get a contradiction. This ends the proof of (4.31).

One easily checks that y satisfies the assumption of Lemma 4.1. Since $\Phi^y(t, t_2, x_\infty) = \Phi(t, 0, \bar{x}_\infty)$, Lemma 4.1 and (4.31) give that

$$\Phi^y(t, 0, \bar{x}_\infty) \in \overline{\Omega} \setminus \overline{\gamma_+}\, \forall t \in [0, t_1],$$

which implies Lemma 4.2.

Let us now prove (ii) of Theorem 2.8. One needs the following lemma, which is analogous to Lemma 3.5 but requires a different proof.

LEMMA 4.3. *Let $T > 0$, let $\omega \in C^0([0, T]; H^{-1}(\Omega)) \cap L^\infty([0, T]; L^\infty(\Omega))$ be a solution of system Σ_1, and let $t \in [0, T]$. Then the closed set $S(t) := \{x \in \omega;\ a_{y_\omega}(t, x) \in \partial\Omega$ and $s_{y_\omega}(t, x) = 0\}$ has measure 0 and, for a.e. $x \in \Omega$,*

$$(4.33)\qquad \omega(t, x) = \omega(0, \Phi^{y_\omega}(0, t, x)) \textit{ if } s_{y_\omega}(t, x) = 0,$$

$$(4.34)\qquad \omega(t, x) = 0 \textit{ if } s_{y_\omega}(t, x) > 0,$$

where s_{y_ω} *is defined by* (3.23)*—recall the convention* Max $\emptyset = 0$. *In particular,* (3.62) *again holds (for a.e.* $(t,x) \in (0,T) \times \Omega$*).*

With this lemma, which is proved at the end of this section, one gets easily (ii) of Theorem 2.8. Indeed, let ω be a maximal solution of system Σ_1 defined on an interval I containing 0 and let us assume that $T :=$ Sup $I < +\infty$. Again, it follows from (i) of Theorem 2.8 that

$$T \notin I. \tag{4.35}$$

Using (3.62), which holds by Lemma 4.3, one gets that $|\omega(t)|_\infty \leqslant |\omega(0)|_\infty$ for any t in $[0,T)$. Then, using the fact that $\partial\omega/\partial t = -$ div (ωy), one gets that $\omega \in H^1((0,T); H^{-1}(\Omega))$. Hence

$$\omega(t) \text{ converges in } H^{-1}(\Omega) \text{ as } t \to T^-. \tag{4.36}$$

Then $\bar\omega : I \cup \{T\} \to C^0(\overline{\Omega})$ defined by $\bar\omega = \omega$ on I and $\bar\omega(T) = \lim_{t\to T} \omega(t)$ is also a solution of system Σ_1, a contradiction with the maximal property of ω and (4.35).

Finally, we briefly sketch the proof of Lemma 4.3. Let us first check that $S(t)$ has measure 0. One has

$$S(t) \subset \{\Phi^{y_\omega}(t,0,\bar x);\ \bar x \in \gamma_-,\ \Phi^{y_\omega}(s,0,\bar x) \in \Omega\ \forall s \in (0,T]\}.$$

Hence, since div $y_\omega = 0$ in $(0,T) \times \Omega$, $S(t)$ has measure 0.

Let us now prove (4.33). Let $U = \{(t,x) \in (0,T) \times \Omega;\ a_{y_\omega}(t,x) \in \Omega\}$. Let us consider the *linear* hyperbolic equation that we call $\mathcal{L}$, where $f : U \to \mathbb{R}$ is the unknown,

$$\frac{\partial f}{\partial t} + \text{ div } (f y_\omega) = 0 \text{ in } U, \tag{4.37}$$

$$f(0,\cdot) = \omega(0,\cdot) \text{ on } \{0\} \times \Omega, \tag{4.38}$$

where (4.37) and (4.38) have to be understood in a weak sense. More precisely a function $f : U \to \mathbb{R}$ is a solution of $\mathcal{L}$ if $f \in L^\infty(U)$ and is such that, for any $\psi \in C^1_0(U \cup (\{0\} \times \Omega))$,

$$\int_U f\left(\frac{\partial\psi}{\partial t} + (y_\omega \cdot \nabla)\,\psi\right) = -\int_\Omega \omega(0,\cdot)\psi(0,\cdot).$$

Clearly ω is a solution of $\mathcal{L}$. Moreover $\bar\omega : U \to \mathbb{R}$, defined by

$$\bar\omega(t,x) = \omega(0, \Phi^{y_\omega}(0,t,x))\,\forall (t,x) \in U,$$

is also a solution of $\mathcal{L}$. Hence, in order to prove (4.33), it suffices to check that $\mathcal{L}$ has a unique solution. When y_ω is of class C^1, this is a classical result due to Oleĭnik [21]; see also [11, Theorem 2.2.1]. When $y_\omega \in L^\infty\left([0,T]; C^0\left(\overline{B_R}; \mathbb{R}^2\right)\right)$—satisfying (3.5)—and such that $t \to q_{\overline{B_R}}(y(t,\cdot)) \in L^\infty(0,T)$ one needs a (very slight) modification of the proof since, for example, (2.2.10) in [11] is no longer true. Let us briefly describe the modification. By standard smoothing procedures, one can construct, for $\varepsilon \in (0,1]$, $y^\varepsilon \in C^1([0,T] \times \overline{B_R}; \overline{B_R})$ be such that, for some $C_7 > 0$,

$$|y^\varepsilon - y_\omega|_{0,[0,T]\times\overline{B_R}} \leqslant C_7 \varepsilon^{3/4}\ \forall \varepsilon \in (0,1], \tag{4.39}$$

$$\text{Support } y^\varepsilon \subset [0,T] \times \overline{B_R/2}\ \forall \varepsilon \in (0,1], \tag{4.40}$$

$$|\nabla y^\varepsilon|_{0,[0,T]\times\overline{B_R}} \leqslant C_7 \varepsilon^{-1/4}\ \forall \varepsilon \in (0,1]. \tag{4.41}$$

Let $\varphi \in C^1([0,T] \times \overline{B_R})$, the support of which is included in U. Let, for $\varepsilon \in (0,1]$, $\psi^\varepsilon \in C^1([0,T] \times \overline{B_R})$ be defined by

$$\psi^\varepsilon(t,x) = -\int_t^T \varphi(s, \Phi^{y^\varepsilon}(s,t,x))ds.$$

One has

$$\psi^\varepsilon(T,\cdot) = 0,$$
$$\frac{\partial \psi^\varepsilon}{\partial t} + (y^\varepsilon \cdot \nabla)\psi^\varepsilon = \varphi,$$

and one easily checks that, at least for ε small enough, Support $\psi^\varepsilon \subset U \cup (\{0\} \times \Omega)$. Therefore, at least for ε small enough,

$$\int_U (\omega - \bar{\omega})\varphi = \int_U (\omega - \bar{\omega})((y^\varepsilon - y_\omega) \cdot \nabla)\psi^\varepsilon. \tag{4.42}$$

But, with (4.41), one gets the existence of C_8 such that

$$|\nabla \Phi^{y^\varepsilon}|_{0,[0,T]\times\overline{B_R}\times\overline{B_R}} \leqslant C_8 \varepsilon^{-1/4} \ \forall \varepsilon \in (0,1].$$

This gives the existence of C_9 such that

$$|\nabla \psi^\varepsilon|_{0,[0,T]\times\overline{B_R}} \leqslant C_9 \varepsilon^{-1/4} \ \forall \varepsilon \in (0,1],$$

which, with (4.39), implies that

$$\lim_{\varepsilon \to 0} |((y^\varepsilon - y_\omega) \cdot \nabla)\psi^\varepsilon|_{0,[0,T]\times\overline{B_R}} = 0. \tag{4.43}$$

From (4.42) and (4.43), one gets

$$\int_U (\omega - \bar{\omega})\varphi = 0,$$

and therefore $\omega = \bar{\omega}$ on U. The proof of (4.34) is similar to the proof of (4.33): consider $U' = \{(t,x) \in (0,T) \times \Omega;\ s_{y_\omega}(t,x) > 0\}$ and the linear hyperbolic equation $\mathcal{L}'$ $(\partial f / \partial t) +$ div $(f y_\omega) = 0$ on U' with, instead of (4.38), the boundary condition $f = 0$ on $(0,T) \times \gamma_-$; as above, one shows that this equation $\mathcal{L}'$ has a unique solution; but ω and 0 are solutions of $\mathcal{L}'$ on U', which proves (4.34). We omit the details.

Appendix A. Proof of Lemma 3.7. Since θ is harmonic on $\overline{\Omega}$, which is simply connected, it admits a harmonic conjugate $\psi \in C^\infty(\overline{\Omega})$. One has, with $x = (x_1, x_2)$,

$$\frac{\partial \psi}{\partial x_1} = \frac{\partial \theta}{\partial x_2},\ \frac{\partial \psi}{\partial x_2} = -\frac{\partial \theta}{\partial x_1}. \tag{A.1}$$

Let $\tau \in C^\infty(\partial\Omega; \mathbb{R}^2)$ be the unit tangent vector field on $\partial\Omega$ such that (τ, n) is a direct basis of $\mathbb{R}^2$ at any point of $\partial\Omega$. From (A.1) one gets

$$\frac{\partial \psi}{\partial \tau} = \frac{\partial \theta}{\partial n} \text{ on } \partial\Omega, \tag{A.2}$$

$$\frac{\partial \psi}{\partial n} = -\frac{\partial \theta}{\partial \tau} \text{ on } \partial\Omega. \tag{A.3}$$

By (2.3), the closed set $\partial\Omega \setminus (\gamma_+ \cup \gamma_-)$ has two connected components, that we call Γ_+ and Γ_-. By (3.48) and (A.2), there are two constants C_+ and C_- such that

$$\psi = C_+ \text{ on } \Gamma_+,\ \psi = C_- \text{ on } \Gamma_-. \tag{A.4}$$

Relabeling, if necessary, Γ_+ and Γ_-, we may assume that

$$C_- \leqslant C_+. \tag{A.5}$$

By (3.48) and (A.2),

$$\frac{\partial \psi}{\partial \tau} < 0 \text{ on } \gamma_-, \ \frac{\partial \psi}{\partial \tau} > 0 \text{ on } \gamma_+,$$

which, with (A.4) and (A.5), implies that

$$C_- < C_+, \tag{A.6}$$

$$\psi(x) \in [C_-, C_+] \ \forall x \in \partial\Omega. \tag{A.7}$$

Using (A.4), (A.6), and (A.7), together with the strong maximum principle applied to the harmonic function ψ, one gets that

$$\frac{\partial \psi}{\partial n} < 0 \text{ on } \Gamma_-, \ \frac{\partial \psi}{\partial n} > 0 \text{ on } \Gamma_+, \tag{A.8}$$

which, with (A.3), implies that

$$\frac{\partial \theta}{\partial \tau} > 0 \text{ on } \Gamma_-, \ \frac{\partial \theta}{\partial \tau} < 0 \text{ on } \Gamma_+. \tag{A.9}$$

From (2.3), (3.48), and (A.9), one gets that

$$\nabla\theta(x) \neq 0 \ \forall x \in \partial\Omega, \tag{A.10}$$

$$\text{degree } (\nabla\theta, \Omega, 0) = 0. \tag{A.11}$$

Let $f : \overline{\Omega} \subset \mathbb{R}^2 \cong \mathbb{C} \to \mathbb{C} \cong \mathbb{R}^2$ be defined by

$$f(x_1 + ix_2) = \frac{\partial \theta}{\partial x_1}(x_1, x_2) - i\frac{\partial \theta}{\partial x_2}(x_1, x_2). \tag{A.12}$$

Then f is holomorphic and, by (A.10), does not vanish on $\partial\Omega$; therefore the degree deg $(f, \Omega, 0)$ is well defined and is equal to the number of zeros of f, counted according to their multiplicity. But, by (A.11) and (A.12), degree $(f, \Omega, 0) = -$ degree $(\nabla\theta, \Omega, 0)$ $= 0$. Therefore f does not vanish on $\overline{\Omega}$, which proves Lemma 3.7.

Appendix B. Proof of Lemma 3.3. Let $\tau_- \in C^\infty(\Gamma_+ \cup \Gamma_-; \mathbb{R}^2)$ be defined by requiring

$$\tau_-(x) \in \{\tau(x), -\tau(x)\} \, \forall x \in \Gamma_+ \cup \Gamma_-, \tag{B.1}$$

$$\tau_-(x) \text{ is pointing outside } \Gamma_+ \cup \Gamma_- \, \forall x \in \partial\gamma_-. \tag{B.2}$$

Note that $\partial\gamma_-$ has two elements and is included in $\Gamma_+ \cup \Gamma_-$. It follows from (A.9) that

$$(\nabla\theta(x)) \cdot \tau_-(x) < 0 \ \forall x \in \partial\gamma_-,$$

which, with standard elliptic estimates, implies the existence of $M_0 > 0$ such that, for any $z \in C^0(\overline{\Omega}; \mathbb{R}^2)$ and for any $x \in \Gamma_+ \cup \Gamma_-$ such that dist $(x, \partial\gamma_-) \leqslant 1/M_0$,

$$\left(\begin{cases} z \cdot n = g \text{ on } \partial\Omega \\ \text{div } z = 0 \text{ in } \partial\Omega \\ M_0 | \text{ curl } z|_{L^\infty(\Omega)} \leqslant 1 \end{cases} \right) \Rightarrow \left(z(x) \cdot \tau_-(x) \leqslant -\frac{1}{M_0} \right). \tag{B.3}$$

Finally, let us remark that M_0 has the property required by Lemma 3.3. Indeed, let $y \in L^\infty\left((0,T);C^0\left(\overline{B_R};\mathbb{R}^2\right)\right)$ satisfying (3.5), (3.6) for some $K>0$, (3.38)–(3.39) for some $\alpha \in L^\infty((0,T);(0,+\infty))$, and (3.40). We argue by contradiction and therefore assume the existence of $(\tilde{t},\tilde{x}) \in (0,T] \times \overline{\gamma_-}$, and $\nu \in (0,\tilde{t})$ such that

$$\Phi^y\left(t,\tilde{t},\tilde{x}\right) \in \overline{\Omega}\, \forall t \in (\tilde{t}-\nu,\tilde{t}). \tag{B.4}$$

For $t \in [\tilde{t}-\nu,\tilde{t}]$, let $k(t)=\Phi^y\left(t,\tilde{t},\tilde{x}\right)$. We claim that there exists $\nu_1 \in (0,\nu]$ such that

$$k(t) \in \partial\Omega\, \forall t \in [\tilde{t}-\nu_1,\tilde{t}]. \tag{B.5}$$

Indeed, (B.5) follows easily from (3.38) and (B.4) if y is smooth enough, e.g., locally Lipschitz with respect to x. As in the proof of (3.41), the case where y is not smooth follows from the main ingredient, due to Wolibner [25], to prove the uniqueness of Φ^y in Theorem 3.1; since the proof is very similar to the proof of (3.41) given above, we omit it. From (3.38) and (B.5), we get that

$$k(t) \in \Gamma_+ \cup \Gamma_-\, \forall t \in [\tilde{t}-\nu_1,\tilde{t}]. \tag{B.6}$$

By (3.38), (3.39), (3.40), and (B.3), there exists $\nu_2 \in (0,\nu_1]$ such that

$$\dot{k}(t)\cdot\tau_-(k(t)) \leqslant -\frac{1}{M_0}\alpha(t)\, \forall t \in [\tilde{t}-\nu_2,\tilde{t}]. \tag{B.7}$$

Since $k(\tilde{t}) \in \overline{\gamma_-}$, and $\alpha(t)>0$ for a.e. $t \in (0,T)$, (B.6) and (B.7) are in contradiction—recall (B.2).

Acknowledgments. This work was completed while the author was visiting the Scuola Normale Superiore di Pisa. We are very thankful to G. Da Prato for his hospitality.

REFERENCES

[1] R. W. Brockett, *Asymptotic Stability and Feedback Stabilization*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Progr. Math. 27, Birkhäuser, Basel, Boston, 1983, pp. 181–191.

[2] J.-M. Coron, *Contrôlabilité exacte frontière de l'équation d'Euler des fluides parfaits incompressibles bidimensionnels*, C. R. Acad. Sci. Paris Sér. I Math., 317 (1993), pp. 271–276.

[3] J.-M. Coron, *On the controllability of the* 2*-D incompressible perfect fluids*, J. Math. Pures Appli., 75 (1996), pp. 155–188.

[4] J.-M. Coron, *On the controllability of the* 2*-D incompressible Navier-Stokes equations with the Navier slip boundary conditions*, ESAIM Control Optim. Calc. Var., 1 (1996), pp. 35–75; also available online from www.emath.fr/cocv/.

[5] J.-M. Coron and A. Fursikov, *Global exact controllabilty of the* 2*D Navier-Stokes equations on a manifold without boundary*, preprint, Russian J. Math. Phys., 4 (1996), pp. 429–448.

[6] A. Fursikov and O. Yu. Imanuvilov, *Local exact controllability of the Navier-Stokes equations*, C. R. Acad. Sci. Paris Ser. I Math., 323 (1996), pp. 275–280.

[7] A. Fursikov and O. Yu. Imanuvilov, *On exact boundary zero controllability of the two-dimensional Navier-Stokes equations*, Acta Appl. Math., 36 (1994), pp. 1–10.

[8] A. Fursikov and O. Yu. Imanuvilov, *Local exact controllability for* 2*-D Navier-Stokes equations*, Sbornik Math., 187 (1996), pp. 1355–1390.

[9] O. Glass, *Contrôlabilité exacte frontière de l'équation d'Euler des fluides parfaits incompressibles en dimension* 3, C. R. Acad. Sci. Paris Sér. I Math., 325 (1997) pp. 987–992.

[10] O. Glass, *Exact boundary controllability of* 3*-D Euler equation*, ESAIM Control Optim. Calc. Var., to appear.

[11] L. Hörmander, *Lectures on Nonlinear Hyperbolic Differential Equations*, Math. Appl. 26, Springer-Verlag, Berlin, Heidelberg, New York, 1997.

[12] O. Yu. Imanuvilov, *On exact controllability for Navier-Stokes equations*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 97–131; also available online from www.emath.fr/cocv/.

[13] J. Kurzweil, *On the inversion of Lyapunov's second theorem on stability of motion*, Ann. Math. Soc. Trans., Ser. 2, 24 (1956), pp. 19–77.

[14] T. Kato, *On classical solutions of the two-dimensional non-stationary Euler equation*, Arch. Rational Mech. Anal., 25 (1967), pp. 188–200.

[15] V. Komornik, *Rapid boundary stabilization of linear distributed systems*, SIAM J. Control Optim., 35 (1997) pp. 1591–1613.

[16] I. Lasiecka and R. Triggiani, *Differential and Algebraic Ricatti Equations with Applications to Boundary/Point Control Problems: Continuous and Approximation Theory*, Lecture Notes in Control and Inform. Sci. 164, Springer-Verlag, Berlin, Heidelberg, New York, 1991.

[17] J.-L. Lions, *Exact controllability, stabilizability, and perturbations for distributed systems*, SIAM Rev., 30 (1988) pp. 1–68.

[18] P.-L. Lions, *Mathematical Topics in Fluid Mechanics*, Vol. 1, Incompressible Models, Oxford Science Publication, Oxford, 1996.

[19] N. G. Lloyd, *Degree Theory*, Cambridge Tracts in Math. 73, Cambridge University Press, Cambridge, 1978.

[20] A. Majda, *Vorticity and the mathematical theory of incompressible fluid flow*, Comm. Pure Appl. Math., 39, special issue, (1986), pp. 187–220.

[21] O. A. Oleĭnik, *Discontinuous Solutions of Non-Linear Differential Equations*, Amer. Math. Soc. Transl. Ser. 2, 26, Providence, RI, 1957.

[22] M. Slemrod, *A note on complete controllability and stabilizability for linear control systems in Hilbert space*, SIAM J. Control, 12 (1974), pp. 500–508.

[23] E. D. Sontag and H. J. Sussmann, *Remarks on continuous feedback*, in Proceedings, IEEE Conference on Decision and Control, Albuquerque, NM, 1980, pp. 916–921.

[24] H. J. Sussmann, *Subanalytic sets and feedback control*, J. Differential Equations, 31 (1979), pp. 31–52.

[25] W. Wolibner, *Un théorème sur l'existence du mouvement plan d'un fluide parfait, homogène, incompressible, pendant un temps infiniment long*, Math. Z., 37 (1933), pp. 698–726.

[26] V. I. Yudovich, *Nonstationnary flows of an ideal incompressible fluid*, Zh. Vyčisl. Mat. i. Mat. Fis., 3 (1963), pp. 1032–1066 (in Russian).

SIAM J. Control Optim.
Vol. 37, No. 6, pp. 1897–1912

SHIFT OPERATOR INDUCED APPROXIMATIONS OF DELAY SYSTEMS*

P. M. MÄKILÄ[†] and J. R. PARTINGTON[‡]

Abstract. Certain shift operator induced finite-dimensional approximations of an important class of delay systems are studied. H_∞, H_2, and L_1 error bounds are given for stable delay systems. Furthermore, a multiple shift formula, generated by a second-order Padé shift, is analyzed further, and an exact asymptotic H_∞ error formula is derived. The studied class of shift operator induced approximations provides an appealing and most transparent approach for determining good finite-dimensional models of delay systems.

Key words. approximation by exponentials, delay systems, Padé shifts, shift operators

AMS subject classifications. 93B36, 93C80, 30E10, 41A21, 47N70

PII. S0363012998339678

1. Introduction.

This paper studies certain shift operator induced approximations of a class of stable delay systems. This belongs to the classic topic of approximation of stable systems by exponentials (i.e., by finite-dimensional systems); see, e.g., Wiener [38], Kautz [15], Horowitz [13], McDonough and Huggins [26], and Kammler [14]. The new approximation approach studied here provides the most transparent approximations of delay systems with minimum effort, which makes it ideally suited for control engineering textbooks and courses. The studied approach should make it possible to reintroduce delay systems as a main type of plant model in modern undergraduate and graduate level control engineering textbooks, as once was the case in classical textbooks on process control.

There are several norms that are of interest as a measure of the approximation accuracy, especially the L_1 norm (see Glover, Curtain, and Partington [7], Mäkilä [19, 20], Partington [28], and Mäkilä [22]), the L_2 norm (see Glover, Curtain, and Partington [7], Glover, Lam, and Partington [10], Mäkilä [20], Wahlberg and Mäkilä [36]), and the H_∞ norm (see Glover, Lam, and Partington [8, 9], Lam [17], Mäkilä [20], Partington [28, 29], and Wahlberg and Mäkilä [36]).

Classically, most of the interest in approximation by exponentials has been associated with certain orthonormal exponential bases, especially exponential Laguerre and Legendre functions (see Szegő [35], Wiener [38], and Horowitz [13]). The classical theory of orthonormal polynomials has been very useful in analyzing Laguerre, Legendre, and Kautz approximations of linear systems, especially of delay systems (see Glader et al. [6], Wahlberg and Mäkilä [36], and Mäkilä [21]).

In the Laplace transform domain the above-mentioned orthonormal basis approximations produce rational approximations with fixed poles. Walsh [37] is a classical reference on fixed pole rational approximation, which has many applications to H_2 and H_∞ approximation and to identification of systems (see Mäkilä and Partington [23], Mäkilä, Partington, and Gustafsson [25], Ninness and Gustafsson [27], and

*Received by the editors May 29, 1998; accepted for publication (in revised form) March 22, 1999; published electronically October 8, 1999.

http://www.siam.org/journals/sicon/37-6/33967.html

[†]Department of Automation Engineering, Tampere University of Technology, P.O. Box 692, FIN-33101 Tampere, Finland (pmakila@ae.tut.fi). The research of this author was supported by the Academy of Finland.

[‡]School of Mathematics, University of Leeds, Leeds LS2 9JT, UK (J.R.Partington@leeds.ac.uk).

Heuberger, Van den Hof, and Bosgra [11]).

More recently there has been growing interest in rational approximations based on rational wavelets (see Pati [31], Pati and Krishnaprasad [32], Dudley, Ward, and Partington [2, 3, 4], and Partington [30]), which allow the choice of the pole locations from certain infinite sets. Rational wavelets are examples of redundant, rich, model sets, or dictionaries, which offer the potential for better degree of approximation than expansions in a single orthonormal basis.

Shift operators are very important in operator theory and in its many applications (see Rosenblum and Rovnyak [34]). It turns out that shift operator techniques are also very useful in approximation of linear systems by finite-dimensional systems (see Wahlberg and Mäkilä [36]). This is true partly because many important orthonormal bases, such as Laguerre and Kautz bases, are known to be induced by corresponding shift operators. It is also known that the delay operator is a shift operator. Certain Padé approximations of delay systems are also known to be associated with shift operators (for general material on Padé approximations see, e.g., Petrushev and Popov [33]). Therefore, shift operator techniques carry much potential for the purpose of modeling linear systems. Very detailed and strong results can be derived for certain shift operator approximations of delay systems. In Mäkilä and Partington [24], certain multiple Laguerre and Kautz shift induced approximations were studied in detail. In the present work a more general class of shift operator induced approximations of a class of delay systems is studied. The studied approach allows one to determine good low-order approximations of delay systems in a most transparent manner without any computations. Due to the elegant simplicity of the approximations, their error behavior can be analyzed to an unusual degree of completeness.

The rest of this paper is organized as follows. In section 2 some mathematical background material is reviewed briefly. Section 3 introduces the class of shift operator induced approximations of delay systems that will be studied in the present work. This approximation technique can also be interpreted in terms of all-pass rational approximations of the Laplace domain delay operator. H_∞ and H_2 error bounds for this technique are given here, too. In addition, L_1 error bounds are derived. In section 4 a special case of the general technique is studied in detail, namely, the so-called Padé-2 shift formula. Some very detailed H_∞ error bounds are given here and an asymptotically exact H_∞ error formula is derived. Section 5 provides a numerical example. Finally some conclusions are drawn in section 6.

2. Mathematical preliminaries.

Let H_p denote the Hardy space of functions f, over the real field, which are analytic in the open right half complex plane such that $|f(\sigma + j\omega)|^p$ is integrable for each $\sigma > 0$, and $\sup_{\sigma>0}[\int_{-\infty}^{\infty} |f(\sigma + j\omega)|^p \, d\omega]^{1/p} < \infty$, where $1 \leq p < \infty$. Here f is defined over the real field, meaning that $f(s) = \overline{f(\bar{s})}$ (f has real-valued coefficients only). Of special interest to us are the Hardy spaces H_2 and H_∞. In H_2 we shall use the norm $\|f\|_2 \equiv \sqrt{1/(2\pi)} \sup_{\sigma>0}[\int_{-\infty}^{\infty} |f(\sigma + j\omega)|^2 \, d\omega]^{1/2}$. The choice of the coefficient $\sqrt{1/(2\pi)}$ in the definition of the H_2 norm is convenient as this makes the Laplace transform operator $\mathcal{L}$ an isometry from $L_2(0, \infty)$ onto H_2. Finally, the Hardy space H_∞ of bounded analytic functions f in the right half plane is equipped with the norm $\|f\|_\infty = \sup_{Re\, s>0} |f(s)| < \infty$. (Note that the notation $\| \cdot \|_p$ is used to denote both the L_p norm and the H_p norm.)

Let H be a Hilbert space with inner product $(\cdot, \cdot)$. The norm for $x \in H$ is defined as $\|x\| \equiv (x, x)^{1/2}$. Let A denote a bounded linear operator from the Hilbert space H into H. Let A^* denote the adjoint operator of A, that is, $(Ax, y) = (x, A^* y)$ for all $x, y \in H$. A bounded operator $S : H \to H$ is said to be a *shift operator* if S is an isometry and $S^{*n} \to 0$ strongly (that is, $\|S^{*n}x\| \to 0$ for any x in H). A subspace E of H is called *cyclic* for the bounded operator A if $\bigvee_{k\geq 0} A^k E = H$. (Here $\bigvee$ denotes linear envelope.) The *multiplicity* of a shift operator S is the minimum dimension of a cyclic subspace for S. Multiplicity of a shift operator is important as two shift operators are unitarily equivalent if and only if they have the same multiplicity. (Unitarily equivalent operators are indistinguishable for many purposes.) Rosenblum and Rovnyak [34] give more mathematical background material on shift operators.

An example of a shift operator on H_2 is the so-called *Laguerre shift* S_L, defined by $S_L : G(s) \to G(s)(s-1/2)/(s+1/2)$ for any $G \in H_2$. (This shift is simply the Laplace transform of the (unitarily equivalent) Laguerre shift on $L_2(0, \infty)$; see Rosenblum and Rovnyak [34].) The multiplicity of the Laguerre shift is one as $\bigvee_{k\geq 0} S_L^k 1/(s+1/2) = H_2$ by a classical closure result based on Laguerre functions (see Szegő [35]).

A fundamental shift operator is the delay operator. Let S_h denote the shift operator on H_2, defined by $(S_h f)(s) = e^{-hs} f(s)$, $f \in H_2$. Here $h > 0$ corresponds to the time delay. Therefore, this operator corresponds to multiplication with e^{-hs}. The presence of this operator in most real problems makes control and systems theory a highly nontrivial theory. Note that the multiplicity of the shift S_h on H_2 is infinite.

Thus the delay operator on H_2 is in a certain sense a complex shift operator. The corresponding right shift operator, S_d, defined on the space of square summable sequences ℓ_2 by $S_d(x_1, x_2, \ldots) = (0, x_1, x_2, \ldots)$, for any $x \in \ell_2$, is clearly of multiplicity one and so has mathematically a rather simple structure. It is for this reason that the theory of the shift operator is rather more difficult in continuous time control and systems theory than in discrete time.

More generally it can be shown that any operator of multiplication on H_2 by a nonconstant inner function u gives rise to a shift operator, which has finite multiplicity if and only if u is a Blaschke product. See, for example, [1, Chapter 1].

Finally, the standard notation $s_n = O(f_n)$ means that the nonnegative sequences $\{s_n\}$, $\{f_n\}_{n\geq 1}$ satisfy $s_n \leq C f_n$, for some constant $C > 0$, for n large enough.

3. A general multiple shift formula.

In this section we consider approximation of stable delay systems by certain multiple shift formulas.

Recall that multiplication with e^{-hs} is a shift operator on H_2. Note that the best rational H_∞ approximation, of any degree, of e^{-hs} is the zero function. The transfer function e^{-hs} corresponds to the pure transmission line, the so-called *delay line*, for which many classical approximation methods such as Padé approximations and Bessel approximations have been used (see Kuo [16], Glover, Lam, and Partington [10], and Glader et al. [6]).

We shall study delay systems of the form $G(s) = e^{-hs}R(s)$, where $h > 0$ and $R \neq 0$ is a stable, strictly proper, rational transfer function. Let $m \geq 1$ denote the relative degree of $R(s)$. (Recall that the relative degree of a rational transfer function $R(s)$ is defined as the difference between the degrees of the denominator polynomial and the numerator polynomial of $R(s)$.)

An even simpler approximation technique than the full Padé approximation is based on the relationship

$$e^{-hs} = \lim_{n\to\infty} \left(\frac{1 - \frac{hs}{2n}}{1 + \frac{hs}{2n}} \right)^n . \tag{3.1}$$

Note that the shift operator $S_L^{(n)}$ defined by

$$S_L^{(n)} f \equiv \left(\frac{1 - \frac{hs}{2n}}{1 + \frac{hs}{2n}} \right)^n f,\ f \in H_2, \tag{3.2}$$

is essentially a multiple Laguerre shift of multiplicity n. Thus this approximation procedure involves approximating the shift operator corresponding to multiplication with e^{-hs} by the shift operator $S_L^{(n)}$.

This suggests a Laguerre shift type approximation of a stable delay system of the earlier mentioned form $G(s) = e^{-hs}R(s)$, namely, the approximation $S_L^{(n)}R(s)$. This type of approximation has been studied in Lam [18] and Mäkilä and Partington [24]. Such approximations are easy to compute and also easy to implement using analog filters.

The Laguerre formula is an example of a much more general construction, which we now analyze.

Let $u(s)$ be an inner function (a stable all-pass function) that lies in H_∞ on the right-hand half plane and is analytic in a neighborhood of the origin. For physical reasons we shall assume that $u(\overline{s}) = \overline{u(s)}$, and for the purposes of rational approximation we shall normally require u to be rational. We shall choose the normalization $u(0) = 1$.

One simple way to construct examples of such functions is to write

$$u(s) = \frac{p(-s)}{p(s)}, \tag{3.3}$$

where p is a real polynomial with no zeroes in the closed right-hand half plane. An example analyzed in Mäkilä and Partington [24] is $p(s) = 1 + s/2$; another important example (based on Padé approximation) is $p(s) = 1 + s/2 + s^2/12$. This latter example will be studied in more detail in the next section.

The intention is to approximate an exponential e^{-hs} by the inner function

$$u_n(hs) := (u(hs/n))^n. \tag{3.4}$$

Note first that, using the Taylor expansion, there is an index $k \geq 1$ and constants A, B, and $C > 0$ such that

$$A|\omega|^k \leq \left| e^{-j\omega} - u(j\omega) \right| \leq B|\omega|^k \qquad \text{for} \quad |\omega| \leq C. \tag{3.5}$$

We also have

$$\|e^{-j\omega} - u(j\omega)\|_\infty \leq 2, \tag{3.6}$$

since both functions are inner.

Note that for nonzero $R(s)$ with relative degree $m (\geq 1)$, there exist positive constants D, E, and F, such that

$$E|\omega|^{-m} \leq |R(j\omega)| \leq F|\omega|^{-m} \qquad \text{for } |\omega| \geq D. \tag{3.7}$$

The following lemma shows that dilations of u provide approximations to the exponential function with, in general, improved error bounds.

LEMMA 3.1. *Under the hypotheses above,*

$$|e^{-jh\omega} - u_n(jh\omega)| \leq \min\left(2, B\frac{h^k|\omega|^k}{n^{k-1}}\right) \qquad \textit{for} \quad |\omega| \leq nC/h. \tag{3.8}$$

Proof. We use the formula $(a^n - b^n) = (a-b)(a^{n-1} + a^{n-2}b + \cdots + ab^{n-2} + b^{n-1})$, with $a = e^{-jh\omega/n}$ and $b = u(jh\omega/n)$. Since $|a-b| \le B|h\omega/n|^k$ for $|h\omega/n| \le C$, and $|a| = |b| = 1$, we obtain the result. □

The following general result shows how the dilations of u provide shift operators which converge strongly to the exponential shift. For a function G defined on the right half plane, we define $S_h G$ by $(S_h G)(s) = e^{-hs}G(s)$ and $S_{u,n,h}G$ by $(S_{u,n,h}G)(s) = u_n(hs)G(s)$. We write A_0 for the closed subspace of H_∞ consisting of functions G continuous on the closed right half plane, such that $G(j\omega) \to 0$ as $\omega \to \pm\infty$.

COROLLARY 3.1. *Let u be a rational inner function satisfying* (3.5) *with $k > 1$. Then for every $G \in H_2$ one has $\|S_{u,n,h}G - S_h G\|_2 \to 0$ as $n \to \infty$; and for every $G \in A_0$ one has $\|S_{u,n,h}G - S_h G\|_\infty \to 0$ as $n \to \infty$.*

Proof. By Lemma 3.1 we see that $u_n(jh\omega) \to e^{-jh\omega}$ pointwise as $n \to \infty$, and hence, since e^{-hs} and $u_n(hs)$ are inner, Lebesgue's dominated convergence theorem implies that $\|S_{u,n,h}G - S_h G\|_2 \to 0$ when $G \in H_2$. Moreover the bound in Lemma 3.1 shows that $S_{u,n,h}G$ tends uniformly to $S_h G$ on the imaginary axis, since, if $\epsilon > 0$ is given, there is an M such that $|G(j\omega)| < \epsilon/2$, whenever $|\omega| \ge M$, and hence $|(e^{-jh\omega} - u_n(jh\omega))G(j\omega)| < \epsilon$ if $|\omega| \ge M$. It is now necessary to show only that the same applies for $|\omega| < M$, if n is sufficiently large, and this follows from Lemma 3.1. □

We define ω_n and ω_n^* to be the solutions of the equations

$$Bh^k\omega_n^k/n^{k-1} = 2 \qquad \text{and} \qquad Bh^k{\omega_n^*}^k/n^{k-1} = \sqrt{2}, \tag{3.9}$$

so that ω_n and ω_n^* grow as $n^{(k-1)/k}$.

Let R be a stable rational function satisfying (3.7). We may choose D large enough that $FD^{-m} \le \|R\|_\infty$.

THEOREM 3.1. *Under the hypotheses above, provided that $D \le \omega_n \le Cn/h$, one has*

$$\|R(s)e^{-sh} - R(s)u_n(sh)\|_\infty \le \max\left\{ \|R\|_\infty \frac{Bh^kD^k}{n^{k-1}}, \max_{\omega\in\{D,\omega_n\}} \frac{FBh^k\omega^{k-m}}{n^{k-1}}, 2F\omega_n^{-m} \right\}. \tag{3.10}$$

Proof. This follows by considering the intervals $[0, D]$, $[D, \omega_n]$, and $[\omega_n, \infty)$ separately, using the estimates in Lemma 3.1. □

Similar analyses can be performed for small n, if $D \le \omega_n \le Cn/h$ is no longer satisfied, by considering intervals in the same way. We shall see specific examples of this later.

To establish some lower bounds, we need an estimate corresponding to that in Lemma 3.1.

LEMMA 3.2. *Under the hypotheses above,*

$$|e^{-jh\omega} - u_n(jh\omega)| \ge \frac{2\sqrt{2}A}{\pi}\frac{h^k|\omega|^k}{n^{k-1}}, \qquad \textit{for} \quad |\omega| \le \min\{\omega_n^*, Cn/h\}. \tag{3.11}$$

Proof. Without loss of generality we can take $\omega \ge 0$. We consider the argument of $u_n(jh\omega)$ and use the standard inequalities

$$\frac{2\sqrt{2}}{\pi}|\theta| \le \left|1 - e^{i\theta}\right| \le |\theta| \qquad \text{for } |\theta| \le \pi/2, \tag{3.12}$$

implying that

$$\frac{2\sqrt{2}}{\pi}|\arg z - \arg w| \le |z - w| \le |\arg z - \arg w| \tag{3.13}$$

whenever $|z| = |w| = 1$ and $|\arg z - \arg w| \leq \pi/2$. Hence

$$A\left(\frac{h\omega}{n}\right)^k \leq \left|\arg(u(j\omega h/n)) - \frac{h\omega}{n}\right| \leq \frac{\pi}{2\sqrt{2}} B\left(\frac{h\omega}{n}\right)^k \leq \frac{\pi}{2n} \tag{3.14}$$

for $|h\omega/n| \leq C$ and $|\omega| \leq \omega_n^*$, and hence

$$|\arg(u_n(j\omega h)) - h\omega| \geq nA\left(\frac{h\omega}{n}\right)^k, \tag{3.15}$$

which implies the asserted inequality. □

We can now derive a lower bound, corresponding to the upper bound in Theorem 3.1.

THEOREM 3.2. *Under the hypotheses above, provided that $D \leq \omega_n^* \leq Cn/h$, one has*

$$\|R(s)e^{-sh} - R(s)u_n(sh)\|_\infty \geq \max_{\omega \in \{D, \omega_n^*\}} \frac{2\sqrt{2}AE}{\pi} \frac{h^k \omega^{k-m}}{n^{k-1}}. \tag{3.16}$$

Proof. This follows at once from Lemma 3.2, using (3.7). □

COROLLARY 3.2. *Under the hypotheses of Theorem* 3.1*, the asymptotic rate of H_∞ convergence of $R(s)u_n(sh)$ to $R(s)e^{-sh}$ as $n \to \infty$ is of order exactly $n^{-\lambda}$, where $\lambda = (k-1)\min(1, m/k)$.*

Proof. This follows immediately from Theorems 3.1 and 3.2, on considering the cases $m \leq k$ and $m \geq k$ separately. □

For example, if $k = 3$, which is the case with the function $u(s) = (1 - s/2)/(1 + s/2)$, analyzed in Mäkilä and Partington [24], then the exact H_∞ convergence rates for $m = 1, 2, 3, 4, \ldots$, are of orders $n^{-2/3}$, $n^{-4/3}$, n^{-2}, $n^{-2}, \ldots$.

The bounds given in Theorems 3.1 and 3.2 can also be used to determine the exact H_2 convergence rates.

THEOREM 3.3. *Under the hypotheses of Theorem* 3.1*, the asymptotic rate of H_2 convergence of $R(s)u_n(sh)$ to $R(s)e^{-sh}$ as $n \to \infty$ is of order exactly $n^{-\mu}$, where $\mu = (k-1)\min(1, (2m-1)/(2k))$.*

Proof. Writing $K_1, K_2, \ldots$ for positive constants which we do not need to write down explicitly, we find that the estimates used in Lemma 3.1 and Theorem 3.1 show that the square of an upper H_2 bound is

$$K_1 \int_0^D \frac{\omega^{2k}}{n^{2k-2}}\, d\omega + K_2 \int_D^{\omega_n} \frac{\omega^{2k-2m}}{n^{2k-2}}\, d\omega + K_3 \int_{\omega_n}^\infty \omega^{-2m}\, d\omega, \tag{3.17}$$

which, using the fact that ω_n is of order $n^{(k-1)/k}$, gives three terms, each of order at most the greater of $n^{-(2k-2)}$ or $n^{-(k-1)(2m-1)/k}$.

Likewise a lower bound can be obtained by using the estimates in Lemma 3.2 and Theorem 3.2 and integrating from D to ω_n^*. This gives as the square of a lower bound the quantity

$$K_4 \int_D^{\omega_n^*} \frac{\omega^{2k-2m}}{n^{2k-2}}\, d\omega, \tag{3.18}$$

which, since ω_n^* is of order $n^{(k-1)/k}$, produces a lower bound the same as the upper bound (to within a factor independent of n). □

For example, if $k = 3$, as for the function $u(s) = (1 - s/2)/(1 + s/2)$, discussed above, the exact H_2 convergence rates for $m = 1, 2, 3, 4, 5, \ldots$, are of orders $n^{-1/3}$, n^{-1}, $n^{-5/3}$, n^{-2}, $n^{-2}, \ldots$, which is a result given in Mäkilä and Partington [24].

We now turn our attention to L_1 error bounds for the corresponding impulse response functions. The following lemma shows that they are not very much worse than the H_∞ bounds, and it applies in a very general setting. We let $\mathcal{L}$ denote the Laplace transform operator, which maps $L_1(0,\infty)$ contractively into H_∞.

LEMMA 3.3. *Let $g \in L_1(0,\infty)$ and let (g_n) be any sequence of degree-n approximants in $L_1(0,\infty)$ such that $\|\mathcal{L}g_n - \mathcal{L}g\|_\infty = O(n^{-\lambda})$ as $n \to \infty$, for some $\lambda > 1$. Then $\|g_n - g\|_1 = O(n^{1-\lambda})$ as $n \to \infty$.*

Proof. It is well known (see also Partington [29]) that for functions $f \in L_1$ whose Laplace transforms are rational of degree n the inequality $\|g\|_1 \le 2n\|\mathcal{L}g\|_\infty$ holds. We therefore see that

$$\|g_{2^r} - g_{2^{r+1}}\|_1 \le 2(2^r + 2^{r+1})\|\mathcal{L}g_{2^r} - \mathcal{L}g_{2^{r+1}}\|_\infty = O(2^{r-\lambda r}), \tag{3.19}$$

implying that (g_{2^r}) is a Cauchy sequence in L_1 whose limit must be g. Hence $\|g_{2^r} - g\|_1 = O(2^{r-\lambda r})$. To obtain the result for general n we choose r such that $2^r \le n < 2^{r+1}$ and use the inequality $\|g_n - g_{2^r}\|_1 \le 2(n+2^r)\|\mathcal{L}g_n - \mathcal{L}g_{2^r}\|_\infty = O(n^{1-\lambda})$. □

A rather better bound can be obtained by using the continuous-time version of the Hardy–Littlewood inequality

$$\|g\|_1 \le K_1\|(\mathcal{L}g)'\|_{H_1} \tag{3.20}$$

(see also [5, Ex. 11.3], Hille and Tamarkin [12]). (In this and the following calculation, we shall use K_1, K_2, etc., to denote positive constants that we do not estimate.) To do this, we need to bound the H_1 norm of the derivative of $R(s)(e^{-hs} - u_n(hs))$. Note that it follows from (3.5) that $|-e^{-s} - u'(s)| \le K_2|s|^{k-1}$ for $|s|$ sufficiently small. It is also clear that $u'(s) \in H_\infty$ if u is rational and inner. Moreover we have the following bound: for $n \ge 2$,

$$|e^{-jh\omega(n-1)/n} - u(jh\omega/n)^{n-1}| \le \min\left(2, K_3\frac{h^k|\omega|^k}{n^{k-1}}\right) \quad \text{for } |\omega| \le nK_4/h, \tag{3.21}$$

analogous to Lemma 3.1. Putting these inequalities together and using the inequality

$$|a^n - b^{n-1}c| \le |a^{n-1}(a-c)| + |(a^{n-1} - b^{n-1})c| \tag{3.22}$$

with $a = e^{-jh\omega/n}$, $b = u(jh\omega/n)$, and $c = -u'(jh\omega/n)$, we obtain a bound

$$\begin{aligned} &|-he^{-jh\omega} - hu(jh\omega/n)^{n-1}u'(jh\omega/n)| \\ &\quad \le \min\left(K_5, K_6\frac{h^k|\omega|^{k-1}}{n^{k-1}}\right) + \min\left(K_7, K_8\frac{h^{k+1}|\omega|^k}{n^{k-1}}\right), \end{aligned} \tag{3.23}$$

valid for $|\omega| \le nK_9/h$.

THEOREM 3.4. *Under the hypotheses of Theorem 3.1, provided that m, the relative degree of $R(s)$, is at least 2, then the L_1 errors between the impulse responses corresponding to $R(s)e^{-hs}$ and $R(s)u_n(hs)$ are of order at most $n^{-\nu}$ as $n \to \infty$, where $\nu = (k-1)\min\{1, (m-1)/k\}$.*

Proof. We bound the H_1 norm of the derivative of $Q(s) := R(s)(e^{-hs} - u_n(hs))$, as above. Since the calculation is similar to several others in this section, we shall give only a brief summary of it. Now

$$\|Q'(s)\|_{H_1} \le \|R'(s)(e^{-hs} - u_n(hs))\|_{H_1} + \|R(s)(-he^{-hs} - hu(hs/n)^{n-1}u'(hs/n))\|_{H_1}. \tag{3.24}$$

Note that $R'(s)$ is rational with relative degree $m+1$, since $R(s)$ is rational with relative degree m. For the first term in (3.24), we have pointwise bounds of orders ω^k/n^{k-1}, ω^{k-m-1}/n^{k-1}, and ω^{-m-1}, valid for various eventually overlapping ranges of ω, as in Theorem 3.1 with m replaced by $(m+1)$; for the second they become ω^{k-1}/n^{k-1}, ω^{k-m}/n^{k-1}, and ω^{-m}. Performing the integrals along the imaginary axis, we obtain the H_1 bounds $n^{-\nu_1}$ and $n^{-\nu_2}$, where $\nu_1 = (k-1)\min\{1, m/k\}$ and $\nu_2 = (k-1)\min\{1, (m-1)/k\} \leq \nu_1$. Finally, using (3.20) we obtain the desired result. □

Thus, for the case $k = 3$ discussed above, the L_1 convergence rates when m takes the values $2, 3, 4, 5, \ldots$ are at worst $O(n^{-2/3})$, $O(n^{-4/3})$, $O(n^{-2})$, $O(n^{-2}), \ldots$, which is an improvement on the bounds that one would obtain from Corollary 3.2 and Lemma 3.3.

4. The multiple Padé-2 shift formula. An important shift operator is the Kautz shift, $S_K : H_2 \to H_2$ (see Wahlberg and Mäkilä [36]), defined by

$$S_K f = \left(\frac{s^2 - bs + c}{s^2 + bs + c}\right) f,\ f \in H_2, \tag{4.1}$$

where $b > 0$ and $b^2 - 4c < 0$. This shift operator has multiplicity 2 and corresponds to multiplication by a second-order rational all-pass function with two complex (strictly nonreal) *stable* poles.

Analogous to the Laguerre formula for e^{-hs}, the Kautz shift gives the following (Kautz) formula for e^{-hs}:

$$e^{-hs} = \lim_{n\to\infty} \left(\frac{1 - \frac{hs}{2n} + \frac{1}{3}\left(\frac{hs}{2n}\right)^2}{1 + \frac{hs}{2n} + \frac{1}{3}\left(\frac{hs}{2n}\right)^2}\right)^n . \tag{4.2}$$

The shift operator $S_P^{(n)} : H_2 \to H_2$ defined by

$$S_P^{(n)} f = \left(\frac{1 - \frac{hs}{2n} + \frac{1}{3}\left(\frac{hs}{2n}\right)^2}{1 + \frac{hs}{2n} + \frac{1}{3}\left(\frac{hs}{2n}\right)^2}\right)^n f,\ f \in H_2, \tag{4.3}$$

is a multiple Kautz shift of multiplicity $2n$. As this shift is motivated by the second-order Padé approximation of e^{-hs}, we shall call it *the multiple Padé-2 shift.*

Introduce the monotonically increasing continuous function

$$\varphi_k(\nu) = \sup_{0\leq\omega\leq\nu} \omega^k |R(j\omega)|,\ \nu \geq 0. \tag{4.4}$$

Note that for nonzero $R(s)$ with relative degree $m(\geq 1)$

$$|R(j\omega)| \leq \|R\|_\infty C^m \omega^{-m}, \quad \omega > C, \tag{4.5}$$

for some constant $C > 0$. We shall next study the approximation of $G(s) = e^{-hs}R(s)$ by $S_P^{(n)} R(s)$.

THEOREM 4.1. *Let $G(s) = e^{-hs}R(s)$, where $h > 0$ and $R(s) \not\equiv 0$ is a stable rational transfer function with relative degree $m \geq 1$. Then*

$$\|G - S_P^{(n)} R\|_\infty \geq \frac{h^5}{1080\pi} \varphi_5(\omega_n(h)) \times n^{-4}, \quad n \geq 1, \tag{4.6}$$

where $\omega_n(h) = 2h^{-1}n^{4/5}$ and φ_5 is defined in (4.4).

Hence, for $m = 1, 2, 3, 4$ *if* $n < (Dh/2)^{5/4}$, *and for* $m \geq 5$ *for any* $n \geq 1$,

$$\|G - S_P^{(n)} R\|_\infty \geq \frac{h^5}{1080\pi} \varphi_5(\omega_1(h)) \times n^{-4}. \tag{4.7}$$

Finally, for $m = 1, 2, 3, 4$ *if* $n \geq (Dh/2)^{5/4}$,

$$\|G - S_P^{(n)} R\|_\infty \geq \frac{4}{135\pi} (h/2)^m E \times n^{-(4m/5)}. \tag{4.8}$$

Proof. Clearly,

$$\|G - S_P^{(n)} R\|_\infty = \sup_{\omega \geq 0} |R(j\omega)| \left| e^{-jh\omega} - \left(\frac{1 - \frac{jh\omega}{2n} + \frac{1}{3}\left(\frac{jh\omega}{2n}\right)^2}{1 + \frac{jh\omega}{2n} + \frac{1}{3}\left(\frac{jh\omega}{2n}\right)^2} \right)^n \right|. \tag{4.9}$$

Furthermore,

$$\left(\frac{1 - \frac{jh\omega}{2n} + \frac{1}{3}\left(\frac{jh\omega}{2n}\right)^2}{1 + \frac{jh\omega}{2n} + \frac{1}{3}\left(\frac{jh\omega}{2n}\right)^2} \right)^n = e^{-j2n \arctan \frac{(h\omega/2n)}{1-(1/3)(h\omega/2n)^2}} \tag{4.10}$$

for $(1/3)(h\omega/2n)^2 < 1$, that is, for $(0 \leq)\omega < 2\sqrt{3}h^{-1}n$. (Note that the range of arctan is $(-\pi/2, \pi/2)$.)

It follows immediately from the formula $|e^{j\theta} - 1| = 2|\sin \frac{\theta}{2}|$ that

$$\begin{aligned} &\left| e^{-jh\omega} - e^{-j2n \arctan \frac{(h\omega/2n)}{1-(1/3)(h\omega/2n)^2}} \right| \\ &\quad = 2 \left| \sin \left[\frac{h\omega}{2} - n \arctan \frac{(h\omega/2n)}{1 - (1/3)(h\omega/2n)^2} \right] \right| \end{aligned} \tag{4.11}$$

for $(0 \leq)\omega < 2\sqrt{3}h^{-1}n$.

We shall need the auxiliary inequalities

$$0 \leq y - \arctan\left(\frac{y}{1 - (1/3)y^2} \right) \leq \frac{1}{45} y^5 \text{ for } 0 \leq y < \sqrt{3}. \tag{4.12}$$

To prove the right-hand side inequality consider the function

$$g(y) = y - \arctan \frac{y}{1 - (1/3)y^2} - \gamma y^5. \tag{4.13}$$

Then clearly $g(0) = 0$ and

$$g'(y) = y^4[(1/9) - 5\gamma - (5/3)\gamma y^2 - (5/9)\gamma y^4]/(1 + (1/3)y^2 + (1/9)y^4). \tag{4.14}$$

Hence $\gamma = (1/45)$ is the best (smallest) constant to get $g(y) \geq 0$ for all $y \geq 0$ up to the first discontinuity of g at $y = \sqrt{3}$. The proof of the left-hand inequality follows by using in g above $\gamma = 0$ and noting that then $g'(y) \geq 0$ for $0 \leq y < \sqrt{3}$.

Denote now $y = h\omega/(2n)$. Then

$$\begin{aligned} 2 \left| \sin \left[\frac{h\omega}{2} - n \arctan \frac{(h\omega/2n)}{1 - (1/3)(h\omega/2n)^2} \right] \right| &= 2 \left| \sin n \left(y - \arctan \frac{y}{1 - (1/3)y^2} \right) \right| \\ &\geq 2(2/\pi) n \left[y - \arctan \left(\frac{y}{1 - (1/3)y^2} \right) \right], \end{aligned} \tag{4.15}$$

if $0 \le n[y - \arctan(\frac{y}{1-(1/3)y^2})] \le (\pi/2)$ and $0 \le y < \sqrt{3}$. (The latter condition means that $0 \le \omega < 2\sqrt{3}h^{-1}n$.) By (4.12) it is seen that (4.15) holds if

$$\frac{1}{45}y^5 \le \frac{\pi}{2}\frac{1}{n} \text{ and } 0 \le y < \sqrt{3}. \tag{4.16}$$

That is, (4.15) holds if

$$y \le (45\pi/2)^{1/5} n^{-1/5} \text{ and } 0 \le y < \sqrt{3}. \tag{4.17}$$

Next we show that the auxiliary inequality

$$y - \arctan\left(\frac{y}{1-(1/3)y^2}\right) \ge \frac{1}{135}y^5 \tag{4.18}$$

is valid for $0 \le y < \sqrt{3}$. Consider the function $g(y)$ as above. From the expression of $g'(y)$ we see that $\gamma = 1/135$ is actually the largest constant for which $g'(y) \ge 0$ for $0 \le y < \sqrt{3}$. This proves the inequality (4.18).

Using (4.18) in (4.15) gives that

$$2\left|\sin\left[\frac{h\omega}{2} - n\arctan\frac{(h\omega/2n)}{1-(1/3)(h\omega/2n)^2}\right]\right| \ge 2 \times (2/\pi) \times n \times \frac{1}{135}\left(\frac{h\omega}{2n}\right)^5, \tag{4.19}$$

for any $n \ge 1$ when $0 \le \omega h \le n^{-1/5} \times 2n = 2n^{4/5}$. Therefore

$$\begin{aligned} \|G - S_P^{(n)}R\|_\infty &\ge \frac{h^5}{1080\pi}\sup_{0\le\omega\le\omega_n(h)} \omega^5|R(j\omega)| \times n^{-4} \\ &= \frac{h^5}{1080\pi}\varphi_5(\omega_n(h)) \times n^{-4}, \end{aligned} \tag{4.20}$$

where $\omega_n(h) = 2h^{-1}n^{4/5}$. This proves the first part of the theorem.

Note that, for any $m \ge 1$,

$$\|G - S_P^{(n)}R\|_\infty \ge \frac{h^5}{1080\pi}\varphi_5(\omega_1(h)) \times n^{-5}, \quad n \ge 1, \tag{4.21}$$

by the first part of the theorem. This gives the second bound of the theorem. (Observe that clearly $\varphi_5(\omega_1(h)) > 0$ as R is a nonzero rational function by assumption.) Also, as $\varphi_5(\nu) \to \infty$ when $\nu \to \infty$ for $m = 1, 2, 3, 4$, we can get a tighter bound, at least for large n, in that case, as follows.

By (3.7) we get that

$$\varphi_5(\nu) \ge E\nu^{5-m} \text{ for } \nu \ge D, \tag{4.22}$$

where $E > 0$. Thus $\varphi_5(\nu)$ grows without limit for $m = 1, 2, 3, 4$ when $\nu \to \infty$. Clearly, for $m \ge 5$

$$\lim_{\nu\to\infty}\varphi_5(\nu) < \infty. \tag{4.23}$$

Therefore, for $m \ge 5$, $\varphi_5(\omega_n(h))$ remains finite even for large n and so the order of magnitude n^{-2} is the tightest order we can get here. However, for $m = 1, 2, 3, 4$, we see that when $D \le \omega_n(h)$ then

$$\varphi_5(\omega_n(h)) \ge E\omega_n(h)^{5-m} = (2h^{-1})^{5-m}En^4n^{-(4m/5)}. \tag{4.24}$$

Thus the first part of the theorem now gives the last (third) bound of the theorem with the coefficient $(4/135\pi)(h/2)^m E$ in front of $n^{-(4m/5)}$. This completes the proof of the theorem. □

Next we shall give some upper bounds for the H_∞ error of the Padé-2 formula.

THEOREM 4.2. *Let $G(s) = e^{-hs}R(s)$, where $h > 0$ and $R(s) \not\equiv 0$ is a stable rational transfer function with relative degree $m \geq 1$. Let $C > 0$ be a constant such that* (4.5) *is satisfied. Furthermore, let $n \geq \max\{(2\sqrt{3})^{-1}Ch, 3\}$. Then, for $m = 1,2,3,4$ if $C < 1440^{1/5}h^{-1}n^{4/5}$,*

$$\|G - S_P^{(n)}R\|_\infty \leq 2\left(\frac{Ch}{1440^{1/5}}\right)^m \|R\|_\infty n^{-4m/5}. \tag{4.25}$$

Finally, for $m \geq 5$ if $n \geq [1440(Ch)^{m-5}/(2\sqrt{3})^m]^{1/(m-4)}$, and for $m = 1,2,3,4$ if $C \geq 1440^{1/5}h^{-1}n^{4/5}$,

$$\|G - S_P^{(n)}R\|_\infty \leq \frac{(Ch)^5}{720}\|R\|_\infty n^{-4}. \tag{4.26}$$

Hence,

$$\|G - S_P^{(n)}R\|_\infty = \max\{O(n^{-4m/5}), O(n^{-4})\}. \tag{4.27}$$

Proof. Now

$$\|G - S_P^{(n)}R\|_\infty \leq 2\max\left\{\sup_{0\leq\omega\leq\omega_n}|R(j\omega)s_n(\omega)|, \sup_{\omega>\omega_n}|R(j\omega)s_n(\omega)|\right\} \tag{4.28}$$

where

$$s_n(\omega) = \sin\left(\frac{h\omega}{2} - n\arctan\frac{(h\omega/2n)}{1-(1/3)(h\omega/2n)^2}\right), \tag{4.29}$$

and $C < \omega_n < 2\sqrt{3}h^{-1}n$ is at this point an otherwise unspecified parameter. Clearly,

$$\sup_{\omega>\omega_n}|R(j\omega)s_n(\omega)| \leq \sup_{\omega>\omega_n}|R(j\omega)| \leq \|R\|_\infty C^m\omega_n^{-m}, \tag{4.30}$$

where the last inequality follows by (4.5).

Also by (4.12)

$$\sup_{0\leq\omega\leq\omega_n}|R(j\omega)s_n(\omega)| \leq \sup_{0\leq\omega\leq\omega_n}|R(j\omega)| \times n \times \frac{1}{45}\left(\frac{\omega h}{2n}\right)^5, \tag{4.31}$$

for $\omega_n < 2\sqrt{3}h^{-1}n$. It follows by (4.5) that

$$\sup_{0\leq\omega\leq\omega_n}|R(j\omega)s_n(\omega)| \leq \frac{\|R\|_\infty C^m h^5}{1440}\sup_{C\leq\omega\leq\omega_n}\omega^{-m+5}n^{-4}. \tag{4.32}$$

Then, for $m \geq 5$,

$$\sup_{0\leq\omega\leq\omega_n}|R(j\omega)s_n(\omega)| \leq \frac{\|R\|_\infty C^m h^5}{1440}C^{-m+5}n^{-4}. \tag{4.33}$$

This gives (4.26) for $m \geq 5$ when

$$C^m\omega_n^{-m} \leq \frac{(Ch)^5}{1440}n^{-4}, \tag{4.34}$$

that is, when

$$n \geq \left(\frac{1440(Ch)^{m-5}}{(2\sqrt{3})^m} \right)^{1/(m-4)}. \tag{4.35}$$

Therefore, it remains to consider the case $m = 1, 2, 3, 4$. Then (4.31), (4.28), and (4.30) give that

$$\|G - S_P^{(n)} R\|_\infty \leq 2 \max \left\{ \frac{\|R\|_\infty C^m h^5}{1440} \omega_n^{-m+5} n^{-4}, \|R\|_\infty C^m \omega_n^{-m} \right\}. \tag{4.36}$$

To determine the smallest upper bound, consider

$$\inf_{C<\omega_n<2\sqrt{3}h^{-1}n} \max \left\{ \frac{h^5}{1440} \omega_n^{-m+5} n^{-4}, \omega_n^{-m} \right\}. \tag{4.37}$$

Note that here the first expression is a monotone increasing function of $\omega_n \geq 0$ and the second a monotone decreasing function of $\omega_n \geq 0$. These expressions attain the same value at the unique point $\omega_n^* = (1440^{1/5}/h) n^{4/5}$, which is the global minimizer of the maximum of these two expressions for $C < \omega_n < 2\sqrt{3}h^{-1}n$ if $C < \omega_n^* < 2\sqrt{3}h^{-1}n$. As the case $\omega_n^* \geq 2\sqrt{3}h^{-1}n$ cannot occur for large n, we have treated this case by the condition $n \geq 3$ in the statement of the theorem. (It is easy to check that $n \geq 3$ guarantees that $\omega_n^* < \sqrt{3}h^{-1}n$.) Hence only two cases remain. The first occurs when $\omega_n^* > C$, which gives (4.25). If $\omega_n^* \leq C$, then the infimum above is given by $[(h^5)/1440] C^{-m+5} n^{-4}$, which gives (4.26). This completes the proof of the theorem. □

The above two results mean that the Padé-2 formula achieves the tight order of magnitude $O(n^{-4m/5})$ for $m = 1, 2, 3, 4$, and $O(n^{-4})$ for $m \geq 5$, for H_∞ approximation of delay systems of the form $e^{-hs}R(s)$, where $h > 0$ and $R \neq 0$ is a stable strictly proper rational transfer function with relative degree m. These order of magnitude estimates are easier to derive with the general result from the previous section, but the above bounds have independent interest. The Padé-2 formula has a surprisingly good degree of approximation when compared with the optimal degree of rational approximation $O(n^{-m})$ (see Glover, Lam, and Partington [9]) and noting that the Padé-2 formula involves no computations to get the approximations! Some explicit (but rather complicated) bounds for the error in Padé $(2n, 2n)$ approximation are given by Lam [17]. These are asymptotically optimal, but other methods may be preferred in practice.

THEOREM 4.3. *Let $G(s) = e^{-hs}R(s)$ be as in Theorem 4.2. Let $R(s)$ have relative degree $m \geq 5$. Then*

$$\|G - S_P^{(n)} R\|_\infty \leq \frac{h^5}{720} \lim_{\nu\to\infty} \varphi_5(\nu) \times n^{-4}, \; n \geq 3. \tag{4.38}$$

Furthermore,

$$\lim_{n\to\infty} \|G - S_P^{(n)} R\|_\infty \times n^4 = \frac{h^5}{720} \lim_{\nu\to\infty} \varphi_5(\nu). \tag{4.39}$$

Proof. Clearly,

$$\|G - S_P^{(n)} R\|_\infty \leq 2 \max \left\{ \sup_{0\leq\omega_n} |R(j\omega) s_n(\omega)|, \sup_{\omega>\omega_n} |R(j\omega)| \right\} \tag{4.40}$$

where s_n is defined in (4.29) and $0 < \omega_n < 2\sqrt{3}h^{-1}n$. Clearly, by the definition of φ_5

$$2 \sup_{\omega > \omega_n} |R(j\omega)| \leq \frac{2\varphi_5(\infty)}{\omega_n^5}, \tag{4.41}$$

where $\varphi_5(\infty) = \lim_{\nu \to \infty} \varphi_5(\nu)$. Note that $\varphi_5(\infty)$ is well defined as $m \geq 5$ by assumption.

Also by (4.12)

$$\begin{aligned} 2 \sup_{0 \leq \omega \leq \omega_n} |R(j\omega)s_n(\omega)| &\leq \frac{h^5}{720} \sup_{0 \leq \omega \leq \omega_n} |R(j\omega)|\omega^5 \times n^{-4} \\ &\leq \frac{h^5}{720}\varphi_5(\infty) \times n^{-4}. \end{aligned} \tag{4.42}$$

Take $\omega_n = 2\alpha\sqrt{3}h^{-1}n$, where $0 < \alpha < 1$. Hence if n is such that

$$\frac{h^5}{720}n^{-4} > \frac{2}{(2\alpha\sqrt{3}h^{-1}n)^5}, \tag{4.43}$$

then the upper bound in the theorem follows. This gives $n > 5/(\alpha\sqrt{3})$. But as $\alpha < 1$ can be taken arbitrarily close to 1, we see that the upper bound of the theorem is indeed valid for $n \geq 3$.

To prove the asymptotically exact error formula, it suffices to show that asymptotically the upper bound of the theorem is also a lower bound for the approximation error. Recall the Taylor series expansions

$$y - \arctan \frac{y}{1 - (\frac{1}{3})y^2} = \frac{1}{45}y^5 + \cdots, \tag{4.44}$$

valid for $|y| < 2/(1 + \sqrt{7/3})$(higher order terms in y than y^5 have not been explicitly indicated above), and $\sin x = x+$ (higher order terms in x). Let $M > 0$ and $0 < \beta < 4/5$. Now clearly

$$\begin{aligned} \|G - S_P^{(n)}R\|_\infty \times n^4 &\geq 2 \sup_{0 \leq \omega \leq Mn^\beta} |R(j\omega)| \left| \sin\left(\frac{\omega h}{2} - n \arctan \frac{\frac{\omega h}{2n}}{1 - \left(\frac{1}{3}\right)\left(\frac{\omega h}{2n}\right)^2}\right)\right| \times n^4 \\ &\to \lim_{n \to \infty} 2 \sup_{0 \leq \omega \leq Mn^\beta} |R(j\omega)| \times n \times \frac{1}{45}\left(\frac{\omega h}{2n}\right)^5 \times n^4 = \frac{h^5}{720}\varphi_5(\infty), \end{aligned} \tag{4.45}$$

when $n \to \infty$. (Here we have utilized the fact that, when taking above the supremum over ω, the argument of the sine function as well as of the arctan function both tend to zero when $n \to \infty$. Hence the indicated Taylor series expansions become exact using the first term of both of the series only.) This completes the proof. □

Thus the function φ_5 appears in a fundamental manner in the H_∞ error of the Padé-2 formula for approximating delay systems of the considered form for $m \geq 5$. The asymptotically accurate formula $\|G - S_P^{(n)}R\|_\infty \approx (h^5/720)\varphi_5(\infty) \times n^{-4}$ provides an elegant way to determine what n value suffices to achieve some given approximation accuracy.

5. Numerical example.

We shall consider the linear system with a transfer function given by

$$G(s) = \frac{\exp(-hs)}{(Ts + 1)^5}, \tag{5.1}$$

TABLE 5.1
H_∞ errors (all times 10^{-3}).

Formula/ℓ	1	2	4	6	8	10
Laguerre	15.49	3.874	0.968	0.430	0.242	0.155
Kautz	-	7.747	1.937	0.861	0.484	0.310
Padé-2	-	1.389	0.087	0.017	0.005	0.002

of which one physical example is that of five perfectly mixed identical tanks with transportation delay, all coupled in series. To fix ideas we take $T = 1$ and $h = 1$.

Table 5.1 shows the H_∞ errors as given by the respective asymptotic error formulae produced by approximating the delay e^{-hs} by, respectively, the Laguerre shift formula

$$e^{-hs} \sim \frac{(1 - hs/(2n))^n}{(1 + hs/(2n))^n}, \tag{5.2}$$

the particular Kautz shift formula

$$e^{-hs} \sim \frac{(1 - hs/(2n) + h^2s^2/(8n^2))^n}{(1 + hs/(2n) + h^2s^2/(8n^2))^n}, \tag{5.3}$$

and the Padé-2 based shift formula

$$e^{-hs} \sim \frac{(1 - hs/(2n) + h^2s^2/(12n^2))^n}{(1 + hs/(2n) + h^2s^2/(12n^2))^n}. \tag{5.4}$$

Specifically, we give in Table 5.1 the values for the asymptotic H_∞ error formulas for multiplicity-ℓ Laguerre, Kautz, and Padé-2 shift formulas, so that $\ell = n$ for the Laguerre shift and $\ell = 2n$ for the others. The corresponding approximations of the delay systems have degree $5 + \ell$ as rational functions.

In this example it is easy to determine the asymptotic H_∞ error formulas for the Laguerre, Kautz, and Padé-2 shift formulas. We note that the function $\varphi_5(\nu)$ in (4.4) is now given as $\varphi_5(\nu) = \nu^5|R(j\nu)|$, $\nu \geq 0$, so that $\varphi_5(\infty) = \lim_{\nu\to\infty} \varphi_5(\nu) = 1$. Furthermore, we get that $\varphi_3(\infty) = (2/5) \times (3/5)^{3/2} \approx 0.4648$. Hence the asymptotic H_∞ error formula for the Laguerre shift formula is $\|G - S_L^{(n)}R\|_\infty \approx 15.49 \cdot 10^{-3}n^{-2}$, for the Kautz shift formula $\|G - S_K^{(n)}R\|_\infty \approx 30.99 \cdot 10^{-3}n^{-2}$ (see Mäkilä and Partington [24]), and for the Padé-2 shift formula $\|G - S_P^{(n)}R\|_\infty \approx 1.389 \cdot 10^{-3}n^{-4}$.

Note that in this example the Laguerre formula provides a fairly small H_∞ error even for $\ell = 1$, i.e., the sixth-order approximation of G so obtained might be satisfactory for many practical purposes (e.g., for robust H_∞ control design). The eleventh-order approximation provided by the Padé-2 formula for $\ell = 6$ would most likely be satisfactory for all practical purposes (e.g., even for simulating time responses of the system).

6. Conclusions. Certain shift operator induced finite-dimensional approximations of a class of delay systems have been studied. The studied approach provides the most transparent way of approximating delay systems. H_∞, H_2, and L_1 approximation error bounds have been derived for the studied class of delay systems. A special feature of the studied approximation approach is that the approximations can be written down without any computations and that their H_∞ error behavior can also be often determined reliably by simple hand computations. Hence this approach should provide an ideal technique for the purpose of treating approximate modeling and control of delay systems in control engineering courses and text books. Furthermore, it provides good low-order models for rapid low-order controller design for this important class of delay systems.

It is possible to use the studied approach for general delay systems, not just for the class of delay systems studied here. Certain generalizations of the error analysis made here are obvious; e.g., removing the restriction to a stable (strictly proper) rational part $R(s)$ in the system transfer function $G(s) = e^{-hs}R(s)$ is easy. If $R(s)$ is strictly proper with no poles on the imaginary axis, then it is clear that instead of H_∞ error bounds, one gets completely analogous L_∞ error bounds on the imaginary axis. The approximations have the same number of unstable poles as $G(s)$. Hence, L_∞ convergence of the frequency responses is enough to guarantee convergence of the studied approximations in the gap metric and in the chordal metric. Therefore the studied approach is also most useful for unstable delay systems.

The case of relative degree $m = 1$ is hard for analyzing the L_1 error behavior of the shift operator induced approximation method. This case is known to be difficult for the L_1 error analysis of several other approximation methods as well.

Acknowledgment. Financial support to one of the authors (P. M. Mäkilä) from the Academy of Finland is gratefully acknowledged.

REFERENCES

[1] H. Bercovici, *Operator Theory and Arithmetic in H^∞*, Math. Surveys Monogr. 26, American Mathematical Society, Providence, Rhode Island, 1988.

[2] N. F. Dudley Ward and J. R. Partington, *Rational wavelet decompositions of transfer functions in Hardy-Sobolev classes*, Math. Control Signals Systems, 8 (1995), pp. 257–278.

[3] N. F. Dudley Ward and J. R. Partington, *Robust identification in the disc algebra using rational wavelets and orthonormal basis functions*, Internat. J. Control, 64 (1996), pp. 409–423.

[4] N. F. Dudley Ward and J. R. Partington, *A construction of rational wavelets and frames in Hardy-Sobolev spaces with applications to system modeling*, SIAM J. Control Optim., 36 (1998), pp. 654–679.

[5] P. L. Duren, *Theory of H^p Spaces*, Academic Press, New York, 1970.

[6] C. Glader, G. Högnäs, P. M. Mäkilä, and H. Toivonen, *Approximation of delay systems—a case study*, Internat. J. Control, 53 (1991), pp. 369–390.

[7] K. Glover, R. F. Curtain, and J. R. Partington, *Realisation and approximation of linear infinite-dimensional systems with error bounds*, SIAM J. Control Optim., 26 (1988), pp. 863–898.

[8] K. Glover, J. Lam, and J. R. Partington, *Rational approximation of a class of infinite-dimensional systems* I*: Singular values of Hankel operators*, Math. Control Signals Systems, 3 (1990), pp. 325–344.

[9] K. Glover, J. Lam, and J. R. Partington, *Rational approximation of a class of infinite-dimensional systems* II: *Optimal convergence rates of L_∞ approximants*, Math. Control Signals Systems, 4 (1991), pp. 233–246.

[10] K. Glover, J. Lam, and J. R. Partington,*Rational approximation of a class of infinite dimensional systems: The L_2 case*, in Progress in Approximation Theory, P. Nevai and A. Pinkus, eds., Academic Press, New York, 1991, pp. 405–440.

[11] P. Heuberger, P. M. J. Van den Hof, and O. Bosgra, *A generalized orthonormal basis for linear dynamical systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 451–465.

[12] E. Hille and J. D. Tamarkin, *On the absolute integrability of Fourier transforms*, Fund. Math., 25 (1935), pp. 329–352.

[13] I. M. Horowitz, *Synthesis of Feedback Systems*, Academic Press, New York, 1963.

[14] D. W. Kammler, *Approximation with sums of exponentials in $L_p[0,\infty)$*, J. Approx. Theory, 16 (1976), pp. 384–408.

[15] W. H. Kautz, *Transient synthesis in the time domain*, I.R.E. Trans. Circuit Theory, CT (1) (3) (1954), pp. 29–39.

[16] F.F. Kuo, *Network Analysis and Synthesis*, 2nd ed., John Wiley & Sons, New York, 1966.

[17] J. Lam, *Convergence of a class of Padé approximations for delay systems*, Internat. J. Control, 52 (1990), pp. 989–1008.

[18] J. Lam, *Analysis on the Laguerre Formula for approximating delay systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 1517–1521.

[19] P. M. Mäkilä, *Approximation of stable systems by Laguerre filters*, Automatica J. IFAC, 26 (1990), pp. 333–345.

[20] P. M. Mäkilä, *Laguerre series approximation of infinite dimensional systems*, Automatica J. IFAC, 26 (1990), pp. 985–995.

[21] P. M. Mäkilä, *Approximation and identification of continuous-time systems*, Internat. J. Control, 52 (1990), pp. 669–687.

[22] P. M. Mäkilä, *On identification of stable systems and optimal approximation*, Automatica J. IFAC, 27 (1991), pp. 663–676.

[23] P. M. Mäkilä and J. R. Partington, *Robust approximate modelling of stable linear systems*, Internat. J. Control, 58 (1993), pp. 665–683.

[24] P. M. Mäkilä and J. R. Partington, *Laguerre and Kautz shift approximations of delay systems*, Internat. J. Control, 72 (1999), pp. 932–946.

[25] P. M. Mäkilä, J. R. Partington, and T. K. Gustafsson, *Worst-case control-relevant identification*, Automatica J. IFAC, 31 (1995), pp. 1799–1819.

[26] R. N. McDonough and W. H. Huggins, *Best least-squares representation of signals by exponentials*, IEEE Trans. Automat. Control, AC-13 (1968), pp. 408–412.

[27] B. Ninness and F. Gustafsson, *A unifying construction of orthonormal bases for system identification*, IEEE Trans. Automat. Control, 42 (1997), pp. 515–521.

[28] J. R. Partington, *Approximation of delay systems by Fourier–Laguerre series*, Automatica J. IFAC, 27 (1991), pp. 569–572.

[29] J. R. Partington, *Approximation of unstable infinite-dimensional systems using coprime factors*, Systems Control Lett., 16 (1991), pp. 89–96.

[30] J. R. Partington, *Interpolation, Identification and Sampling*, Oxford University Press, Oxford, UK, 1997.

[31] Y. C. Pati, *Wavelets and Time-Frequency Methods in Linear Systems and Neural Networks*, Ph.D. thesis, University of Maryland, College Park, MD, 1992.

[32] Y. C. Pati and P. S. Krishnaprasad, *Rational wavelets in approximation and identification of stable linear systems*, in Proceedings 31st IEEE Conf. Dec. Control, Tucson, AZ, 1992, pp. 1502–1507.

[33] P. P. Petrushev and V. A. Popov, *Rational Approximation of Real Functions*, Cambridge University Press, Cambridge, UK, 1987.

[34] M. Rosenblum and J. Rovnyak, *Hardy Classes and Operator Theory*, Oxford University Press, Oxford, UK, 1985.

[35] G. Szegő, *Orthogonal Polynomials*, American Mathematical Society, Providence, RI, 1975.

[36] B. Wahlberg and P. M. Mäkilä, *On approximation of stable linear dynamical systems using Laguerre and Kautz functions*, Automatica J. IFAC, 32 (1996), pp. 693–708.

[37] J. L. Walsh, *Interpolation and Approximation by Rational Functions in the Complex Domain*, American Mathematical Society, New York, 1935.

[38] N. Wiener, *The Theory of Prediction*, Modern Mathematics for Engineers, Bechenbach, McGraw–Hill, New York, 1956.

SIAM J. Control Optim.
Vol. 37, No. 6, pp. 1913–1945

THE VELOCITY TRACKING PROBLEM FOR NAVIER–STOKES FLOWS WITH BOUNDED DISTRIBUTED CONTROLS*

M. D. GUNZBURGER† AND S. MANSERVISI‡

Abstract. We present some systematic approaches to the mathematical analysis and numerical approximation of the time dependent optimal control problem of tracking the velocity for Navier–Stokes flows in bounded two-dimensional domains with bounded distributed controls. We study the existence of optimal solutions and derive an optimality system from which optimal solutions may be determined. We also define and analyze semidiscrete-in-time and fully space-time discrete approximations of the optimality system and a gradient method for the solution of the fully discrete system. The results of some computational experiments are provided.

Key words. optimal control, Navier–Stokes equations, fluid mechanics

AMS subject classifications. 35B40, 35B37, 35Q30, 65M60

PII. S0363012998337400

1. Introduction.

The possibility of steering a velocity field to a target velocity field over time has a wide range of applications in engineering and science. In the literature, examples are found related to combustion, chemical reacting flows, design problems, reduction of turbulence, controllability, drag reduction, etc.; see, e.g., [5], [7], [17], [27], and [31].

The controls are some physical parameters or functions that can be adjusted in practice, e.g., the velocity at the boundary or the body force in some special settings. Here, we will consider distributed or body force controls $\vec{f}$. One method for effecting such control is through a magnetic field acting on an ionized fluid or on a liquid metal. Although such controls are technologically difficult and sometimes impossible to realize in practice, they are mathematically more tractable and thus serve as a first setting for the study of optimal control problems for the nonlinear Navier–Stokes system.

A common means for effecting the velocity tracking is to monitor the quadratic functional

$$\mathcal{J}(\vec{u}(\vec{f})) = \frac{\alpha}{2}\int_0^T \int_\Omega |\vec{u}-\vec{U}|^2 \, d\vec{x}dt + \frac{\gamma}{2}\int_\Omega \left|\vec{u}(T)-\vec{U}(T)\right|^2 d\vec{x}\,, \tag{1.1}$$

where $\vec{u}$ denotes the velocity field of the flow and $\vec{U}$ the desired velocity field, i.e., the velocity field that we would like to match. Also, Ω denotes the flow domain and $(0,T)$ denotes the time interval over which the tracking is to be affected. The α and γ constants can be chosen to adjust the relative importance of the two terms appearing in (1.1). The first term in (1.1) measures, in the L^2-norm in space and time, the distance between the given target velocity $\vec{U}$ and the state velocity $\vec{u}$ over the interval $(0,T)$. The second term measures this distance, in the $L^2(\Omega)$-norm, at the time $t = T$.

*Received by the editors April 15, 1998; accepted for publication (in revised form) March 15, 1999; published electronically October 8, 1999.
http://www.siam.org/journals/sicon/37-6/33740.html

†Department of Mathematics, Iowa State University, Ames, IA 50011-2064 (gunzburg@iastate.edu). The research of this author was supported in part by the Air Force Office of Scientific Research under grant F49620-95-1-0407.

‡Mathematical Institute, University of Oxford, 24-29 St.Giles, Oxford OX1 3LB, UK (manservi@maths.ox.ac.uk).

The inclusion of the second term is found to be useful in practice since otherwise the flow field $\vec{u}$ may be driven far from the target velocity $\vec{U}$ near $t = T$.

The state of the system is described by the velocity-pressure pair $(\vec{u}, p)$, which is the solution of the Navier–Stokes system

$$\left\{\begin{array}{l} \partial_t \vec{u} + (\vec{u}\cdot\nabla)\vec{u} - \nu\Delta\vec{u} + \nabla p = \vec{f} \qquad \text{in } (0,T)\times\Omega\,, \\ \nabla\cdot\vec{u} = 0 \qquad \text{in } (0,T)\times\Omega\,, \\ \vec{u} = \vec{0} \qquad \text{on } (0,T)\times\Gamma\,, \\ \vec{u} = \vec{u}_0 \qquad \text{on } \Omega\,, \\ \displaystyle\int_\Omega p\,d\vec{x} = 0 \qquad \text{on } (0,T)\,, \end{array}\right. \tag{1.2}$$

where Γ denotes the boundary of the flow domain Ω, ∂_t the partial derivative with respect to time, and $\vec{u}_0$ a given initial velocity field. For simplicity, we have chosen homogeneous velocity boundary conditions; our results can be easily extended, with some notational complications, to other boundary conditions. In (1.2), $\vec{f}$ denotes the distributed control; thus, in our study, the distributed control is the body force per unit mass.

The size of the control is limited by technological constraints such as the amount of energy available for controlling the system. The simplest way to define such a constraint on the size of the control is to explicitly require that some norm of the control be bounded by a given constant, e.g., for all t,

$$\left(\int_\Omega |\vec{f}\,|^2\,d\vec{x}\right)^{1/2} \le K\,, \tag{1.3}$$

where the positive constant K represents, e.g., a measure of the maximum power that is available to control the system.

The velocity tracking will be effected by minimizing the cost functional (1.1) subject to the constraints (1.2) and (1.3). In this paper, we present and analyze a precise mathematical formulation of this problem and its numerical approximation for bounded, two-dimensional domains. We prove the existence of optimal controls and states and characterize such optimal solutions by deriving the first-order necessary condition associated with the problem. We formulate and analyze semidiscrete and fully discrete finite element-based approximations of the optimality system and study a gradient method for the solution of the discrete equations. We also present the results of some computational experiments. The results presented in this paper may be extended, with little difficulty, to many other objective functionals; however, the extension to boundary controls is more difficult and will be the subject of a separate paper.

Mathematical theories and approximation techniques for optimal control problems for the Navier–Stokes equations have been developed in various settings; see, e.g., [1], [10], [11], [12], [13], [14], [15], [21], [25], [26], [29], [30], and [32]. Some numerical methods for solving control problems for unsteady flows have been proposed and tested; see, e.g., [21], [24], [25], [26], [28], and [30]. Many of these deal with distributed controls and take a different approach to limiting the size of the control. Instead of imposing the explicit bound (1.3), the size of the control is limited by introducing a penalty term into the functional, e.g., instead of (1.1), a functional such as

$$\mathcal{J}(\vec{u}(\vec{f})) + \frac{\beta}{2}\int_0^T\int_\Omega |\vec{f}\,|^2\,d\vec{x}dt \tag{1.4}$$

is minimized subject to (1.2). Such an approach has been shown to be effective. However, it is obvious that the minimizer of (1.3) does not in general minimize (1.4), which is the real objective. Furthermore, when one minimizes (1.4) one does not know what the actual size of the control will be; one merely knows that by adjusting the constant β relative to α and γ, one can make more or less control available for effecting the velocity tracking; i.e., reducing β allows for more control and vice versa. However, unlike the explicit bound (1.3), one cannot ever be sure that the size of the control will indeed be within the technological constraints. For this reason, the approach taken in this paper is the more practical one; the penalty approach is easier to treat mathematically, which accounts for its popularity in the literature.

2. The optimal control problem for velocity tracking. Let $\Omega \subset \mathbb{R}^2$ be a bounded, open set. We shall use the standard notations for the Sobolev spaces (and their vector-valued, i.e., $\mathbb{R}^2$-valued, counterparts) $H^m(\Omega)$ with norm $\|\cdot\|_m$; we also use the notations $L^2(\Omega) = H^0(\Omega)$ and $\|\cdot\| = \|\cdot\|_0$. Let $H_0^m(\Omega)$ denote the closure of $C_0^\infty(\Omega)$ under the norm $\|\cdot\|_m$ and $H_0^{-m}(\Omega)$ denote the dual space of $H_0^m(\Omega)$. We introduce the solenoidal spaces

$$\begin{aligned}\mathcal{V}(\Omega) &= \{\vec{u} \in C_0^\infty(\Omega) : \nabla\cdot\vec{u} = 0\},\\ V(\Omega) &= \{\vec{u} \in H_0^1(\Omega) : \nabla\cdot\vec{u} = 0\},\\ W(\Omega) &= \{\vec{u} \in L^2(\Omega) : \nabla\cdot\vec{u} = 0\}\,.\end{aligned}$$

The spaces $V(\Omega)$ and $W(\Omega)$ are the closures of $\mathcal{V}(\Omega)$ in $L^2(\Omega)$ and $H^1(\Omega)$, respectively. The dual space of $V(\Omega)$ is denoted by $V(\Omega)^*$ while the dual space of $W(\Omega)$ can be identified with itself. Also, we define

$$L_0^2(\Omega) = \{p \in L^2(\Omega) : \int_\Omega p\, d\vec{x} = 0\}\,.$$

For details about these spaces, see, e.g., [2] and [16]. Given T, we introduce the notation $L^p((0,T); H^m(\Omega))$ for the temporal-spatial function spaces defined on $(0,T)\times\Omega$ with the norm

$$\|\vec{u}\|_{L^p((0,T);X)} = \left(\int_0^T \|\vec{u}\|_X^p\, dt\right)^{1/p}\,.$$

In order to define the weak form of the Navier–Stokes equations we introduce two continuous bilinear forms,

$$a(\vec{u},\vec{v}) = 2\nu \sum_{i,j=1}^n \int_\Omega D_{ij}(\vec{u}) D_{ij}(\vec{v})\, d\vec{x} \qquad \forall\, \vec{u},\vec{v} \in H^1(\Omega)\,, \tag{2.1}$$

$$b(\vec{v},q) = -\int_\Omega q \nabla\cdot\vec{v}\, d\vec{x} \qquad \forall q \in L^2(\Omega), \quad \forall\, \vec{v} \in H^1(\Omega), \tag{2.2}$$

where $D_{ij}(\vec{v}) = \frac{1}{2}(\partial v_i/\partial x_j + \partial v_j/\partial x_i)$ and the continuous trilinear form

$$c(\vec{w};\vec{u},\vec{v}) = \sum_{i,j=1}^n \int_\Omega w_j \left(\frac{\partial u_i}{\partial x_j}\right) v_i d\vec{x} \qquad \forall\, \vec{w},\vec{u},\vec{v} \in H^1(\Omega)\,.$$

We set $\vec{g} = \vec{f}/K$ so that the constraint (1.3) on the control can be written in the form

$$\|\vec{g}\|^2(t) = \int_\Omega |\vec{g}|^2(t)\, d\vec{x} \le 1 \qquad \text{for } t \in [0,T]\,. \tag{2.3}$$

We denote by $\mathcal{B}$ the set of $\vec{g} \in L^2((0,T);L^2(\Omega))$ that satisfies the constraint (2.3) which implies that $\|\vec{g}\|_{L^\infty((0,T);L^2(\Omega))} \leq 1$. The set $\mathcal{B}$ is a convex closed set in $L^2((0,T);L^2(\Omega))$. The constraint choice (2.3) reflects the many practical situations in which at any time one can or must, due to technological constraints, limit the total size of the control.

Other types of constraints can also be handled by the techniques used in this paper. For example, one can choose $\|\vec{g}\|_{L^2((0,T);L^2(\Omega))} \leq K$. However, constraints such as this are not very useful in practice since they allow for forces whose size, at any given time, can exceed technological limits. Also, they reflect a situation similar to that encountered when one seeks optimal controls that minimize the functional (1.4). The regularization parameter β in (1.4) can be interpreted as a Lagrange multiplier corresponding to a constraint such as $\|\vec{g}\|_{L^2((0,T);L^2(\Omega))} \leq K$ with some $K = K(\beta)$. In the context of infinite-dimensional parameter identification problems, this duality is explored by, e.g., [4] and [36]. The effect of using such constraints would be very similar to that for the case in which one optimizes with respect to the penalized functional (1.4), a case that has already been treated in [21]. For these reasons, it seems that the constraint choice (2.3) is more appropriate.

A weak formulation of the Navier–Stokes problem is defined as follows: seek $(\vec{u},p) \in L^2((0,T);H^1_0(\Omega)) \times L^2((0,T);L^2_0(\Omega))$ satisfying

$$\text{(2.4)} \qquad \begin{cases} \langle \partial_t \vec{u}, \vec{v}\rangle + a(\vec{u},\vec{v}) + c(\vec{u};\vec{u},\vec{v}) + b(\vec{u},p) = K\langle \vec{g},\vec{v}\rangle \quad \forall\, \vec{v} \in H^1_0(\Omega), \\ b(\vec{u},p) = 0 \quad \forall\, q \in L^2_0(\Omega), \\ \vec{u}(0,\cdot) = \vec{u}_0(\cdot) \in V(\Omega)\,, \end{cases}$$

where $\langle \cdot,\cdot\rangle$ denotes the duality pairing between $H^{-1}(\Omega)$ and $H^1_0(\Omega)$. The homogeneous boundary conditions for the velocity and the zero-mean condition for the pressure are satisfied due to the choice of spaces in which solutions are sought. For a simulation problem, $\vec{g} \in H^{-1}(\Omega)$ as well as $\vec{u}_0(\cdot)$ are given functions; in our optimal control setting, $\vec{g} \in L^2(\Omega)$ is the control to be determined through the optimization process.

In terms of our current notation, the functional (1.1) can be expressed as

$$\text{(2.5)} \qquad \mathcal{J}(\vec{u}(\vec{g})) = \frac{\alpha}{2}\int_0^T \|\vec{u} - \vec{U}\|^2\, dt + \frac{\gamma}{2}\|\vec{u}(T) - \vec{U}(T)\|^2\,,$$

where $\vec{U}$ is the desired target velocity. The set of all possible target velocities $L^\infty((0,T);L^2(\Omega))$ is denoted by $\mathcal{U}_{ad}$. There are no particular requirements on the target velocity $\vec{U}$ other than the fact that the functional must be bounded. The target velocity field need not be a solution of the Navier–Stokes equations. In particular, nonsolenoidal fields or fields that satisfy initial and boundary conditions different from those in (1.2) can be used as desired target velocities.

Given $K > 0$, $T > 0$, $\vec{u}_0 \in V(\Omega)$, and $\vec{U} \in \mathcal{U}_{ad}$, the set of *admissible states and controls* is defined by

$$\begin{aligned} \mathcal{A}_{ad} = \Big\{ (\vec{u}, \vec{g}) \in L^2((0,T); H_0^1(\Omega)) \times L^2((0,T); L^2(\Omega)) \\ \text{such that } \mathcal{J}(\vec{u}(\vec{g})) < \infty, \text{ (2.3) is satisfied, and} \\ \text{there exists } p \in L^2((0,T); L_0^2(\Omega)) \text{ such that (2.4) is satisfied} \Big\}. \end{aligned}$$

Actually, our notion of optimality will be a local one; i.e., we will have to be content to find optimal solutions $(\vec{u}, \vec{g})$ defined as follows:

$$\begin{aligned} &\text{given } K > 0,\ T > 0,\ \vec{u}_0 \in V(\Omega), \text{ and } \vec{U} \in \mathcal{U}_{ad}, \text{ find } (\vec{u}, \vec{g}) \in \mathcal{A}_{ad} \\ &\text{such that } \ \mathcal{J}(\vec{u}(\vec{g})) \le \mathcal{J}(\vec{w}(\vec{h})) \ \ \forall \ \ (\vec{w}, \vec{h}) \in \mathcal{A}_{ad} \text{ satisfying} \\ &\qquad \|\vec{h} - \vec{g}\|_{L^2((0,T);L^2(\Omega))} \le \epsilon \ \text{ for some } \epsilon > 0. \end{aligned} \tag{2.6}$$

2.1. Existence of the optimal solutions. In the following theorem we prove the existence of solutions for the optimal control problem. First, a few words should be said concerning known results about the unique solvability of the Navier–Stokes system. If Γ is Lipschitz continuous, $\vec{u}_0 \in W(\Omega)$, and $\vec{g} \in L^2((0,T); B(\Omega))$, where $B(\Omega)$ is the closed unit ball in $L^2(\Omega)$, then the unique solution $\vec{u}$ of (2.4) belongs to $C([0,T]; W(\Omega)) \cap L^2((0,T); V(\Omega))$ and $\vec{u}_t \in L^2((0,T); H^{-1}(\Omega))$. Furthermore, if $\Gamma \in C^2$ and $\vec{u}_0 \in V(\Omega)$, then the unique solution belongs to $C([0,T]; V(\Omega)) \cap L^2((0,T); H^2 \cap H_0^1)$ and $\vec{u}_t \in L^2((0,T); W(\Omega))$. See, e.g., [9] or [33].

THEOREM 2.1. *Given $K > 0$, $T > 0$, $\vec{u}_0 \in V(\Omega)$, and $\vec{U} \in \mathcal{U}_{ad}$, there exists a solution $\widehat{g} \in L^2((0,T); B(\Omega))$ and $\widehat{u} \in C([0,T]; W(\Omega)) \cap L^2((0,T); V(\Omega))$ of the optimal control problem* (2.6).

Proof. The admissible set $\mathcal{A}_{ad}$ is bounded and not empty, e.g., the Navier–Stokes system has a solution for $\vec{g} = \vec{0}$. Let $\{\vec{u}_n, \vec{g}_n\}$ be a minimizing sequence in $\mathcal{A}_{ad}$. Let p_n be the corresponding pressure such that $(\vec{u}_n, p_n, \vec{g}_n)$ satisfies (2.4). The sequence $\{\vec{g}_n\}$ is uniformly bounded in $L^2((0,T); L^2(\Omega))$ by T. From well-known theorems for solutions of the two-dimensional, unsteady Navier–Stokes equations (see, e.g., [33]), it follows that the corresponding sequence $\vec{u}_n$ is bounded in $C([0,T]; W(\Omega)) \cap L^2((0,T); V(\Omega))$ and the corresponding sequence p_n is bounded in $L^2((0,T); L_0^2(\Omega))$. Thus, there exist a $(\widehat{u}, \widehat{p}, \widehat{g})$ and a subsequence of $(\vec{u}_n, p_n, \vec{g}_n)$ that converges weakly to $(\widehat{u}, \widehat{p}, \widehat{g})$. We abuse the notation and write again that

$$\begin{array}{lllll} \vec{g}_n & \rightarrow & \widehat{g} & in & L^2((0,T); L^2(\Omega)) \quad \text{weakly}, \\ \vec{u}_n & \rightarrow & \widehat{u} & in & L^2((0,T); V(\Omega)) \quad \text{weakly}, \\ \vec{u}_n & \rightarrow & \widehat{u} & in & L^\infty((0,T); W(\Omega)) \quad \text{*-weakly}, \\ p_n & \rightarrow & \widehat{p} & in & L^2((0,T); L_0^2(\Omega)) \quad \text{weakly}. \end{array}$$

The sequence $\{\vec{g}_n\} \in \mathcal{B}$ converges weakly in the topology of $L^2((0,T); L^2(\Omega))$ to an element of $L^2((0,T); L^2(\Omega))$, and now we must prove that the limit $\widehat{g}$ belongs to $\mathcal{B}$. If we prove that the set $\mathcal{B}$ is weakly closed in the topology of $L^2((0,T); L^2(\Omega))$, then this implies that the sequence $\{\vec{g}_n\}$ converges in $\mathcal{B}$. We recall that every closed convex set is also weakly closed. The set $\mathcal{B}$ is convex and closed in $L^2((0,T); L^2(\Omega))$ and thus is weakly closed in this topology; then we have that $\widehat{g} \in \mathcal{B}$.

Now the pair $(\widehat{u}, \widehat{g})$ satisfies the Navier–Stokes equations (2.4) and minimizes the functional (2.5). In fact, by the lower semicontinuity of the functional (2.5) we have

$$\mathcal{J}(\vec{u}(\widehat{g})) \le \liminf_{n \to \infty} \mathcal{J}(\vec{u}_n(\vec{g}_n)) .$$

Furthermore, a priori estimates (see [9] or [33]) for $\vec{u}$ in a fractional time-order Sobolev space yields in our case that $\vec{u}_n$ converges strongly to $\vec{u} \in L^2((0,T);W(\Omega))$. Now we consider the weak Navier–Stokes system (2.4) with state $(\vec{u}_n, p_n)$ and control $\vec{g}_n$. We let $\vec{v} = \psi(t)\vec{w}$ in that system, where $\vec{w} \in \mathcal{V}(\Omega)$ and $\psi(t)$ is a continuously differentiable function on $(0,T)$ such that $\psi(T) = 0$. We then integrate the result with respect to time over the interval $(0,T)$. Then, integrating by parts with respect to time, we obtain

$$- \int_0^T \langle \vec{u}_n, \psi'(t)\vec{w} \rangle \, dt + \nu \int_0^T a(\vec{u}_n, \psi(t)\vec{w}) \, dt + \int_0^T c(\vec{u}_n; \vec{u}_n, \psi(t)\vec{w}) \, dt \\ + \int_0^T b(\psi(t)\vec{w}, p_n) \, dt = (\vec{u}_0, \psi(0)\vec{w}) + \int_0^T (\vec{g}_n, \psi(t)\vec{w}) \, dt \, .$$

We can pass to the limit inside the linear and the nonlinear terms. In fact, if $\vec{u}_n$ converges to $\widehat{u}$ in $L^2((0,T);V(\Omega))$ weakly and $L^2((0,T);W(\Omega))$ strongly, then for any $\vec{z} \in C^1((0,T);\mathcal{D}(\Omega))$, we have

$$\lim_{n\to\infty} \int_0^T c(\vec{u}_n; \vec{u}_n, \vec{z}(t)) \, dt = \int_0^T c(\widehat{u}; \widehat{u}, \vec{z}(t)) \, dt \, ;$$

see, e.g., [33]. If $\psi \in \mathcal{D}(0,T)$, then $(\widehat{u}, \widehat{p}, \widehat{g})$ satisfies the Navier–Stokes equation (2.4) in the distribution sense. Since $\mathcal{V}(\Omega)$ is dense in $V(\Omega)$, then this is still true for any $\vec{w} \in V(\Omega)$ by a continuity argument. □

We remark that this theorem implies the existence of a global optimal solution but it does not preclude the existence of other local optimal solutions. How such local optimal solutions are determined is discussed in the next section.

2.2. First-order necessary condition. In this section we derive the first-order necessary condition associated with the optimal control problem (2.6). If the Gateaux derivative of the functional exists, then the optimal solution must satisfy this standard first-order necessary condition; see, e.g., [3] or [35].

THEOREM 2.2. *Let $\vec{u}_0 \in V(\Omega)$. If $(\widehat{u}, \widehat{g})$ is an optimal pair and the functional $\mathcal{J}(\vec{u}(\vec{g}))$ is Gateaux differentiable, then the necessary condition for $\widehat{g}$ to be a minimizer of $\mathcal{J}(\vec{u}(\vec{g}))$ is*

$$\left(\frac{d\mathcal{J}(\vec{u}(\widehat{g}))}{d\vec{g}} \cdot (\vec{h} - \widehat{g}) \right) \geq 0 \quad \forall \vec{h} \in \mathcal{B} \, . \tag{2.7}$$

It is clear that the set $\mathcal{B}$ is in $L^\infty((0,T);L^2(\Omega))$; however, it is usual and convenient to work with the standard topology of $L^2((0,T);L^2(\Omega))$ in order to use well-known results. We recall that the solution of the Navier–Stokes system defines a mapping $\vec{u} = \vec{u}(\vec{g})$ from $L^2((0,T);L^2(\Omega))$ to $L^2((0,T);W(\Omega))$ which is Gateaux differentiable, and thus Theorem 2.2 can be applied.

THEOREM 2.3. *Let $\vec{u}_0 \in V(\Omega)$. The mapping $\vec{u} = \vec{u}(\vec{g})$ from $L^2((0,T);L^2(\Omega))$ to $L^2((0,T);V(\Omega))$ has a Gateaux derivative $(d\vec{u}/d\vec{g}) \cdot \vec{z}$ for every $\vec{z}$ in $L^2((0,T);L^2(\Omega))$. Furthermore, $\widetilde{w}(\vec{z}) = (d\vec{u}/d\vec{g}) \cdot \vec{z}$ is the solution of the problem*

$$\begin{cases} \langle \partial_t \widetilde{w}, \vec{v} \rangle + \nu a(\widetilde{w}, \vec{v}) + c(\vec{u}(\vec{g}); \widetilde{w}, \vec{v}) + c(\widetilde{w}, \vec{u}(\vec{g}), \vec{v}) \\ \qquad\qquad + \, b(\widetilde{w}, \widetilde{r}) = K(\vec{z}, \vec{v}) \quad \forall \, \vec{v} \in H_0^1(\Omega) \, , \\ b(\widetilde{w}, q) = 0 \quad \forall \, q \in L_0^2(\Omega) \, , \\ \widetilde{w}(t, \vec{x}) = 0 \quad \text{on } (0,t) \times \Gamma \, , \\ \widetilde{w}(0, \vec{x}) = 0 \quad \text{on } \Omega \end{cases} \tag{2.8}$$

for some $\widetilde{r} \in L^2((0,T); L^2_0(\Omega))$. *Furthermore, we have that* $\widetilde{w} \in L^\infty((0,T); W(\Omega)) \cap L^2((0,T); V(\Omega))$.

Proof. See, e.g., [1] or [21]. □

Of course, $\vec{g} + \mathcal{B} \subset L^2((0,T); L^2(\Omega))$ so that the mapping $\vec{u} = \vec{u}(\vec{g})$ is Gateaux differentiable for all $\vec{h} - \vec{g}$ with $\vec{h}$ in the admissible set $\mathcal{B}$. Note that $\widetilde{r}$ may be interpreted as the derivative $(dp/d\vec{g}) \cdot \vec{z}$ of the pressure $p(\vec{g})$.

In order to characterize optimal solutions, we introduce the adjoint problem to the linearized Navier–Stokes equations.

THEOREM 2.4. *Let* $\vec{u}_0 \in V(\Omega)$ *and let* $(\widehat{u}, \widehat{g})$ *be a solution of the optimal control problem. Let* $\vec{w} \in L^2((0,T), V(\Omega)) \cap L^\infty((0,T), L^2(\Omega))$, *and* $r \in L^2((0,T); L^2_0(\Omega))$ *denote the solution of the adjoint problem*

$$\left\{\begin{array}{l} -\langle \partial_t \vec{w}, \vec{v}\rangle + \nu a(\vec{v}, \vec{w}) + c(\widehat{u}; \vec{v}, \vec{w}) + c(\vec{v}; \widehat{u}, \vec{w}) \\ \qquad\qquad + \, b(\vec{v}, r) = \alpha(\vec{u} - \vec{U}, \vec{v}) \quad \forall\, \vec{v} \in H^1_0(\Omega)\,, \\ b(\vec{w}, q) = 0 \quad \forall\, q \in L^2_0(\Omega)\,, \\ \vec{w}(T, \cdot) = \gamma(\widehat{u}(T,\cdot) - \vec{U}(T,\cdot)) \quad \text{in } \Omega\,, \\ \vec{w} = \vec{0} \quad \text{on } (0,T) \times \Gamma\,. \end{array}\right. \tag{2.9}$$

Let $S \subset (0,T)$ *be the set in which* $\vec{w}$ *is different from zero. Then, we have*

$$\|\vec{w}\|\widehat{g} = -\vec{w} \qquad \text{for } t \in S\,. \tag{2.10}$$

On $(0,T)\backslash S$, $\widehat{u} = \vec{U}$ *and* $\widehat{g}$ *is uniquely determined up to a gradient by* $(1/K)\vec{F}$, *where* $\vec{F} = \partial_t \vec{U} + (\vec{u} \cdot \nabla)\vec{U} - \nu \Delta \vec{U}$.

Proof. Let $(\widehat{u}, \widehat{g})$ be a solution of the optimal control problem. We can compute the Gateaux derivative of the functional $\mathcal{J}(\vec{u}(\widehat{g}))$ in the direction of $\vec{z}$. We have

$$\frac{d\mathcal{J}(\vec{u}(\widehat{g}))}{d\vec{g}} \cdot \vec{z} = \alpha \int_0^T \int_\Omega (\widehat{u} - \vec{U}) \cdot \widetilde{w}\, d\vec{x}dt + \gamma \int_\Omega (\widehat{u}(T) - \vec{U}(T)) \cdot \widetilde{w}(T)\, d\vec{x},$$

where $\widetilde{w} = (d\vec{u}/d\vec{g}) \cdot \vec{z}$. Now, using (2.8) and (2.9) and then integrating the result by parts with respect to time, we obtain

$$\begin{aligned} \frac{d\mathcal{J}(\vec{u}(\widehat{g}))}{d\vec{g}} \cdot \vec{z} = \int_0^T & \Big(-\langle \partial_t \vec{w}, \widetilde{w}\rangle + \nu a(\widetilde{w}, \vec{w}) + c(\widehat{u}; \widetilde{w}, \vec{w}) + c(\widetilde{w}; \widehat{u}, \vec{w})\Big)\, dt \\ & + \int_\Omega \vec{w}\widetilde{w}\, d\vec{x}\Big|_0^T \\ = \int_0^T & \Big(\langle \partial_t \widetilde{w}, \vec{w}\rangle + \nu a(\widetilde{w}, \vec{w}) + c(\widehat{u}; \widetilde{w}, \vec{w}) + c(\widetilde{w}; \widehat{u}, \vec{w})\Big)\, dt\,. \end{aligned}$$

Hence, again using (2.8) and (2.9), we obtain

$$\frac{d\mathcal{J}(\vec{u}(\widehat{g}))}{d\vec{g}} \cdot \vec{z} = K \int_0^T \int_\Omega \vec{w} \cdot \vec{z}\, dt = K(\vec{w}, \vec{z})_{L^2((0,T);L^2(\Omega))}\,, \tag{2.11}$$

where $\vec{w}$ is the solution of the system (2.9).

In order to show (2.10), we use the first-order necessary condition (2.7) which, as a result of (2.11), is equivalent to

$$(\vec{w}, \vec{h} - \widehat{g})_{L^2((0,T);L^2(\Omega))} \geq 0 \quad \forall\, \vec{h} \in \mathcal{B}\,. \tag{2.12}$$

Let $J \subset S$ be the subset on which $\|\widehat{g}\| < 1$. Let

$$\epsilon = \frac{1}{2}\left(\int_\Omega \widehat{g}\frac{\vec{w}}{\|\vec{w}\|}\,d\vec{x} + \sqrt{\left(\int_\Omega \widehat{g}\frac{\vec{w}}{\|\vec{w}\|}\,d\vec{x}\right)^2 + 1 - \|\widehat{g}\|^2}\right) > 0\,.$$

We choose $\vec{h} = -\epsilon(\vec{w}/\|\vec{w}\|) + \widehat{g}$ on J and $\vec{h} = \widehat{g}$ elsewhere. In this case, $\|\vec{h}\| \le 1$ and $(\vec{w}, \vec{h} - \widehat{g})_{L^2((0,T);L^2(\Omega))} = -\epsilon\|\vec{w}\|_{L^2((0,T);L^2)} < 0$, which contradicts the first-order necessary condition (2.12). Thus, the set J has zero measure and $\|\widehat{g}\| = 1$ a.e. on S.

Next, again let $\widehat{g}$ be an optimal control for t almost everywhere (a.e.) in S; we have already determined that $\|\widehat{g}\| = 1$. Let $I \subset S$ be the subset on which $\widehat{g} \neq -\vec{w}/\|\vec{w}\|$. Choosing $\vec{h} = -\vec{w}/\|\vec{w}\|$ in (2.12) implies that

$$-\int_I (\vec{w}, \widehat{g})\,dt \ge \int_I \|\vec{w}\|\,dt\,. \tag{2.13}$$

By contradicting (2.13), we will show that I has zero measure so that $\|\vec{w}\|\widehat{g} = -\vec{w}$ on S.

First, we show that $\widehat{g} \neq \vec{w}/\|\vec{w}\|$ on the set I so that, together with its definition, we conclude that $\widehat{g}$ is *not* proportional to $\vec{w}$ on I. If we set $\vec{h} = 0$ in (2.12) we have that $(\vec{w}, -\widehat{g})_{L^2((0,T);L^2(\Omega))} \ge 0$ and thus $\widehat{g}$ cannot be equal to $\vec{w}/\|\vec{w}\|$. Therefore, the control $\widehat{g}$ is not proportional to $\vec{w}$ on the set I, and thus the strict Schwarz inequality $|(\vec{w}, \widehat{g})| < \|\vec{w}\|$ holds for t a.e. I.

From the strict Schwarz inequality, we then have that

$$-\int_I\int_\Omega \vec{w}\cdot\widehat{g}\,d\vec{x}dt \le \int_I |(\vec{w}, \widehat{g})|\,dt < \int_I \|\vec{w}\|\,dt\,.$$

But this contradicts (2.13) so that the set I must have zero measure. Hence, if $\vec{w} \neq 0$, we have $\|\vec{w}\|\widehat{g} = -\vec{w}$ for t a.e. in S.

On $(0,T)\backslash S$, we need only concern ourselves with sets of positive measure on which $\vec{w} = \vec{0}$, since otherwise we can define $\widehat{g}$ by continuity from neighboring times at which $\vec{w} \neq \vec{0}$. Now $\vec{w} = \vec{0}$ on a set of positive measure in $(0,T)$ if and only if $\widehat{u} = \vec{U}$ a.e. on that set. Consequently, on that set there must exist P (which is not necessarily the same as $\widehat{p}$) such that $\widehat{g} = (1/K)(\vec{F} + \nabla P)$ and $\|\widehat{g}\| \le 1$. □

Remark 2.5. Clearly, even if there exists a measurable set in $(0,T)$ such that there exists a P such that $\|\vec{g}_U\| \le 1$, where $\vec{g}_U = (1/K)(\vec{F} + \nabla P)$, we may not have that $\vec{w} = \vec{0}$ on that set. For this to be true, we also need a requirement on $(\widehat{u} - \vec{U})$ sometime on that set, e.g., $\widehat{u} = \vec{U}$ at some point in that set.

Remark 2.6. It is also clear that it is indeed possible for $\vec{w} = \vec{0}$ on measurable sets. For example, if $\vec{u}_0(\cdot) = \vec{U}(0,\cdot)$ and if there exists a P such that $\|\vec{F} + \nabla P\| \le K$, then we have that $\widehat{u} = \vec{U}$, $\vec{w} = \vec{0}$, and $\widehat{g} = (1/K)(\vec{F} + \nabla P)$ for almost all t.

Remark 2.7. Clearly, a joint rescaling of α and γ does not affect the optimal solution. Indeed, if we multiply both of these parameters by the same positive number, then the adjoint variable $\vec{w}$ is scaled by that same multiple, but, from (2.10), the optimal control $\widehat{g}$ remains unchanged as does, by (2.4), the optimal state $(\widehat{u}, \widehat{p})$. Of course, this is to be expected since multiplying the functional $\mathcal{J}(\vec{u}(\vec{g}))$ of (2.5) by a positive number cannot change its extremal points.

2.3. The optimality system. We have seen that all local optimal solutions are among the solutions of the optimality system (neglecting the possibility of $\vec{w} = \vec{0}$)

$$
\left\{
\begin{aligned}
&\langle \partial_t \widehat{u}, \vec{v}\rangle + \nu a(\widehat{u}, \vec{v}) + c(\widehat{u}; \widehat{u}, \vec{v}) + b(\vec{v}, \widehat{p}) = K(\widehat{g}, \vec{v}) \quad \forall\, v \in H_0^1(\Omega)\,,\\
&b(\widehat{u}, q) = 0 \quad \forall\, q \in L_0^2(\Omega)\\
&-\langle \partial_t \vec{w}, \vec{v}\rangle + \nu a(\vec{w}, \vec{v}) + c(\vec{w}; \widehat{u}, \vec{v}) + c(\widehat{u}; \vec{w}, \vec{v})\\
&\qquad\qquad + b(\vec{v}, r) = \alpha(\widehat{u} - \vec{U}, \vec{v}) \quad \forall\, v \in H_0^1(\Omega)\,,\\
&b(\vec{w}, q) = 0 \quad \forall\, q \in L_0^2(\Omega)\,,\\
&\|\vec{w}\| \vec{g} = -\vec{w} \quad \text{in } (0,T)\times\Omega
\end{aligned}
\right.
\tag{2.14}
$$

with initial velocity $\widehat{u}(0, \vec{x}) = \vec{u}_0(\vec{x})$, final condition $\vec{w}(T, \vec{x}) = \gamma(\widehat{u}(T) - \vec{U}(T))$ on Ω, and homogeneous boundary condition on $(0,T)\times\Gamma$ for both $\widehat{u}$ and $\vec{w}$. The above system of equations is a weak formulation of the system

$$
\left\{
\begin{aligned}
&\partial_t \widehat{u} - \nu\Delta\widehat{u} + (\widehat{u}\cdot\nabla)\widehat{u} + \nabla\widehat{p} = K\widehat{g}\,,\\
&\nabla\cdot\widehat{u} = 0\,,\\
&-\partial_t\vec{w} - \nu\Delta\vec{w} + (\nabla\widehat{u})^T\vec{w} - (\widehat{u}\cdot\nabla)\vec{w} + \nabla r = \alpha(\widehat{u} - \vec{U})\,,\\
&\nabla\cdot\vec{w} = 0\,,\\
&\|\vec{w}\|\widehat{g} = -\vec{w}
\end{aligned}
\right.
$$

in $(0,T)\times\Omega$ with the same initial, final, and boundary conditions.

REMARK 2.8. From a computational standpoint, this is a very difficult system to solve and, among the solutions of the above system, there might be solutions which do not have the property of local optimality. Therefore, how one solves this system and determines an optimal solution is a rather important question. In the following sections, we will discretize in time and then in space-time.

3. Semidiscrete-in-time approximation.

3.1. Formulation of the semidiscrete optimal control problem. Let $\sigma_N = \{t_n\}_{n=0}^N$ be a partition of $[0,T]$ into equal intervals of duration $\Delta t = T/N$ with $t_0 = 0$ and $t_N = T$. We will denote by $\mathbf{q}$ the vector $(q^{(1)}, q^{(2)}, \ldots, q^{(N)})$ of functions belonging to $\mathbf{X} = X^N$ and defined discretely with respect to time. Note that only the time levels t_n, $n = 1, 2, \ldots, N$, enter into this definition. The associated continuous, piecewise (with respect to t) linear function $q_N = q_N(t, \vec{x})$ is defined by the interpolating conditions $q_N(t_n, \vec{x}) = q^{(n)}(\vec{x})$ for $n = 1, 2, \ldots, N$. For each fixed Δt (or N) and for every function $\psi(t, \cdot)$ defined over $(0,T)$, we can associate the corresponding set $\{\psi^{(n)}(\cdot)\}_{n=1}^N$, where $\psi^{(n)}(\vec{x}) = \psi(t_n, \vec{x})$. Thus, we may define, for example, $\vec{U}_N$ to be a continuous, piecewise linear function with respect to t defined by $\vec{U}_N(t_n, \vec{x}) = \vec{U}^{(n)}(\vec{x}) = \vec{U}(t_n, \vec{x})$ for all $n = 1, 2, \ldots, N$.

The vector $\mathbf{r}$ of functions in the dual space $\mathbf{X}^*$ are defined with respect to the dual grid t_n, $n = 0, 1, \ldots, N-1$, i.e., $\mathbf{r} = (r^{(0)}, r^{(1)}, \ldots, r^{(N-1)})$. $\langle \mathbf{r}, \mathbf{p}\rangle = \sum_{n=1}^N \langle r^{(n-1)}, p^{(n)}\rangle$ denotes a duality pairing between elements of the spaces $\mathbf{X}$ and its dual $\mathbf{X}^*$ defined with respect to the discrete-time grid.

The state variables $\vec{u}^{(n)} \in H_0^1(\Omega)$ and $p^{(n)} \in L_0^2(\Omega)$ are constrained to satisfy the semidiscrete Navier–Stokes equations,

$$
\left\{
\begin{aligned}
&\frac{1}{\Delta t}(\vec{u}^{(n)} - \vec{u}^{(n-1)}, \vec{v}) + \nu a(\vec{u}^{(n)}, \vec{v}) + c(\vec{u}^{(n)}; \vec{u}^{(n)}, \vec{v})\\
&\qquad\qquad + b(\vec{v}, p^{(n)}) = K(\vec{g}^{(n)}, \vec{v}) \quad \forall\, \vec{v} \in H_0^1(\Omega), \quad \text{for } n = 1, 2, \ldots, N\,,\\
&b(\vec{u}^{(n)}, q) = 0 \quad \forall\, q \in L_0^2(\Omega), \quad \text{for } n = 1, 2, \ldots, N\,,\\
&\vec{u}^{(n)} = 0 \quad \text{on } \Gamma, \quad \text{for } n = 1, 2, \ldots, N\,,\\
&\vec{u}^{(0)} = \vec{u}_0(\vec{x}) \in V(\Omega),
\end{aligned}
\right. \tag{3.1}
$$

obtained from (1.2) by a backward Euler discretization in time.

Since we are requiring the exact control $\vec{f}$ to be bounded in norm by a real positive number K, we again require that

$$\|\vec{g}^{(n)}\| \le 1 \qquad \text{for } n = 1, 2, \ldots, N\,. \tag{3.2}$$

If we denote the unit ball of $\mathbf{L}^2(\Omega)$ by $\mathbf{B}$, then the above constraint simply implies that $\mathbf{g} \in \mathbf{B}$.

Optimization is achieved by means of the minimization of the discretized-in-time functional

$$\mathcal{J}^N(\mathbf{u}(\mathbf{g})) = \frac{\alpha \Delta t}{2} \sum_{n=1}^{N} \|\vec{u}^{(n)} - \vec{U}^{(n)}\|^2 + \frac{\gamma}{2}\|\vec{u}^{(N)} - \vec{U}^{(N)}\|^2\,. \tag{3.3}$$

This functional results from applying the right-point discretization rule in time to the continuous functional (2.5). We also introduce the admissibility set

$$
\begin{aligned}
\mathcal{A}_{ad}^N = \Big\{ &(\mathbf{u}, \mathbf{g}) \in \mathbf{H}_0^1(\Omega)) \times \mathbf{L}^2(\Omega)) \text{ such that } \mathcal{J}^N(\mathbf{u}(\mathbf{g})) < \infty, \ \mathbf{g} \in \mathbf{B},\\
&\qquad \text{and there exists } \mathbf{p} \in \mathbf{L}_0^2(\Omega)) \text{ such that (3.1) is satisfied} \Big\}.
\end{aligned}
$$

The formulation of the *semidiscrete-in-time optimal control problem* is then given by

> given $K > 0$, $T > 0$, $\vec{u}_0 \in V(\Omega)$, $\vec{U} \in \mathcal{U}_{ad}$, and N, find $(\mathbf{u}, \mathbf{g}) \in \mathcal{A}_{ad}^N$ such that the functional (3.3) is minimized.

Of course, we really seek local minimizers of the functional (3.3) in the same sense as (2.6).

The value of $\vec{g}^{(0)}$ is not defined by this formulation; it can be arbitrarily chosen. For example, one could simply set $\vec{g}^{(0)} = \vec{g}^{(1)}$ or use some other extension of the continuous, piecewise linear (with respect to time) function $\vec{g}_N$ in $C((0, T); L^2(\Omega))$. More complicated discretizations of the functional can explicitly involve the value $\vec{g}^{(0)}$. Note that this is exactly the same situation as is encountered for the pressure in the backward Euler semidiscretization of the Navier–Stokes equations, i.e., $p^{(0)}$ is not defined by (3.1).

3.2. Existence and first-order necessary condition. Now we state and prove the existence of solutions of the semidiscrete optimal control problem in an open, bounded, two-dimensional domain Ω with Lipschitz-continuous boundary Γ.

THEOREM 3.1. *Given $K > 0$, $T > 0$, $\Delta t = T/N$, $\vec{u}_0 \in V(\Omega)$, and $\vec{U} \in \mathcal{U}_{ad}$, there exists at least one solution, $(\widehat{\mathbf{u}}, \widehat{\mathbf{g}}) \in \mathbf{V}(\Omega) \times \mathbf{L}^2(\Omega)$, of the semidiscrete-in-time optimal control problem.*

Proof. Given N, let $\{\mathbf{g}_k\}_{k=1}^{\infty}$ be a minimizing sequence. Each component of the minimizing sequence is uniformly bounded by 1, and the corresponding solution $\vec{\mathbf{u}}_k$ is uniformly bounded in $\mathbf{V}(\Omega)$. Now we can proceed with a weakly convergent subsequence for each component and show that these subsequences converge to the solution of the optimal control problem in the semidiscrete approximation. Using the fact that the injection of $V(\Omega)$ into $L^2(\Omega)$ is compact, the subsequence $\mathbf{u}_k$ converges strongly. This allows us to pass to the limit in the semidiscrete Navier–Stokes equation, and the proof follows as in the continuous case. The components of $\mathbf{g}_k$ are bounded in the unit ball, and so the weak limit satisfies the bound imposed on the control. □

In the rest of this section we derive the first-order necessary condition that optimal solutions of the semidiscrete-in-time optimal control problem must satisfy. We will follow standard procedures. First, we introduce auxiliary variables that allow us to transform the inequality constraints into equalities and then invoke well-known derivations for equality constrained minimization problems; see, e.g., [3] or [35].

We begin by replacing (3.2) with

$$\|\vec{g}^{(n)}\|^2 - 1 + (s^{(n)})^2 = 0 \qquad \text{for } n = 1, 2, \ldots, N\,, \tag{3.4}$$

for some $\mathbf{s} \in \mathbb{R} = \mathbb{R}^N$. Clearly, if (3.4) is satisfied, so is (3.2). Then, the constraints in the semidiscrete-in-time optimal control problem are all equality constraints and are given by (3.1) and (3.4). Also, note that if $(\widehat{\mathbf{u}}, \widehat{\mathbf{p}}, \widehat{\mathbf{g}})$ is a solution of the semidiscrete-in-time optimal control problem, then there exists $\widehat{\mathbf{s}}$ such that $\widehat{\mathbf{g}}$ and $\widehat{\mathbf{s}}$ satisfy (3.4).

REMARK 3.2. It is worthwhile pointing out that the variables $s^{(n)}$ are introduced only as a theoretical tool and not for computational purposes. Computationally, these variables are known to be a poor tool since they may introduce many local minima.

Let $\mathbf{B}_1 = \mathbf{H}_0^1(\Omega) \times \mathbf{L}_0^2(\Omega) \times \mathbf{L}^2(\Omega) \times \mathbb{R}$ and $\mathbf{B}_2 = \mathbf{H}^{-1}(\Omega) \times \mathbf{L}_0^2(\Omega) \times \mathbb{R}$. We equip these product spaces with the usual graph norms. We define the nonlinear mapping $M : \mathbf{B}_1 \to \mathbf{B}_2$ by $M(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s}) = (\mathbf{f}, \mathbf{z}, \mathbf{l})$ for $(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s}) \in \mathbf{B}_1$ and $(\mathbf{f}, \mathbf{z}, \mathbf{l}) \in \mathbf{B}_2$ if and only if

$$\left\{\begin{array}{l} \dfrac{1}{\Delta t}(\vec{u}^{(n)} - \vec{u}^{(n-1)}, \vec{v}) + \nu a(\vec{u}^{(n)}, \vec{v}) + c(\vec{u}^{(n)}, \vec{u}^{(n)}, \vec{v}) + b(\vec{v}, p^{(n)}) \\ \qquad\qquad -K(\vec{g}^{(n)}, \vec{v}) = (\vec{f}^{(n)}, \vec{v}) \quad \forall\, \vec{v} \in H_0^1(\Omega)\,, \quad \text{for } n = 1, 2, \ldots, N\,, \\ b(\vec{u}^{(n)}, q) = (z^{(n)}, q) \quad \forall\, q \in L_0^2(\Omega)\,, \quad \text{for } n = 1, 2, \ldots, N\,, \\ \frac{1}{2}[(\vec{g}^{(n)}, \vec{g}^{(n)}) - 1 + (s^{(n)})^2] = l^{(n)}, \quad \text{for } n = 1, 2, \ldots, N\,, \\ \vec{u}^{(n)} = 0 \quad \text{on } \Gamma\,, \quad \text{for } n = 1, 2, \ldots, N\,, \\ \vec{u}^{(0)} = \vec{u}_0(\vec{x}) \in V(\Omega)\,. \end{array}\right. \tag{3.5}$$

Thus, the constraint equations (3.1) and (3.4) in the semidiscrete-in-time optimal control problem can be expressed as $M(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s}) = (\mathbf{0}, \mathbf{0}, \mathbf{0})$.

For any fixed $\widehat{\mathbf{g}} \in \mathbf{L}^2(\Omega)$ and $\mathbf{u}(\widehat{\mathbf{g}}) \in \mathbf{H}_0^1(\Omega)$, we define another nonlinear mapping $Q : \mathbf{B}_1 \to \mathbb{R} \times \mathbf{B}_2$ by $Q(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s}) = (a, \mathbf{f}, \mathbf{z}, \mathbf{l})$ if and only if

$$\begin{pmatrix} \mathcal{J}^N(\mathbf{u}(\mathbf{g})) - \mathcal{J}^N(\mathbf{u}(\widehat{\mathbf{g}})) \\ M(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s}) \end{pmatrix} = \begin{pmatrix} a \\ (\mathbf{f}, \mathbf{z}, \mathbf{l}) \end{pmatrix}. \tag{3.6}$$

These mappings are strictly differentiable, as is shown in the following lemma. We recall the notion of a strict differentiability, which is a slightly stronger concept than Frechet differentiability; see [35]. Let X and Y denote Banach spaces, and let the mapping $\varphi(x) : X \to Y$ be Frechet differentiable at x. Denote its Frechet derivative by D; of course, D is a bounded, linear mapping from X to Y. Then, $\varphi(x)$ is strictly differentiable at $x \in X$ if for any $\epsilon > 0$ there exists a $\delta > 0$ such that whenever $\|x - x_1\|_X < \delta$ and $\|x - x_2\|_X < \delta$ for $x_1, x_2 \in X$,

$$\|\varphi(x_1) - \varphi(x_2) - D(x_1 - x_2)\|_Y \le \epsilon \|x_1 - x_2\|_X \,.$$

The strict derivative D at the point $x \in X$, if it exists, will often be denoted by $D = \varphi'(x)$. The value of this mapping on an element $\widetilde{x} \in X$ will often be denoted by $\varphi'(x) \cdot \widetilde{x}$.

Lemma 3.3. *Let the nonlinear mappings $M : \mathbf{B}_1 \to \mathbf{B}_2$ and $Q : \mathbf{B}_1 \to \mathbb{R} \times \mathbf{B}_2$ be defined by* (3.5) *and* (3.6), *respectively. Then, these mappings are strictly differentiable on all of $\mathbf{B}_1$. Furthermore, the strict derivative of M at a point $(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s}) \in \mathbf{B}_1$ is given by the bounded linear operator $M'(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s}) : \mathbf{B}_1 \to \mathbf{B}_2$, where $M'(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s}) \cdot (\widetilde{\mathbf{w}}, \widetilde{\mathbf{r}}, \widetilde{\mathbf{h}}, \widetilde{\mathbf{s}}) = (\overline{\mathbf{f}}, \overline{\mathbf{z}}, \overline{\mathbf{l}})$ for $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{r}}, \widetilde{\mathbf{h}}, \widetilde{\mathbf{s}}) \in \mathbf{B}_1$ and $(\overline{\mathbf{f}}, \overline{\mathbf{z}}, \overline{\mathbf{l}}) \in \mathbf{B}_2$ if and only if*

$$(3.7) \qquad \begin{cases} \dfrac{1}{\Delta t}(\widetilde{\vec{w}}^{(n)} - \widetilde{\vec{w}}^{(n-1)}, \vec{v}) + \nu a(\widetilde{\vec{w}}^{(n)}, \vec{v}) + c(\widetilde{\vec{w}}^{(n)}, \vec{u}^{(n)}, \vec{v}) \\ \qquad\qquad + c(\vec{u}^{(n)}, \widetilde{\vec{w}}^{(n)}, \vec{v}) + b(\vec{v}, \widetilde{r}^{(n)}) - K(\widetilde{\vec{h}}^{(n)}, \vec{v}) = (\overline{\vec{f}}^{(n)}, \vec{v}) \\ \qquad\qquad\qquad\qquad \forall\, \vec{v} \in H^1_0(\Omega)\,, \quad \textit{for } n = 1, 2, \ldots, N\,, \\ b(\widetilde{\vec{w}}^{(n)}, q) = (\overline{z}^{(n)}, q) \quad \forall\, q \in L^2_0(\Omega)\,, \quad \textit{for } n = 1, 2, \ldots, N\,, \\ 2(\widetilde{\vec{h}}^{(n)}, \vec{g}^{(n)}) + 2s^{(n)}\widetilde{s}^{(n)} = \overline{l}^{(n)}, \quad \textit{for } n = 1, 2, \ldots, N\,, \\ \widetilde{\vec{w}}^{(n)} = \vec{0} \quad \textit{on } \Gamma\,, \quad \textit{for } n = 1, 2, \ldots, N\,, \\ \widetilde{\vec{w}}^{(0)} = \vec{0} \quad \textit{on } \Omega\,. \end{cases}$$

Moreover, the strict derivative of Q at a point $(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s}) \in \mathbf{B}_1$ is given by the bounded linear operator $Q'(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s}) : \mathbf{B}_1 \to \mathbb{R} \times \mathbf{B}_2$, where $Q'(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s}) \cdot (\widetilde{\mathbf{w}}, \widetilde{\mathbf{r}}, \widetilde{\mathbf{h}}, \widetilde{\mathbf{s}}) = (\overline{a}, \overline{\mathbf{f}}, \overline{\mathbf{z}}, \overline{\mathbf{l}})$ for $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{r}}, \widetilde{\mathbf{h}}, \widetilde{\mathbf{s}}) \in \mathbf{B}_1$ and $(\overline{a}, \overline{\mathbf{f}}, \overline{\mathbf{z}}, \overline{\mathbf{l}}) \in \mathbb{R} \times \mathbf{B}_2$ if and only if

$$(3.8) \qquad \begin{pmatrix} (\mathcal{J}^N)'(\mathbf{u}(\widehat{\mathbf{g}})) \cdot (\widetilde{\mathbf{w}}, \widetilde{\mathbf{r}}, \widetilde{\mathbf{h}}, \widetilde{\mathbf{s}}) \\ M'(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s}) \cdot (\widetilde{\mathbf{w}}, \widetilde{\mathbf{r}}, \widetilde{\mathbf{h}}, \widetilde{\mathbf{s}}) \end{pmatrix} = \begin{pmatrix} \overline{a} \\ (\overline{\mathbf{f}}, \overline{\mathbf{z}}, \overline{\mathbf{l}}) \end{pmatrix},$$

where

$$(\mathcal{J}^N)'(\mathbf{u}(\widehat{\mathbf{g}})) \cdot (\widetilde{\mathbf{w}}, \widetilde{\mathbf{r}}, \widetilde{\mathbf{h}}, \widetilde{\mathbf{s}}) = \alpha \Delta t \sum_{n=1}^{N} \int_\Omega (\vec{u}^{(n)} - \vec{U}^{(n)}) \cdot \widetilde{\vec{w}}^{(n)} \, d\vec{x} + \gamma \int_\Omega (\vec{u}^{(N)} - \vec{U}^{(N)}) \cdot \widetilde{\vec{w}}^{(N)} \, d\vec{x}.$$

Proof. The linearity of the operator $M'(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s})$ is obvious and its boundedness follows easily from the continuity of the forms $a(\cdot, \cdot)$, $b(\cdot, \cdot)$, and $c(\cdot, \cdot, \cdot)$. Likewise, the linearity and boundedness of the operator $Q'(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s})$ are obvious.

The fact that $M'(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s})$ is the strict derivative of the mapping $M(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s})$ follows also from the continuity of the trilinear form $c(\cdot, \cdot, \cdot)$. Indeed, we have that for any $(\mathbf{u},\mathbf{p},\mathbf{g},\mathbf{s}) \in X = \mathbf{B}_1$ and for all $(\mathbf{w},\mathbf{r},\boldsymbol{\beta}) \in \mathbf{B}_2^* = \mathbf{H}_0^1(\Omega) \times \mathbf{L}_0^2(\Omega) \times \mathbb{R}$,

$$\begin{aligned}
&\Big\langle(\mathbf{w},\mathbf{r},\boldsymbol{\beta}), M(\mathbf{u}_1,\mathbf{p}_1,\mathbf{g}_1,\mathbf{s}_1) - M(\mathbf{u}_2,\mathbf{p}_2,\mathbf{g}_2,\mathbf{s}_2)\\
&\qquad\qquad -M'(\mathbf{u},\mathbf{p},\mathbf{g},\mathbf{s}) \cdot (\mathbf{u}_1-\mathbf{u}_2,\mathbf{p}_1-\mathbf{p}_2,\mathbf{g}_1-\mathbf{g}_2,\mathbf{s}_1-\mathbf{s}_2)\Big\rangle\\
&\quad = \sum_{n=1}^{N} \beta^{(n)}\Big((g_1^{(n)},g_1^{(n)}) + (s_1^{(n)})^2 - (g_2^{(n)},g_2^{(n)}) - (s_2^{(n)})^2\\
&\qquad\qquad -2(g^{(n)},g_1^{(n)}-g_2^{(n)}) - 2s^{(n)}(s_1^{(n)}-s_2^{(n)})\Big)\\
&\qquad + \sum_{n=1}^{N}\Big(c(\vec{u}_1^{(n)},\vec{u}_1^{(n)},\vec{w}^{(n)}) - c(\vec{u}_2^{(n)},\vec{u}_2^{(n)},\vec{w}^{(n)})\\
&\qquad\qquad -c(\vec{u}_1^{(n)}-\vec{u}_2^{(n)},\vec{u}^{(n)},\vec{w}^{(n)}) - c(\vec{u}^{(n)},\vec{u}_1^{(n)}-\vec{u}_2^{(n)},\vec{w}^{(n)})\Big)\\
&\quad = -\sum_{n=1}^{N}\Big(c(\vec{u}_1^{(n)}-\vec{u}_2^{(n)},\vec{u}^{(n)}-\vec{u}_1^{(n)},\vec{w}^{(n)}) + c(\vec{u}^{(n)}-\vec{u}_2^{(n)},\vec{u}_1^{(n)}-\vec{u}_2^{(n)},\vec{w}^{(n)})\Big)\\
&\qquad - \sum_{n=1}^{N}\beta^{(n)}\Big((g_1^{(n)}-g_2^{(n)},g^{(n)}-g_1^{(n)}) + (g_1^{(n)}-g_2^{(n)},g^{(n)}-g_2^{(n)})\\
&\qquad\qquad +(s_1^{(n)}-s_2^{(n)})(s^{(n)}-s_1^{(n)}) + (s_1^{(n)}-s_2^{(n)})(s^{(n)}-s_2^{(n)})\Big),
\end{aligned}$$

so that using the continuity of the form $c(\cdot,\cdot,\cdot)$, we have, for some constants $C_1, C_2 > 0$,

$$\begin{aligned}
&\|M(\mathbf{u}_1,\mathbf{p}_1,\mathbf{g}_1,\mathbf{s}_1) - M(\mathbf{u}_2,\mathbf{p}_2,\mathbf{g}_2,\mathbf{s}_2)\\
&\qquad -M'(\mathbf{u},\mathbf{p},\mathbf{g},\mathbf{s}) \cdot (\mathbf{u}_1-\mathbf{u}_2,\mathbf{p}_1-\mathbf{p}_2,\mathbf{g}_1-\mathbf{g}_2,\mathbf{s}_1-\mathbf{s}_2)\|_{\mathbf{B}_2}\\
&\quad \le C_1\Big(\|\mathbf{u}_1-\mathbf{u}_2\|_1(\|\mathbf{u}-\mathbf{u}_1\|_1 + \|\mathbf{u}-\mathbf{u}_2\|_1)\\
&\qquad +\|\mathbf{g}_1-\mathbf{g}_2\|(\|\mathbf{g}-\mathbf{g}_1\| + \|\mathbf{g}-\mathbf{g}_2\|)\\
&\qquad +|\mathbf{s}_1-\mathbf{s}_2|(|\mathbf{s}-\mathbf{s}_1| + |\mathbf{s}-\mathbf{s}_2|)\Big)\\
&\quad \le C_2\|(\mathbf{u}_1-\mathbf{u}_2,\mathbf{p}_1-\mathbf{p}_2,\mathbf{g}_1-\mathbf{g}_2,\mathbf{s}_1-\mathbf{s}_2)\|_{\mathbf{B}_1}\\
&\qquad \Big(\|(\mathbf{u}-\mathbf{u}_1,\mathbf{p}-\mathbf{p}_1,\mathbf{g}-\mathbf{g}_1,\mathbf{s}-\mathbf{s}_1)\|_{\mathbf{B}_1}\\
&\qquad +\|(\mathbf{u}-\mathbf{u}_2,\mathbf{p}-\mathbf{p}_2,\mathbf{g}-\mathbf{g}_2,\mathbf{s}-\mathbf{s}_2)\|_{\mathbf{B}_1}\Big)\,.
\end{aligned}$$

Then, for any $\epsilon > 0$, by choosing $\delta = \epsilon/(2C_2)$ we have that whenever $\|(\mathbf{u}-\mathbf{u}_1,\mathbf{p}-\mathbf{p}_1,\mathbf{g}-\mathbf{g}_1,\mathbf{s}-\mathbf{s}_1)\|_{\mathbf{B}_1} < \delta$ and $\|(\mathbf{u}-\mathbf{u}_2,\mathbf{p}-\mathbf{p}_2,\mathbf{g}-\mathbf{g}_2,\mathbf{s}-\mathbf{s}_2)\|_{\mathbf{B}_1} < \delta$,

$$\begin{aligned}
&\|M(\mathbf{u}_1,\mathbf{p}_1,\mathbf{g}_1,\mathbf{s}_1) - M(\mathbf{u}_2,\mathbf{p}_2,\mathbf{g}_2,\mathbf{s}_2)\\
&\qquad -M'(\mathbf{u},\mathbf{p},\mathbf{g},\mathbf{s}) \cdot (\mathbf{u}_1-\mathbf{u}_2,\mathbf{p}_1-\mathbf{p}_2,\mathbf{g}_1-\mathbf{g}_2,\mathbf{s}_1-\mathbf{s}_2)\|_{\mathbf{B}_2}\\
&\quad \le \epsilon\|(\mathbf{u}_1-\mathbf{u}_2,\mathbf{p}_1-\mathbf{p}_2,\mathbf{g}_1-\mathbf{g}_2,\mathbf{s}_1-\mathbf{s}_2)\|_{\mathbf{B}_1}\,.
\end{aligned}$$

Thus, the mapping M is strictly differentiable on all of $\mathbf{B}_1$ and its strict derivative is given by M'.

Using the strict differentiability of the mapping M, it is then easy to show that the mapping Q is also strictly differentiable and that its strict derivative is given by Q'. □

Next, we prove some further properties of the derivatives of the mappings M and Q. We will be assuming that the optimal control does not vanish, i.e., that $\widehat{g}^{(n)} \neq 0$ for all $n = 1, \dots, N$. This is permissible since if $\widehat{g}^{(n)} = 0$ for some n, then it is clear that the constraint $\|\widehat{g}^{(n)}\| \le 1$ is satisfied everywhere in a neighborhood of $\widehat{g}^{(n)} \in L^2(\Omega)$. Since in any case our notion of an optimal control is a local one, see, e.g., (2.6), we can delete that constraint since it is automatically satisfied.

LEMMA 3.4. *Given $T > 0$, $K > 0$, $\Delta t = T/N$ for $N > 0$, $\vec{u}_0 \in V(\Omega)$, and $\vec{U} \in \mathcal{U}_{ad}$. Let $(\widehat{\mathbf{u}}, \widehat{\mathbf{p}}, \widehat{\mathbf{g}}, \widehat{\mathbf{s}}) \in \mathbf{B}_1$ denote a solution of the semidiscrete-in-time optimal control problem. Then, if $\widehat{g}^{(n)} \neq 0$ for all $n = 1, \dots, N$, we have*

(i) *the operator $M'(\widehat{\mathbf{u}}, \widehat{\mathbf{p}}, \widehat{\mathbf{g}}, \widehat{\mathbf{s}})$ has closed range in $\mathbf{B}_2$;*

(ii) *the operator $Q'(\widehat{\mathbf{u}}, \widehat{\mathbf{p}}, \widehat{\mathbf{g}}, \widehat{\mathbf{s}})$ has closed range in $\mathbb{R} \times \mathbf{B}_2$;*

(iii) *the operator $Q'(\widehat{\mathbf{u}}, \widehat{\mathbf{p}}, \widehat{\mathbf{g}}, \widehat{\mathbf{s}})$ is not onto $\mathbb{R} \times \mathbf{B}_2$;*

(iv) *for Δt sufficiently small the operator $M'(\widehat{\mathbf{u}}, \widehat{\mathbf{p}}, \widehat{\mathbf{g}}, \widehat{\mathbf{s}})$ is onto $\mathbf{B}_2$.*

Proof. Let the bounded linear operator $S : \mathbf{B}_1 \to \mathbf{B}_2$ be defined as follows: $S \cdot (\widetilde{\mathbf{w}}, \widetilde{\mathbf{r}}, \widetilde{\mathbf{h}}, \widetilde{\mathbf{s}}) = (\overline{\mathbf{f}}, \overline{\mathbf{z}}, \overline{\mathbf{l}})$ for $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{r}}, \widetilde{\mathbf{h}}, \widetilde{\mathbf{s}}) \in \mathbf{B}_1$ and $(\overline{\mathbf{f}}, \overline{\mathbf{z}}, \overline{\mathbf{l}}) \in \mathbf{B}_2$ if and only if

$$(3.9)\quad \begin{cases} \dfrac{1}{\Delta t}(\widetilde{w}^{(n)} - \widetilde{w}^{(n-1)}, \vec{v}) + \nu a(\widetilde{w}^{(n)}, \vec{v}) \\ \qquad\qquad + b(\vec{v}, \widetilde{r}^{(n)}) - K(\widetilde{h}^{(n)}, \vec{v}) = (\overline{f}^{(n)}, \vec{v}) \\ \qquad\qquad\qquad\qquad \forall\, \vec{v} \in H_0^1(\Omega)\,, \quad \text{for } n = 1, 2, \dots, N\,, \\ b(\widetilde{w}^{(n)}, q) = (\overline{z}^{(n)}, q) \quad \forall\, q \in L_0^2(\Omega)\,, \quad \text{for } n = 1, 2, \dots, N\,, \\ 2(\widetilde{h}^{(n)}, \widehat{g}^{(n)}) + 2\widetilde{s}^{(n)}\widehat{s}^{(n)} = \overline{l}^{(n)}, \quad \text{for } n = 1, 2, \dots, N\,, \\ \widetilde{w}^{(n)} = \vec{0} \quad \text{on } \Gamma\,, \quad \text{for } n = 1, 2, \dots, N\,, \\ \widetilde{w}^{(0)} = \vec{0} \quad \text{on } \Omega\,. \end{cases}$$

The third equation poses no difficulty. In fact, for all $n = 1, 2, \dots, N$, we may choose $\widetilde{s}^{(n)} = 0$ and $\widetilde{h}^{(n)} = \overline{l}^{(n)}\widehat{g}^{(n)}/(2\|\widehat{g}^{(n)}\|^2)$ which is permissible since, by hypothesis, $\|\widehat{g}^{(n)}\| \neq 0$. Then the third equation in (3.9) is satisfied for any choice of $\overline{l}^{(n)} \in \mathbb{R}$. Thus, the question of the closeness of the range of the operator $S : \mathbf{B}_1 \to \mathbf{B}_2$ reduces to the like question for the inhomogeneous semidiscrete-in-time Stokes operator $\widetilde{S} : \mathbf{H}_0^1(\Omega) \times \mathbf{L}_0^2(\Omega) \to \mathbf{H}^{-1}(\Omega) \times \mathbf{L}_0^2(\Omega)$ defined as follows: $\widetilde{S} \cdot (\widetilde{\mathbf{w}}, \widetilde{\mathbf{r}}) = (\widetilde{\mathbf{f}}, \overline{\mathbf{r}})$ for $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{r}}) \in \mathbf{H}_0^1(\Omega) \times \mathbf{L}_0^2(\Omega)$ and $(\widetilde{\mathbf{f}}, \overline{\mathbf{r}}) \in \mathbf{H}^{-1}(\Omega) \times \mathbf{L}_0^2(\Omega)$ if and only if

$$\begin{cases} \dfrac{1}{\Delta t}(\widetilde{w}^{(n)} - \widetilde{w}^{(n-1)}, \vec{v}) + \nu a(\widetilde{w}^{(n)}, \vec{v}) + b(\vec{v}, \widetilde{r}^{(n)}) \\ \qquad\qquad = (\widetilde{f}^{(n)}, \vec{v}) \qquad \forall\, \vec{v} \in H_0^1(\Omega)\,, \quad \text{for } n = 1, 2, \dots, N\,, \\ b(\widetilde{w}^{(n)}, q) = (\overline{z}^{(n)}, q) \quad \forall\, q \in L_0^2(\Omega)\,, \quad \text{for } n = 1, 2, \dots, N\,, \\ \widetilde{w}^{(n)} = \vec{0} \quad \text{on } \Gamma\,, \quad \text{for } n = 1, 2, \dots, N\,, \\ \widetilde{w}^{(0)} = \vec{0} \quad \text{on } \Omega\,. \end{cases}$$

The fact that the operator $\widetilde{S}$ has closed range in $\mathbf{H}^{-1}(\Omega) \times \mathbf{L}_0^2(\Omega)$ follows easily form well-known results for the semidiscrete-in-time Stokes equations; see, e.g., [33]. We can then conclude that the operator S has closed range in $\mathbf{B}_2$, and since the operator $M'(\widehat{\mathbf{u}}, \widehat{\mathbf{p}}, \widehat{\mathbf{g}}, \widehat{\mathbf{s}})$ is a compact perturbation of the operator S, we have, from the Fredholm theory, that $M'(\widehat{\mathbf{u}}, \widehat{\mathbf{p}}, \widehat{\mathbf{g}}, \widehat{\mathbf{s}})$ itself has closed range in $\mathbf{B}_2$.

Starting from (i), the proof of (ii) and (iii) can be found in [18], [19], or [20].

For (iv), we must show that (3.7), with $(\mathbf{u}, \mathbf{p}, \mathbf{g}, \mathbf{s}) = (\widehat{\mathbf{u}}, \widehat{\mathbf{p}}, \widehat{\mathbf{g}}, \widehat{\mathbf{s}})$, has a solution $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{r}}, \widetilde{\mathbf{h}}, \widetilde{\mathbf{s}}) \in \mathbf{B_1}$ for all $(\overline{\mathbf{f}}, \overline{\mathbf{z}}, \overline{\mathbf{l}}) \in \mathbf{B}_2$. Since by (i) we have that $Ran(M'(\widehat{\mathbf{u}}, \widehat{\mathbf{p}}, \widehat{\mathbf{g}}, \widehat{\mathbf{s}}))$

is closed in $\mathbf{B}_2$ and since then $Ran(M'(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}},\widehat{\mathbf{s}})) = \mathrm{Ker}(M'^*(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}},\widehat{\mathbf{s}}))^\perp$, we can instead show that $\mathrm{Ker}(M'^*(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}},\widehat{\mathbf{s}})) = \{0\}$, i.e., that $(\mathbf{w},\mathbf{r},\boldsymbol{\beta}) = (\mathbf{0},\mathbf{0},\mathbf{0})$ is the only solution of

$$(3.10)\qquad \begin{cases} -\dfrac{1}{\Delta t}(\vec{w}^{(n)} - \vec{w}^{(n-1)},\vec{v}) + \nu a(\vec{w}^{(n-1)},\vec{v}) + c(\vec{w}^{(n-1)},\widehat{u}^{(n)},\vec{v}) \\ \qquad\qquad + c(\widehat{u}^{(n)},\vec{w}^{(n-1)},\vec{v}) + b(\vec{v},r^{(n-1)}) = 0 \\ \qquad\qquad\qquad\qquad \forall\,\vec{v}\in H^1_0(\Omega)\,, \quad \text{for } n=1,2,\ldots,N\,, \\ b(\vec{w}^{(n-1)},q) = 0 \quad \forall\, q\in L^2_0(\Omega)\,, \quad \text{for } n=1,2,\ldots,N\,, \\ \beta^{(n-1)}(\widehat{g}^{(n)},h) - K(\vec{w}^{(n-1)},h) = 0 \\ \qquad\qquad\qquad\qquad \forall\, h\in L^2(\Omega)\,, \quad \text{for } n=1,2,\ldots,N\,, \\ \widehat{s}^{(n)}\beta^{(n-1)} = 0, \quad \text{for } n=1,2,\ldots,N\,, \\ \vec{w}^{(n-1)} = \vec{0} \quad \text{on } \Gamma\,, \quad \text{for } n=1,2,\ldots,N\,, \\ \vec{w}^{(N)} = \vec{0} \quad \text{on } \Omega\,. \end{cases}$$

It is well known (see, e.g., [34]) that in two dimensions, the trilinear form $c(\cdot,\cdot,\cdot)$ satisfies, for some constant $K>0$ depending only on Ω,

$$c(\vec{w},\vec{u},\vec{v}) \le K\|\vec{u}\|_1\|\vec{w}\|^{1/2}\|\vec{w}\|_1^{1/2}\|\vec{v}\|^{1/2}\|\vec{v}\|_1^{1/2} \qquad \forall\,\vec{w},\vec{u},\vec{v}\in H^1(\Omega)\,.$$

Then, from (3.10), one obtains, for $n=1,2,\ldots,N$,

$$(3.11)\quad \left(1-\frac{K^2\Delta t}{\epsilon}\|\widehat{u}^{(n)}\|_1^2\right)\|\vec{w}^{(n-1)}\|^2 - \|\vec{w}^{(n)}\|^2 + (2\nu-\epsilon)\Delta t\|\vec{w}^{(n-1)}\|_1^2 \le 0\,.$$

Let

$$G^{(n)} = 1 - \Delta t\left(\frac{K^2}{\epsilon}\|\widehat{u}^{(n)}\|_1^2\right)$$

so that (3.11) yields that

$$(3.12)\qquad G^{(n)}\|\vec{w}^{(n-1)}\|^2 + (2\nu-\epsilon)\Delta t\|\vec{w}^{(n-1)}\|_1^2 \le \|\vec{w}^{(n)}\|^2\,.$$

We choose ϵ sufficiently small so that $(2\nu-\epsilon)>0$. Now, if

$$(3.13)\qquad \Delta t \le \frac{\epsilon}{K^2\|\widehat{u}^{(n)}\|_1^2} \qquad \text{for } n=1,2,\ldots,N\,,$$

then $G^{(n)}>0$ and thus, from (3.12) and $\vec{w}^{(N)} = \vec{0}$, we have that $\|\vec{w}^{(n-1)}\| = 0$ for $n=1,2,\ldots,N$. Then, it follows that $\boldsymbol{\beta}$ and $\mathbf{r}$ must also vanish so that $\mathrm{Ker}(M'^*(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}},\widehat{\mathbf{s}}))^\perp = 0$, and thus the operator $M'(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}},\widehat{\mathbf{s}})$ is onto $\mathbf{B}_2$. □

The first-order necessary condition follows easily from the fact that the operator $Q'(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}},\widehat{\mathbf{s}})$ is not onto $\mathbb{R}\times\mathbf{B}_2$; see, e.g., [18], [19], or [20].

THEOREM 3.5. *Given $T>0$, $K>0$, $\Delta t = T/N$ for $N>0$, $\vec{u}_0\in V(\Omega)$, and $\vec{U}\in\mathcal{U}_{ad}$. If $(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}},\widehat{\mathbf{s}})\in\mathbf{B}_1$ is a solution of the semidiscrete-in-time optimal control problem, then if $\widehat{g}^{(n)}\neq 0$ for all $n=1,\ldots,N$, we have that*

(i) *there exists a nonzero Lagrange multiplier $(\lambda,\mathbf{w},\mathbf{r},\boldsymbol{\beta})\in\mathbb{R}\times\mathbf{H}^1_0(\Omega)\times\mathbf{L}^2_0(\Omega)\times\mathbb{R}$ satisfying the Euler equations*

$$(3.14)\quad \begin{aligned} &\lambda\,(\mathcal{J}^N)'(\widehat{\mathbf{u}})\cdot(\widetilde{\mathbf{w}},\widetilde{\mathbf{r}},\widetilde{\mathbf{h}},\widetilde{\mathbf{s}}) + \Big\langle(\mathbf{w},\mathbf{r},\boldsymbol{\beta}), M'(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}},\widehat{\mathbf{s}})\cdot(\widetilde{\mathbf{w}},\widetilde{\mathbf{r}},\widetilde{\mathbf{h}},\widetilde{\mathbf{s}})\Big\rangle = 0 \\ &\qquad\qquad\qquad\qquad\qquad\qquad \forall\,(\widetilde{\mathbf{w}},\widetilde{\mathbf{r}},\widetilde{\mathbf{h}},\widetilde{\mathbf{s}})\in\mathbf{B}_1\,, \end{aligned}$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing between $\mathbf{B}_2$ and $\mathbf{B}_2^*$;

(ii) *if* $(M'(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}},\widehat{\mathbf{s}}))$ *is onto* $\mathbf{B}_2$*, e.g., if* Δt *is sufficiently small so that* (3.13) *holds, then* $\lambda \neq 0$*.*

Remark 3.6. If $\lambda \neq 0$, then it is clear that we may assign an arbitrary value to it.

Remark 3.7. If $\lambda = 0$, then we have, from (3.14), that $M'^*(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}},\widehat{\mathbf{s}})\cdot(\mathbf{w},\mathbf{r},\boldsymbol{\beta}) = 0$; i.e., $0 \neq (\mathbf{w},\mathbf{r},\boldsymbol{\beta}) \in \operatorname{Ker}(M'^*(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}},\widehat{\mathbf{s}}))$.

3.3. The optimality system. Next, we examine the first-order necessary condition (3.14) to derive an optimality system from which semidiscrete-in-time optimal states and controls may be determined. We will find, as in the continuous case, that the optimal control $\widehat{\mathbf{g}}$ is proportional to the solution of an adjoint system of equations.

Theorem 3.8. *Given* $T > 0$*,* $K > 0$*,* $\Delta t = T/N$ *for* $N > 0$*,* $\vec{u}_0 \in V(\Omega)$*, and* $\vec{U} \in \mathcal{U}_{ad}$*. Let* $(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}},\widehat{\mathbf{s}}) \in \mathbf{B}_1$ *denote a solution of the semidiscrete-in-time optimal control problem. Assume that* $\widehat{g}^{(n)} \neq 0$ *for all* $n = 1,\dots,N$ *and* $\operatorname{Ker}(M'^*(\widehat{\mathbf{u}},\widehat{\mathbf{p}},\widehat{\mathbf{g}},\widehat{\mathbf{s}})) = 0$*. Then, if* $\vec{w}^{(n-1)} \neq \vec{0}$*,*

$$\|\vec{w}^{(n-1)}\|\widehat{g}^{(n)} = -\vec{w}^{(n-1)} \quad \textit{for } n = 1,2,\dots,N\,, \tag{3.15}$$

where $(\mathbf{w},\mathbf{r}) \in \mathbf{H}^1_0(\Omega) \times \mathbf{L}^2_0(\Omega)$ *satisfies the adjoint problem*

$$\begin{cases} -\dfrac{1}{\Delta t}(\vec{w}^{(n)} - \vec{w}^{(n-1)},\vec{v}) + \nu a(\vec{v},\vec{w}^{(n-1)}) + c(\vec{v},\widehat{u}^{(n)},\vec{w}^{(n-1)}) \\ \qquad + c(\widehat{u}^{(n)},\vec{v},\vec{w}^{(n-1)}) + b(\vec{v},r^{(n-1)}) = \alpha(\widehat{u}^{(n)} - \vec{U}^{(n)},\vec{v}) \\ \qquad\qquad \forall\, \vec{v} \in H^1_0(\Omega)\,, \quad \textit{for } n = 1,,\dots,N\,, \\ b(\vec{w}^{(n-1)},q) = 0 \quad \forall\, q \in L^2_0(\Omega)\,, \quad \textit{for } n = 1,\dots,N\,, \\ \vec{w}^{(n-1)} = \vec{0} \quad \text{on } \Gamma\,, \quad \textit{for } n = 1,\dots,N\,, \\ \vec{w}^{(N)} = \gamma(\vec{u}^{(N)} - \vec{U}^{(N)}) \quad \text{in } \Omega\,. \end{cases} \tag{3.16}$$

Proof. The first-order necessary condition (3.14) is equivalent to

$$\begin{aligned} &\lambda\left(\alpha\Delta t\sum_{n=1}^{N}\left((\widehat{u}^{(n)} - \vec{U}^{(n)}),\widetilde{w}^{(n)}\right) + \gamma\left((\widehat{u}^{(N)} - \vec{U}^{(N)}),\widetilde{w}^{(N)}\right)\right) \\ &\quad + \Delta t\sum_{n=1}^{N}\beta^{(n-1)}\left[\left(\widehat{g}^{(n)},\widetilde{h}^{(n)}\right) + \widehat{s}^{(n)}\widetilde{s}^{(n)}\right] + \Delta t\sum_{n=1}^{N}\Big[\frac{1}{\Delta t}(\widetilde{w}^{(n)} - \widetilde{w}^{(n-1)},\vec{w}^{(n-1)}) \\ &\qquad + \nu a(\widetilde{w}^{(n)},\vec{w}^{(n-1)}) + c(\widetilde{w}^{(n)},\widehat{u}^{(n)},\vec{w}^{(n-1)}) + c(\widehat{u}^{(n)},\widetilde{w}^{(n)},\vec{w}^{(n-1)}) \\ &\qquad + b(\vec{w}^{(n-1)},\widetilde{r}^{(n)}) - K(\widetilde{h}^{(n)},\vec{w}^{(n-1)}) + b(\widetilde{w}^{(n)},r^{(n-1)})\Big] = 0 \end{aligned}$$

for all $(\widetilde{\mathbf{w}},\widetilde{\mathbf{r}},\widetilde{\mathbf{h}},\widetilde{\mathbf{s}}) \in \mathbf{B}_1$. Since (iv) of Lemma 3.4 holds, we have, by Theorem 3.5, that $\lambda \neq 0$. We are free to choose $\lambda = -1$ (see Remark 3.6.)

Choosing $\widetilde{w}^{(n)} = 0$, $\widetilde{s}^{(n)} = 0$, and $\widetilde{h}^{(n)} = 0$ for all n in the first-order necessary condition yields, since $\widetilde{r}^{(n)}$ is arbitrary in $L^2_0(\Omega)$, the second equation in (3.16).

Choosing $\widetilde{w}^{(0)} = \vec{0}$, we have that

$$\sum_{n=1}^{N}(\widetilde{w}^{(n)} - \widetilde{w}^{(n-1)},\vec{w}^{(n-1)}) = -\sum_{n=1}^{N}(\widetilde{w}^{(n)},\vec{w}^{(n)} - \vec{w}^{(n-1)}) + (\widetilde{w}^{(N)},\vec{w}^{(N)})$$

so that choosing $\widetilde{r}^{(n)} = 0$, $\widetilde{s}^{(n)} = 0$, and $\widetilde{h}^{(n)} = 0$ for all n in the first-order necessary condition yields (recalling that $\lambda = -1$), for all $n = 1, 2, \dots, N$,

$$\begin{aligned}&\Big(-\gamma(\widehat{u}^{(N)} - \vec{U}^{(N)}) + \vec{w}^{(N)}, \widetilde{w}^{(N)}\Big)\\ &\quad -\Delta t \sum_{n=1}^{N} \Big[-\frac{1}{\Delta t}(\widetilde{w}^{(n)}, \vec{w}^{(n)} - \vec{w}^{(n-1)}) + \nu a(\widetilde{w}^{(n)}, \vec{w}^{(n-1)}) + c(\widetilde{w}^{(n)}, \widehat{u}^{(n)}, \vec{w}^{(n-1)})\\ &\quad + c(\widehat{u}^{(n)}, \widetilde{w}^{(n)}, \vec{w}^{(n-1)}) + b(\widetilde{w}^{(n)}, r^{(n-1)}) - \alpha\left((\widehat{u}^{(n)} - \vec{U}^{(n)}), \widetilde{w}^{(n)}\right)\Big] = 0\end{aligned}$$

from which the first and last equations in (3.16) follow. The third equation in (3.16) holds trivially since $\mathbf{w} \in \mathbf{H}_0^1(\Omega)$. Thus, we have shown that (3.16) holds.

Choosing $\widetilde{w}^{(n)} = 0$, $\widetilde{r}^{(n)} = 0$, and $\widetilde{h}^{(n)} = 0$ for all n in the first-order necessary condition yields, since $\widetilde{s}^{(n)}$ is arbitrary in $\mathbb{R}$,

$$\beta^{(n-1)}\widehat{s}^{(n)} = 0 \qquad \text{for } n = 1, 2, \dots, N\,. \tag{3.17}$$

Choosing $\widetilde{w}^{(n)} = 0$, $\widetilde{r}^{(n)} = 0$, and $\widetilde{s}^{(n)} = 0$ in the first-order necessary condition yields, since $\widetilde{h}^{(n)}$ is arbitrary in $\mathbf{L}^2(\Omega)$,

$$\beta^{(n-1)}\widehat{g}^{(n)} = K\vec{w}^{(n-1)} \qquad \text{for } n = 1, 2, \dots, N\,. \tag{3.18}$$

If $\vec{w}^{n-1} \neq \vec{0}$, then (3.18) requires that $\beta^{(n-1)} \neq 0$ (since $\widehat{g}^{(n)} \neq \vec{0}$) and then (3.17) yields $\widehat{s}(n) = 0$. Then the third equation in (3.5) yields that $\|\widehat{g}^{(n)}\| = 1$ and then (3.18) yields $|\beta^{(n-1)}| = K\|\vec{w}^{(n-1)}\|$, i.e., $\widehat{g}^{(n)} = \pm\,\vec{w}^{(n-1)}/\|\vec{w}^{(n-1)}\|$. The first-order necessary condition cannot determine the sign since one corresponds to a local maximum and the other to a local minimum. To determine the sign, we examine the functional $\mathcal{J}^N(\mathbf{u}(\mathbf{g}))$ in the neighborhood of an optimal point $\widehat{\mathbf{g}}$. Let $\mathbf{g} = \widehat{\mathbf{g}} + \epsilon\widetilde{\mathbf{h}}$ for $\epsilon > 0$, where $\widetilde{\mathbf{h}} \in \mathbf{L}^2(\Omega)$ is an arbitrary admissible variation. Since the admissible states and controls satisfy the discretized Navier–Stokes equations, we must have that the corresponding variations, i.e., the sensitivities, $(\widetilde{\mathbf{w}}, \widetilde{\mathbf{r}})$ satisfy

$$\begin{cases} \dfrac{1}{\Delta t}(\widetilde{w}^{(n)} - \widetilde{w}^{(n-1)}, \vec{v}) + \nu a(\widetilde{w}^{(n)}, \vec{v}) + c(\widetilde{w}^{(n)}, \widehat{u}^{(n)}, \vec{v})\\ \qquad + c(\widehat{u}^{(n)}, \widetilde{w}^{(n)}, \vec{v}) + b(\vec{v}, \widetilde{r}^{(n)}) = K(\widetilde{h}^{(n)}, \vec{v})\\ \qquad\qquad \forall\, \vec{v} \in H_0^1(\Omega)\,, \quad \text{for } n = 1, 2, \dots, N\,,\\ b(\widetilde{w}^{(n)}, q) = 0 \quad \forall\, q \in L_0^2(\Omega)\,, \quad \text{for } n = 1, 2, \dots, N\,,\\ \widetilde{w}^{(n)} = \vec{0} \quad \text{on } \Gamma\,, \quad \text{for } n = 1, 2, \dots, N\,,\\ \widetilde{w}^{(0)} = \vec{0} \quad \text{on } \Omega\,. \end{cases} \tag{3.19}$$

Note that so far we have not imposed the requirement that admissible variations $\widetilde{\mathbf{h}}$ keep $\mathbf{g} = \widehat{\mathbf{g}} + \epsilon\widetilde{\mathbf{h}}$ within the constraint set $\|\mathbf{g}\| \leq 1$. The gradient of the functional $\mathcal{J}^N(\mathbf{u}(\widehat{\mathbf{g}}))$ in the direction of $\widetilde{\mathbf{h}}$ is then given by

$$\begin{aligned}(\mathcal{J}^N)'(\mathbf{u}(\widehat{\mathbf{g}})) \cdot \widetilde{\mathbf{h}} = \alpha\Delta t \sum_{n=1}^{N} \int_\Omega (\widehat{u}^{(n)} - \vec{U}^{(n)}) \cdot \widetilde{w}^{(n)}\, d\vec{x}\\ + \gamma \int_\Omega (\widehat{u}^{(N)} - \vec{U}^{(N)}) \cdot \widetilde{w}^{(N)}\, d\vec{x}\,.\end{aligned} \tag{3.20}$$

Combining (3.16), (3.19), and (3.20) yields

$$(\mathcal{J}^N)'(\mathbf{u}(\widehat{\mathbf{g}})) \cdot \widetilde{\mathbf{h}} = K\langle \mathbf{w}, \widetilde{\mathbf{h}} \rangle \,. \tag{3.21}$$

Now, choose $\widetilde{h}^{(j)} = -\widehat{g}^{(j)}$ for some j such that $\vec{w}^{(j-1)} \neq \vec{0}$ and $\widetilde{h}^{(n)} = \vec{0}$ for $n \neq j$ so that, since $\|\widehat{g}^{(j)}\| = 1$, for ϵ small enough we have that if $\vec{g}^{(j)} = \widehat{g}^{(j)} + \epsilon \widetilde{h}^{(j)}$, then $\|\vec{g}^{(j)}\| \leq 1$ so that $\vec{g}^{(j)}$ is admissible. (Note that $\widetilde{h}^{(j)} = \widehat{g}^{(j)}$ would take us out of the constraint set so that we need not consider that case here.) With $\widetilde{h}^{(j)} = -\widehat{g}^{(j)}$ and $\widetilde{h}^{(n)} = \vec{0}$ otherwise, we have from (3.18) and (3.21), that

$$\begin{aligned}(\mathcal{J}^N)'(\mathbf{u}(\widehat{\mathbf{g}}) \cdot \widetilde{\mathbf{h}} = K\langle \mathbf{w}, \widetilde{\mathbf{h}} \rangle = K \sum_{n=1}^{N} (\vec{w}^{(n-1)}, \widetilde{h}^{(n)}) \\ = -K(\vec{w}^{(j-1)}, \widehat{g}^{(j)}) = -\beta^{(j-1))} \|\widehat{g}^{(j)}\|^2 = -\beta^{(j-1)} \,.\end{aligned}$$

But, since we are entering the admissibility set from the optimal solution $\widehat{g}^{(j)}$, we must have that $(\mathcal{J}^N)'(\mathbf{u}(\widehat{\mathbf{g}})) \cdot \widetilde{\mathbf{h}} \geq 0$ for this choice of $\widetilde{\mathbf{h}}$ so that $\beta^{(j-1)} < 0$ and therefore $\widehat{g}^{(j)} = -\vec{w}^{(j-1)}/\|\vec{w}^{(j-1)}\|$. Since j is arbitrary such that $\vec{w}^{(j-1)} \neq 0$, we have that (3.15) holds. □

REMARK 3.9. If $\vec{w}^{(n-1)} = \vec{0}$ and $\vec{w}^{(n)} = \vec{0}$ for some n, then $\widehat{u}^{(n)} = \vec{U}^{(n)}$ and we can determine $\widehat{g}^{(n)}$ from the semidiscrete Navier–Stokes equations, i.e., there must exist a $P^{(n)}$ such that $(\vec{U}^{(n)}, P^{(n)})$ satisfy (3.1) for that particular value of n. If $\vec{w}^{(n-1)} = \vec{0}$ and $\vec{w}^{(n)} \neq \vec{0}$, then $\widehat{u}^{(n)} \neq \vec{U}^{(n)}$ and we can extend the control from the right and set $\widehat{g}^{(n)} = \widehat{g}^{(n+1)}$ or, if $n = N$, $\widehat{g}^{(N)} = -\vec{w}^{(N)}/\|\vec{w}^{(N)}\|$.

REMARK 3.10. Again, a joint rescaling of α and γ does not affect the solution of the semidiscrete-in-time optimal control problem.

We have seen that one may obtain a solution of the semidiscrete-in-time optimal control problem by solving the coupled system (if $\vec{w}^{(n-1)} \neq \vec{0}$ for $n = 1, 2, \ldots, N$)

$$\left\{ \begin{aligned}
&\frac{1}{\Delta t}(\widehat{u}^{(n)} - \widehat{u}^{(n-1)}, \vec{v}) + \nu a(\widehat{u}^{(n)}, \vec{v}) + c(\widehat{u}^{(n)}, \widehat{u}^{(n)}, \vec{v}) + b(\vec{v}, \widehat{p}^{(n)}) \\
&\qquad\qquad = K(\widehat{g}^{(n)}, \vec{v}) \quad \forall\, \vec{v} \in H_0^1(\Omega)\,, \qquad \text{for } n = 1, 2, \ldots, N\,, \\
&b(\widehat{u}^{(n)}, q) = 0 \quad \forall\, q \in L_0^2(\Omega)\,, \qquad \text{for } n = 1, 2, \ldots, N\,, \\
&\widehat{u}^{(n)} = 0 \quad \text{on } \Gamma\,, \qquad \text{for } n = 1, 2, \ldots, N\,, \\
&\widehat{u}^{(0)} = \widehat{u}_0(\vec{x}) \in V(\Omega)\,, \\
&\|\vec{w}^{(n-1)}\| \widehat{g}^{(n)} = -\vec{w}^{(n-1)} \quad \text{in } \Omega\,, \quad \text{for } n = 1, 2, \ldots, N\,, \\
&-\frac{1}{\Delta t}(\vec{w}^{(n)} - \vec{w}^{(n-1)}, \vec{v}) + \nu a(\vec{v}, \vec{w}^{(n-1)}) + c(\vec{v}, \widehat{u}^{(n)}, \vec{w}^{(n-1)}) \\
&\qquad + c(\widehat{u}^{(n)}, \vec{v}, \vec{w}^{(n-1)}) + b(\vec{v}, r^{(n-1)}) = \alpha(\widehat{u}^{(n)} - \vec{U}^{(n)}, \vec{v}) \\
&\qquad\qquad\qquad \forall\, \vec{v} \in H_0^1(\Omega)\,, \quad \text{for } n = 1, \ldots, N\,, \\
&b(\vec{w}^{(n-1)}, q) = 0 \quad \forall\, q \in L_0^2(\Omega)\,, \quad \text{for } n = 1, \ldots, N\,, \\
&\vec{w}^{(n-1)} = \vec{0} \quad \text{on } \Gamma\,, \quad \text{for } n = 1, \ldots, N\,, \\
&\vec{w}^{(N)} = \gamma(\widehat{u}^{(N)} - \vec{U}^{(N)}) \quad \text{in } \Omega\,.
\end{aligned} \right. \tag{3.22}$$

The above system of equations is a weak formulation of the system

$$\begin{cases} \dfrac{1}{\Delta t}(\widehat{u}^{(n)} - \widehat{u}^{(n-1)}) + (\widehat{u}^{(n)} \cdot \nabla)\widehat{u}^{(n)} - \nu\Delta\widehat{u}^{(n)} + \nabla\widehat{p}^{(n)} \\ \qquad\qquad\qquad = K\widehat{g}^{(n)} \quad \text{in } \Omega\,, \quad \text{for } n = 1, 2, \dots, N\,, \\ \nabla \cdot \widehat{u} = 0 \quad \text{in } \Omega\,, \quad \text{for } n = 1, 2, \dots, N\,, \\ \widehat{u}^{(n)} = 0 \quad \text{on } \Gamma\,, \quad \text{for } n = 1, 2, \dots, N\,, \\ \widehat{u}^{(0)} = \widehat{u}_0 \quad \text{in } \Omega\,, \\ \|\vec{w}^{(n-1)}\|\widehat{g}^{(n)} = -\vec{w}^{(n-1)} \quad \text{in } \Omega\,, \quad \text{for } n = 1, 2, \dots, N\,, \\ -\dfrac{1}{\Delta t}(\vec{w}^{(n)} - \vec{w}^{(n-1)}) - \nu\Delta\vec{w}^{(n-1)} + (\nabla\widehat{u}^{(n)})^T\vec{w}^{(n-1)} - (\widehat{u}^{(n)} \cdot \nabla)\vec{w}^{(n-1)} \\ \qquad\qquad +\nabla r^{(n-1)} = \alpha(\widehat{u}^{(n)} - \vec{U}^{(n)}) \quad \text{in } \Omega\,, \quad \text{for } n = 1, 2, \dots, N\,, \\ \nabla \cdot \vec{w}^{(n-1)} = 0 \quad \text{in } \Omega\,, \quad \text{for } n = 1, 2, \dots, N\,, \\ \vec{w}^{(n-1)} = \vec{0} \quad \text{on } \Gamma\,, \quad \text{for } n = 1, \dots, N\,, \\ \vec{w}^{(N)} = \gamma(\widehat{u}^{(N)} - \vec{U}^{(N)}) \quad \text{in } \Omega\,. \end{cases}$$

Remark 3.11. If $\lambda = 0$, then (3.15) still holds where $\vec{\mathbf{w}}$ is any nontrivial solution of the homogeneous system

$$\begin{cases} -\dfrac{1}{\Delta t}(\vec{w}^{(n)} - \vec{w}^{(n-1)}, \vec{v}) + \nu a(\vec{v}, \vec{w}^{(n-1)}) + c(\vec{v}, \widehat{u}^{(n)}, \vec{w}^{(n-1)}) \\ \qquad\qquad + c(\widehat{u}^{(n)}, \vec{v}, \vec{w}^{(n-1)}) + b(\vec{v}, r^{(n-1)}) = \vec{0} \\ \qquad\qquad\qquad\qquad \forall\, \vec{v} \in H_0^1(\Omega)\,, \quad \text{for } n = 1,, \dots, N\,, \\ b(\vec{w}^{(n-1)}, q) = 0 \quad \forall\, q \in L_0^2(\Omega)\,, \quad \text{for } n = 1, \dots, N\,, \\ \vec{w}^{(n-1)} = \vec{0} \quad \text{on } \Gamma\,, \quad \text{for } n = 1, \dots, N\,, \\ \vec{w}^{(N)} = \vec{0} \quad \text{in } \Omega\,; \end{cases}$$

see Remark (3.7).

We now examine the convergence of solutions of the semidiscrete-in-time optimal control system.

Theorem 3.12. *Given $K > 0$, $T > 0$, $\Delta t = T/N$, $\vec{u}_0 \in V(\Omega)$, and $\vec{U} \in \mathcal{U}_{ad}$. Then, for $\Delta T \to 0$, i.e., $N \to \infty$, the solution $\{(\widehat{u}^{(n)}, \widehat{g}^{(n)})\}_{n=1}^N$ of the system* (3.22) *converges to the solution $(\widehat{u}, \widehat{g})$ of the system* (2.14).

Proof. Let $\Delta t = T/N$, and let $\vec{u}\,'^N$ denote the piecewise linear function such that $\vec{u}\,'^{(n)} = (\vec{u}^{(n)} - \vec{u}^{(n-1)})/\Delta t$ for $n = 1, 2, \dots, N$. From the well-known theory on semidiscrete-in-time approximations of the Navier–Stokes equations (see, e.g., [33]), the sequences $\{\widehat{u}^N\}_{N=1}^\infty$, $\{\widehat{g}^N\}_{N=1}^\infty$, and $\{\widehat{u}\,'^N\}_{N=1}^\infty$ are uniformly bounded in $L^2((0,T); V(\Omega)) \cap L^\infty((0,T); W(\Omega))$, $\mathcal{B}$, and $L^2((0,T); V^*(\Omega))$, respectively. Then we can extract subsequences, which we again denote by $\{\widehat{u}^N\}$ and $\{\widehat{g}^N\}$ such that

$$\begin{cases} \widehat{u}^N \to \vec{u} & L^2(0, T, V(\Omega)) \text{ weakly}\,, \\ \widehat{u}^N \to \vec{u} & L^\infty(0, T, H(\Omega)) \text{ *-weakly}\,, \\ \widehat{g}^N \to \vec{g} & L^2(0, T, B(\Omega)) \text{ weakly}\,, \\ \widehat{u}\,'^N \to \vec{u}\,' & \quad L^2(0, T, V^*(\Omega)) \text{ weakly}\,. \end{cases}$$

The set $\mathcal{B}$ is a convex, closed set and thus weakly closed so that the limit $\vec{g}$ satisfies the bound $\|\vec{g}\| \leq 1$ a.e on $(0,T)$. Let $\mathcal{H}((0,T);n,m;V,V^*) = \{\vec{v} \in L^n((0,T);V) : \vec{v}\,' \in L^m((0,T);V^*)\}$. We have that $V(\Omega) \subset W(\Omega) \subset V^*(\Omega)$, where V and V^* are reflexive, the injections are continuous, and, from the Sobolev imbedding theorem, the imbedding $V(\Omega) \to W(\Omega)$ is compact. Then, the injection from $\mathcal{H}((0,T);2,2,V,V^*)$ into $L^2((0,T);W)$ is compact. We note that if the sequence $\vec{u}_k$ converges weakly in $L^2((0,T);V(\Omega))$ and the sequence $\vec{u}\,'_k$ converges weakly in $L^2((0,T);V^*(\Omega))$, then the sequence $\vec{u}_k$ converges strongly in $L^2((0,T);W(\Omega))$. (A proof of this result can be found in, e.g., [33].) As a consequence, the sequence $\widehat{u}^N$ converges strongly in $L^2((0,T);W(\Omega))$. Now we can pass to the limit in the system of equations and in the functional using standard techniques. The fact that the sequence $\widehat{u}^N$ converges weakly in $L^2((0,T);V(\Omega))$ and strongly in $L^2((0,T);W(\Omega))$ allows us to pass to the limit in the nonlinear term and thus show that the limit $(\vec{u},\vec{g})$ is a solution of the continuous optimal control problem. Thus, as $N \to \infty$, the solution of the semidiscrete optimal control problem converges to the corresponding solution of the continuous optimal control problem. □

4. Fully discrete time-space approximation.

4.1. Preliminaries. We consider only conforming finite element approximations. Let $X^h \subset H^1(\Omega)$ and $S^h \subset L^2(\Omega)$ be two families of finite dimensional subspaces parameterized by h that tends to zero. We also denote $X_0^h = X^h \cap H_0^1(\Omega)$, $S_0^h = S^h \cap L_0^2(\Omega)$, and $S_B^h = S^h \cap B(\Omega)$. We make the following assumptions on X^h and S^h.

Approximation hypotheses: there exist an integer l and a constant C, independent of h, $\vec{u}$, and p, such that for $1 \leq k \leq l$ we have

$$\inf_{\vec{u}_h \in X_0^h} \|\vec{u}_h - \vec{u}\|_1 \leq C h^k \|\vec{u}\|_{k+1} \quad \forall \vec{u} \in H^{k+1}(\Omega) \cap H_0^1(\Omega)\,, \tag{4.1}$$

$$\inf_{p_h \in S_0^h} \|p - p_h\| \leq C h^k \|p\|_k \qquad \forall p \in H^k(\Omega) \cap L_0^2(\Omega)\,. \tag{4.2}$$

The inf-sup condition or LBB condition: there exists a constant C', independent of h, such that

$$\inf_{0 \neq q_h \in S_0^h} \; \sup_{0 \neq \vec{u}_h \in X_0^h} \frac{\int_\Omega q_h \nabla \cdot \vec{u}_h}{\|\vec{u}_h\|_1 \|q_h\|} \geq C' > 0\,. \tag{4.3}$$

This condition assures the stability of the Navier–Stokes discrete solutions.

To preserve the antisymmetry of the trilinear form $c(\vec{u};\vec{v},\vec{w})$ on the finite element spaces we introduce the modified trilinear form (see [33])

$$\widetilde{c}(\vec{u};\vec{v},\vec{w}) = \frac{1}{2}\{c(\vec{u};\vec{v},\vec{w}) - c(\vec{u};\vec{w},\vec{v})\} \qquad \forall \vec{u},\vec{v},\vec{w} \in H_0^1(\Omega).$$

We can recall some useful formulas and inequalities that hold in a two-dimensional domain Ω :

$$\left.\begin{aligned} &\widetilde{c}(\vec{u};\vec{v},\vec{w}) = -\widetilde{c}(\vec{u};\vec{w},\vec{v})\,, \\ &\widetilde{c}(\vec{u};\vec{v},\vec{v}) = 0 \end{aligned}\right\} \quad \forall\, \vec{u};\vec{v},\vec{w} \in H^1(\Omega)$$

and (see, e.g., [33])

$$\left.\begin{aligned} &|\widetilde{c}(\vec{u};\vec{v},\vec{w})| \leq K_1 \|\nabla \vec{u}\| \cdot \|\vec{v}\|_{L^4(\Omega)} \|\nabla \vec{w}^{(n)}\|\,, \\ &|\widetilde{c}(\vec{u};\vec{v},\vec{w})| \leq K_2 \|\vec{u}\|^{\frac{1}{2}} \|\nabla \vec{u}\|^{\frac{1}{2}} \|\nabla \vec{v}\| \cdot \|\vec{w}\|^{\frac{1}{2}} \|\nabla \vec{w}\|^{\frac{1}{2}} \end{aligned}\right\} \quad \forall\, \vec{u};\vec{v},\vec{w} \in H^1(\Omega)\,.$$

We remark that the last inequality is true in the framework of the conforming finite element approximation and only in the two-dimensional case; see [33].

4.2. Formulation of the fully discrete optimal control approximation. Let $\sigma_N = \{t_n\}_{n=0}^N$ be a partition of $[0, T]$ into equal intervals $\Delta t = T/N$ with $t_0 = 0$ and $t_N = T$. On the finite element spaces $X_0^h \subset H_0^1(\Omega)$ and $S_0^h \subset L_0^2(\Omega)$ we assume that the hypotheses (4.1)–(4.3) hold. For each fixed Δt (or N) and for every involved quantity $q(t, \vec{x})$ we associate the corresponding set $\{q_h^{(n)}\}_{n=1}^N$ and a continuous linear function $\vec{q}_h^N(t, \vec{x})$ such that $\vec{q}_h^N(t_n, \vec{x}) = q_h(t_n, \vec{x})$ for all $n = 1, 2, \ldots, N$. We denote by $\mathbf{q}$ the vector $(q^{(1)}, q^{(2)}, \ldots, q^{(N)})$ of the discrete-time components. We assume that the control $\vec{f}_h^{(n)}$ is bounded by K so that $\|\vec{g}_h^{(n)}\| \leq 1$ for all $n = 1, 2, \ldots, N$.

Given $\Delta t = T/N$, $\{\vec{g}_h^{(n)}\}_{n=1}^N \in \mathbf{B}(\Omega)$, and $\vec{u}_0 \in V(\Omega)$, then $(\vec{\mathbf{u}}_h, \mathbf{p}_h) \in (\mathbf{X}_0^h \times \mathbf{S}_0^h)$ is said to be generalized solution for the fully discrete time-space approximation of the Navier–Stokes equations if $(u_h^{(n)}, p_h^{(n)})$ satisfies

$$(4.4)\qquad \begin{cases} \dfrac{1}{\Delta t}(\vec{u}_h^{(n)} - \vec{u}_h^{(n-1)}, \vec{v}_h) + \nu a(\vec{u}_h^{(n)}, \vec{v}_h) + \widetilde{c}(\vec{u}_h^{(n)}; \vec{u}_h^{(n)}, \vec{v}_h) \\ \qquad\qquad + \, b(\vec{v}_h, p^{(n)}) = K(\vec{g}_h^{(n)}, \vec{v}_h) \quad \forall \vec{v}_h \in X_0^h(\Omega)\,, \\ b(\vec{v}_h^{(n)}, q_h) = 0 \quad \forall q_h \in S_0^h(\Omega) \end{cases}$$

for $n = 1, 2, \ldots, N$ with initial velocity $\vec{u}_h^{(0)} = \Pi^h \vec{u}_0(\vec{x})$, where Π^h denotes the projection of the initial data $\vec{u}_0$ onto X^h.

The fully discretized functional is given by

$$(4.5)\qquad \mathcal{J}_h^N(\mathbf{u}_h(\mathbf{g})) = \frac{\alpha}{2}\Delta t \sum_{n=1}^N \|\vec{u}_h^{(n)} - \vec{U}^{(n)}\|^2 + \frac{\gamma}{2}\|\vec{u}_h^{(N)} - \vec{U}^{(N)}\|^2 \,.$$

The formulation of the optimal problem in the fully discrete approximation is then given by

> *given $\Delta t = T/N$, $\vec{u}_0 \in V(\Omega)$, and $\vec{U} \in U_{ad}$, find $(\mathbf{u}_h, \mathbf{p}_h, \mathbf{g}_h)$, a sequence in $(\mathbf{X}_0^h(\Omega) \times \mathbf{S}_0^h(\Omega) \times \mathbf{S}_B^h(\Omega))$, such that $(\mathbf{u}_h, \mathbf{p}_h)$ is the solution of* (4.4) *and minimizes the cost function in* (4.5).

The existence and convergence of the solution of the fully discrete optimal control problem can be proved in the same way as for semidiscrete-in-time case if we limit our analysis to conforming finite element methods. We can state that the fully discrete optimal solution $(\widehat{\mathbf{u}}_h, \widehat{g}_h)$ converges to the optimal solution $(\widehat{u}, \widehat{g})$ of the continuous problem as $\Delta t \to 0$, i.e., as $N \to \infty$ and $h \to 0$. The well-known corresponding theorems on fully discrete approximations of the Navier–Stokes equations can be found in, e.g., [33]. The optimal necessary condition can be found using the same techniques that were used for the semidiscrete-in-time case. Finally for completeness we can state the theorem that gives the control as a solution of the adjoint equation

THEOREM 4.1. *Let $\vec{u}_0 \in V(\Omega)$, $\vec{U} \in \mathcal{U}_{ad}$, and $\Delta t = T/N$. If $(\widehat{\mathbf{u}}_h, \widehat{\mathbf{g}}_h)$ is a solution of the fully discrete optimal control problem, then, if Δt is sufficiently small and if $\vec{w}_h^{(n-1)} \neq 0$ or all $n = 1, 2, \ldots, N$, we have that $\|\vec{w}_h^{(n-1)}\|\vec{g}_h^{(n)} = -\vec{w}_h^{(n-1)}$ for all $n = 1, 2, \ldots, N$, where, for $n = 1, 2 \ldots, N$, the functions $\vec{w}_h^{(n-1)} \in X_0^h$ are the solution* *of the discrete adjoint problem*

$$(4.6)\qquad \begin{cases} -\dfrac{1}{\Delta t}(\vec{w}_h^{(n)} - \vec{w}_h^{(n-1)}, \vec{v}_h) + \nu a(\vec{v}_h, \vec{w}_h^{(n-1)}) + \widetilde{c}(\vec{v}_h; \widehat{u}_h^{(n)}, \vec{w}_h^{(n-1)}) \\ \qquad\qquad + \ \widetilde{c}(\widehat{u}_h^{(n)}; \vec{v}_h, \vec{w}_h^{(n-1)}) + b(\vec{v}_h, r_h^{(n-1)}) \\ \qquad\qquad\qquad = \alpha(\widehat{u}_h^{(n)} - \vec{U}^{(n)}, \vec{v}_h) \quad \forall \vec{v}_h \in X_0^h(\Omega) \\ b(\vec{w}_h^{(n-1)}, q_h) = 0 \quad \forall q_h \in S_0^h(\Omega) \end{cases}$$

for $n = 1, 2, \ldots, N$ along with the terminal condition $\vec{w}_h^{(N)} = \gamma(\widehat{u}_h^{(N)} - \vec{U}^{(N)})$.

5. A projected gradient method. An approximation to the optimal control, state, and adjoint variable may be obtained by solving the fully discrete optimality system which consists of the following (after dropping the $\widehat{(\cdot)}$ notation to denote the optimal solution):

The discrete Navier–Stokes system

$$(5.1)\qquad \begin{cases} \dfrac{1}{\Delta t}(\vec{u}_h^{(n)} - \vec{u}_h^{(n-1)}, \vec{v}_h) + \nu a(\vec{u}_h^{(n)}, \vec{v}_h) + c(\vec{u}_h^{(n)}; \vec{u}_h^{(n)}, \vec{v}_h) \\ \qquad\qquad\qquad + \ b(\vec{v}_h, p_h^{(n)}) = K(\vec{g}_h^{(n)}, \vec{v}_h) \qquad \forall\, \vec{v}_h \in X_0^h(\Omega)\,, \\ b(\vec{u}_h^{(n)}, q_h) = 0 \qquad \forall\, q_h \in S_0^h(\Omega) \end{cases}$$

for $n = 1, 2, \ldots, N$, with initial velocity $\vec{u}_h^{(0)} = \pi^h \vec{u}_0$ and homogeneous boundary condition.

The discrete adjoint system

$$(5.2)\qquad \begin{cases} -\dfrac{1}{\Delta t}(\vec{w}_h^{(n)} - \vec{w}_h^{(n-1)}, \vec{v}_h) + \nu a(\vec{w}_h^{(n-1)}, \vec{v}_h) \\ \qquad\qquad + \ c(\vec{u}_h^{(n)}; \vec{v}_h, \vec{w}_h^{(n-1)}) + c(\vec{v}_h; \vec{u}_h^{(n)}; \vec{w}_h^{(n-1)}) \\ \qquad\qquad + \ b(\vec{v}_h, r_h^{(n-1)}) = \alpha(\vec{u}_h^{(n)} - \vec{U}^{(n)}, \vec{v}_h) \quad \forall\, \vec{v}_h \in X_0^h(\Omega)\,, \\ b(\vec{w}_h^{(n-1)}, q_h) = 0 \qquad \forall\, q_h \in S_0^h(\Omega) \end{cases}$$

for $n = 1, 2, \ldots, N$, with final condition $\vec{w}_h^{(N)} = \gamma(\vec{u}_h^{(N)} - \pi^h \vec{U}^{(N)})$ and homogeneous boundary condition.

The discrete optimality condition

$$(5.3)\qquad \|\vec{w}_h^{(n-1)}\| \vec{g}_h^{(n)} = -\vec{w}_h^{(n-1)}$$

for $n = 1, 2, \ldots, N$.

In practice, one cannot solve the systems (5.1)–(5.3) simultaneously; note that not only are all three systems coupled to each other, but the discrete Navier–Stokes system (5.1) marches *forward* in time starting from an initial condition, while the adjoint system (5.2) marches *backward* in time starting from a terminal condition.

We now consider a *gradient method* for the solution of the discrete optimality system (5.1)–(5.3). At each iteration the method requires the sequential solution of (5.1) and (5.2).

Let $\mathcal{J}_h^N(k) = \mathcal{J}_h^N(\mathbf{u}_h(\mathbf{g}_h(k)))$, where $\mathcal{J}_h^N(\cdot,\cdot)$ is given by (4.5), k is the iteration counter of the gradient algorithm, and $\mathbf{g}_h(k)$ is the kth iterate for the optimal control. In the algorithm, τ will denote a prescribed tolerance used to test for the convergence of the functional. The gradient algorithm proceeds as follows.

```
(a) initialization:
    (i) choose $\tau$ and $\vec{\mathbf{g}}_h(0)$; set $k = 0$ and $\epsilon = 1$;
    (ii) solve for the starting velocity field $\mathbf{u}_h(0)$ from (5.1) with $\mathbf{g}_h = \mathbf{g}_h(0)$;
    (iii) evaluate $\mathcal{J}_h^N(0)$;.
(b) main loop:
    (iv) set $k = k + 1$;
    (v) solve for $\mathbf{w}_h(k)$ from (5.2) with $\mathbf{u}_h = \mathbf{u}_h(k-1)$;
    (vi) for $n = 1, 2, \dots, N$,
        – set $\vec{q} = \vec{g}_h^{(n)}(k-1) - \epsilon \dfrac{\vec{w}_h^{(n-1)}(k)}{\|\vec{w}_h^{(n-1)}(k)\|}$;
        – if $\|\vec{g}_h^{(n)}(k)\| \le 1$, set $\vec{g}_h^{(n)}(k) = \vec{q}$ ;
        – if $\|\vec{g}_h^{(n)}(k)\| > 1$, set $\vec{g}_h^{(n)}(k) = \dfrac{\vec{q}}{\|\vec{q}\|}$;
    (vii) solve for $\mathbf{u}_h(k)$ from (5.1) with $\mathbf{g}_h = \mathbf{g}_h(k)$;
    (viii) evaluate $\mathcal{J}_h^N(k)$;
    (ix) if $\mathcal{J}_h^N(k) \ge \mathcal{J}_h^N(k-1)$, set $\epsilon = .5\epsilon$ and go to (vi); otherwise continue;
    (x) if $|\mathcal{J}_h^N(k) - \mathcal{J}_h^N(k-1)|/|\mathcal{J}_h^N(k)| > \tau$, set $\epsilon = 1.5\epsilon$ and go to (iv);
        otherwise stop.
```

REMARK 5.1. The bulk of the computational costs are found in the backward-in-time solution of the discrete adjoint system in step (v) and the forward-in-time solution of the discrete state system in step (vii).

REMARK 5.2. The gradient of the fully discrete functional (4.5) is given by, for $n = 1, 2, \dots, N$,

$$\frac{d\mathcal{J}_h^N}{d\vec{g}_h^{(n)}}\bigg|_{\mathbf{g}_h} = K\vec{w}_h^{(n-1)} \,;$$

this follows from the equation analogous to (3.21) for the semidiscrete functional. Then it is clear that the above algorithm is based on the iteration

$$\vec{g}_h^{(n)}(k+1) = \Pi_B \left(\vec{g}_h^{(n)}(k) - \rho_k \frac{d\mathcal{J}_h^N}{d\vec{g}_h^{(n)}}\bigg|_{\mathbf{g}_h(k)} \right) \tag{5.4}$$

for $n = 1, 2, \dots, N$, where Π_B denotes the projection onto the unit ball in X_h, viewed as a subspace of $L^2(\Omega)$.

The convergence property of the projected gradient algorithm is given in the following result.

THEOREM 5.3. *Let* $(\mathbf{u}_h(k), \mathbf{w}_h(k), \mathbf{p}_h(k), \mathbf{r}_h(k), \mathbf{g}_h(k))$ *denote the kth iterate of the projected gradient algorithm, and let* $(\widehat{\mathbf{u}}_h, \widehat{\mathbf{w}}_h, \widehat{\mathbf{p}}_h, \widehat{\mathbf{r}}_h, \widehat{\mathbf{g}}_h)$ *denote a solution of the fully discrete optimality system* (5.1)–(5.3). *Let* $\mathbf{B}_h = (B_h)^N$, *where* B_h *denotes the unit ball in* X_h (*with respect to* $L^2(\Omega)$.) *Then, for* Δt *sufficiently small,* K *sufficiently large, and any* $\mathbf{g}_h(0) \in \mathbf{B}_h$,

$$(\mathbf{u}_h(k), \mathbf{w}_h(k), \mathbf{p}_h(k), \mathbf{r}_h(k), \mathbf{g}_h(k)) \to (\widehat{\mathbf{u}}_h, \widehat{\mathbf{w}}_h, \widehat{\mathbf{p}}_h, \widehat{\mathbf{r}}_h, \widehat{\mathbf{g}}_h) \quad \text{as } k \to \infty\,.$$

Proof. We will make use of the following classical result; see, e.g., [1], [6], [22], or [8]. Let X be a Hilbert space with norm $\|\cdot\|$, and let $\mathcal{K}(\cdot)$ be a real-valued functional on X. Suppose that $\mathcal{K}(\cdot)$ is of class C^2; suppose that $\widehat{x}$ is a local minimizer of $\mathcal{K}(\cdot)$; suppose that there exists a ball B of X, centered at $\widehat{x}$, such that there exist two real numbers c_1 and c_2 such that for all $\widetilde{g} \in B$ and all $\delta x_1, \delta x_2 \in X$,

$$\mathcal{J}''(\widetilde{x})(\delta x_1, \delta x_2) \le c_1 \|\delta x_1\| \|\delta x_2\| \qquad \text{and} \qquad c_2 \|\delta x_1\|^2 \le \mathcal{J}''(\widetilde{x})(\delta x_1, \delta x_1)\,, \tag{5.5}$$

where $\mathcal{J}''(\widetilde{x})(\delta x_1, \delta x_2)$ is the bilinear form associated with the second derivatives of $\mathcal{J}(\cdot)$; and suppose that ρ_k is chosen so that

$$0 < a \le \rho_k \le b < \frac{2c_2}{c_1} \qquad \text{for all } k \tag{5.6}$$

for some positive numbers a and b. Then the iterates of the gradient algorithm

$$x(k+1) = x(k) - \rho_k \nabla\mathcal{K}(x(k)) \quad k = 0, 1, 2, \dots,$$

converge to $\widehat{x}$ for any initial iterate $x(0) \in B$.

For our setting, let $\Delta t = T/N$ and let

$$\mathcal{K}_h^N(\mathbf{u}, \mathbf{g}, \boldsymbol{\beta}, \mathbf{s}) = \mathcal{J}_h^N(\mathbf{u}) - \sum_{n=1}^{N} \frac{\beta^{(n)}}{2} \left(\|\vec{g}_h^{(n)}\|^2 - 1 + (s^{(n)})^2 \right).$$

The functional $\mathcal{K}_h^N$ could have more minimizers than $\mathcal{J}_h^N$, but every optimal solution of our problem is also a minimizer of $\mathcal{K}_h^N$. If $(\widehat{u}_h^{(n)}, \widehat{g}_h^{(n)})$ is an optimal solution, then it minimizes $\mathcal{J}_h^N$ with $\|\widehat{g}_h^{(n)}\| = 1$, $\widehat{\beta}^{(n)} = (\widehat{g}_h^{(n)}, \widehat{w}_h^{(n-1)}) = -K\|\widehat{w}_h^{(n-1)}\|$, and $\widehat{s}^{(n)} = 0$ for all $n = 1, \dots, N$ and $\mathcal{K}_h^N = \mathcal{J}_h^N$ so that the optimal solution minimizes $\mathcal{K}_h^N$. Since we are looking for a minimum point, we search for $\beta^{(n)} \le 0$. The first variation of $\mathcal{K}_h^N$ can be written as

$$D\mathcal{K}_h^N(\mathbf{u}_h(\mathbf{g}_h), \mathbf{g}_h, \boldsymbol{\beta}, \mathbf{s})(\delta\mathbf{g}_h, \delta\boldsymbol{\beta}, \delta\mathbf{s}) = \Delta t \sum_{n=1}^{N} \int_\Omega (K\vec{w}_h^{(n-1)} - \beta^{(n)}\vec{g}_h^{(n)}) \cdot \delta\vec{g}_h^{(n)} \, d\vec{x}$$

$$-\Delta t \sum_{n=1}^{N} \left(\beta^{(n)} s^{(n)} \delta s^{(n)} + \left(\|\vec{g}_h^{(n)}\|^2 - 1 + (s^{(n)})^2\right)\delta\beta^{(n)} \right). \tag{5.7}$$

The search could be in the $\mathbf{g}, \boldsymbol{\beta}$, and $\mathbf{s}$ direction with the gradient given by

$$\begin{aligned} \left.\frac{d\mathcal{K}_h^N}{d\vec{g}_h^{(n)}}\right|_{(\mathbf{g}_h, \boldsymbol{\beta}, \mathbf{s})} &= K\vec{w}_h^{(n-1)} - \beta^{(n)}\vec{g}_h^{(n)}, \\ \left.\frac{d\mathcal{K}_h^N}{d\beta^{(n)}}\right|_{(\mathbf{g}_h, \boldsymbol{\beta}, \mathbf{s})} &= \|\vec{g}_h^{(n)}\|^2 - 1 + (s^{(n)})^2, \\ \left.\frac{d\mathcal{K}_h^N}{ds^{(n)}}\right|_{(\mathbf{g}_h, \boldsymbol{\beta}, \mathbf{s})} &= \beta^{(n)} s^{(n)} \end{aligned}$$

for $n = 1, 2, \dots, N$. We can reduce the number of the directions to only one if we set $\|\vec{g}^{(n)}(k)\|^2 - 1 + s^{2(n)}(k) = 0$ and $s^{(n)}(k) = 0$ if $\|\widetilde{g}^{(n)}(k-1)\|^2 \ge 1$ or $\beta^{(n)}(k) = 0$ if $\|\widetilde{g}^{(n)}\|^2(k-1) < 1$, where k is the kth iteration of the gradient algorithm. If $s^{(n)}(k) = 0$, we can fix $\beta^{(n)}(k)$ equal to $(\vec{g}_h^{(n)}(k), \vec{w}_h^{(n)}(k))$, assuming that $\vec{g}_h^{(n)}(k)$ is orthogonal to the gradient direction $d\mathcal{K}_h^N/d\vec{g}_h^{(n)}$. The variation $\delta\mathbf{g}_h$ can be set proportional to the adjoint solution and $\mathcal{K}_h^N$ decreases for all iterations.

Now the gradient algorithm can be applied to $\mathcal{K}_h^N$ only in the $\mathbf{g}$ direction. For each $\widetilde{\mathbf{g}}_h \in L^2((0,T); \mathbf{X}^h)$, the nonvanishing terms of the second Frechet derivative of $\mathcal{K}_h^N(\widetilde{\mathbf{u}}_h(\widetilde{\mathbf{g}}_h), \widetilde{\mathbf{g}}_h)$ are given by

$$D^2\mathcal{K}_h^N(\widetilde{\mathbf{u}}_h(\widetilde{\mathbf{g}}_h), \widetilde{\mathbf{g}}_h)(\delta\mathbf{g}_{1h}, \delta\mathbf{g}_{2h}) = -\Delta t \sum_{n=1}^{N} \widetilde{\beta}^{(n)} \int_\Omega \delta\vec{g}_{1h}^{(n)} \cdot \delta\vec{g}_{2h}^{(n)} d\vec{x},$$

$$\begin{aligned}&\alpha\Delta t\sum_{n=1}^{N}\int_\Omega \widetilde{w}_{1h}^{(n)}\cdot\widetilde{w}_{2h}^{(n)}\,d\vec{x}+\gamma\int_\Omega \widetilde{w}_{1h}^{(N)}\cdot\widetilde{w}_{2h}^{(N)}d\vec{x}\\ &+\alpha\Delta t\sum_{n=1}^{N}\int_\Omega(\vec{u}_h^{(n)}-\vec{U}_h^{(n)})\cdot\widetilde{z}_h^{(n)}\,d\vec{x}+\gamma\int_\Omega(\vec{u}_h^{(N)}-\vec{U}_h^{(N)})\cdot\widetilde{z}_h^{(N)}\,d\vec{x}\,,\end{aligned}\tag{5.8}$$

where $\widetilde{w}_h^{(n)}\in X_0^h$, $\widetilde{u}_h^{(n)}\in X_0^h$, $\widetilde{w}_{ih}^{(n)}\in X_0^h$, and $\widetilde{z}_h^{(n)}\in X_0^h$ are the solution of the adjoint equations, the solution of the Navier–Stokes equations, and the first and the second variation of $\vec{u}_h^{(n)}$, respectively. By following arguments similar to those used in [21], we have that, for some constants $C_1, C_2 > 0$,

$$\begin{aligned}&|D^2\mathcal{K}_h^N(\widetilde{\mathbf{u}}_h(\widetilde{\mathbf{g}}_h),\widetilde{\mathbf{g}}_h)(\delta\mathbf{g}_{1h},\delta\mathbf{g}_{2h})|\\ &\quad\le\Big(\alpha f_2(\|\widetilde{\mathbf{g}}_h-\widehat{\mathbf{g}}_h\|)\\ &\qquad+\big(\alpha+\gamma+\alpha C_1 f_2(\|\widetilde{\mathbf{g}}_h-\widehat{\mathbf{g}}_h\|)\big)f_1^2(\|\widetilde{\mathbf{g}}_h-\widehat{\mathbf{g}}_h\|)\Big)\|\delta\mathbf{g}_{1h}\|\|\delta\mathbf{g}_{2h}\|\end{aligned}\tag{5.9}$$

and

$$\begin{aligned}&|D^2\mathcal{K}_h^N(\widetilde{\mathbf{u}}_h(\widetilde{\mathbf{g}}_h),\widetilde{\mathbf{g}}_h)(\delta\mathbf{g}_{1h},\delta\mathbf{g}_{1h})|\\ &\quad\ge\Delta t\sum_{n=1}^{N}\Big(-\widetilde{\beta}^{(n)}-C_2\alpha f_2(\|\widetilde{\mathbf{g}}_h-\widehat{\mathbf{g}}_h\|)f_1^2(\|\widetilde{\mathbf{g}}_h-\widehat{\mathbf{g}}_h\|)\Big)\|\delta\mathbf{g}_{1h}\|^2\,,\end{aligned}\tag{5.10}$$

where $f_1(\cdot)$ and $f_1(\cdot)$ are continuous functions. Equation (5.10) can give the correct estimate only if $\widehat{\beta}^{(n)}\neq 0$, which means $\widehat{s}^{(n)}=0$ and $\|\widehat{g}^{(n)}\|=1$. Hence, let $\widehat{w}^{(n-1)}$ be different from zero for all $n=1,\dots,N$ and define $\eta=\min\{\|\widehat{w}^{(n-1)}\| : n=1,\dots,N\}$. We can write

$$\widetilde{\beta}^{(n)}=(\widetilde{g}_h^{(n)},\widetilde{w}_h^{(n-1)})=\widehat{\beta}^{(n)}-(\widehat{g}_h^{(n)}-\widetilde{g}_h^{(n)},\widehat{w}_h^{(n-1)})-(\widetilde{g}_h^{(n)},\widehat{w}_h^{(n-1)}-\widetilde{w}_h^{(n-1)}),$$

and, for K large enough, the estimate

$$|\widetilde{\beta}^{(n)}|\ge K\eta-\alpha(f_2(0)\|\widetilde{\mathbf{g}}_h-\widehat{\mathbf{g}}_h\|+f_2(\|\widetilde{\mathbf{g}}_h-\widehat{\mathbf{g}}_h\|)+f_2(0))\,.$$

Now, assume that $\|\widehat{\mathbf{g}}_h-\widetilde{\mathbf{g}}_h\|\le\xi$ so that, consequently, $f_1(\|\widehat{\mathbf{g}}_h-\widetilde{\mathbf{g}}_h\|)\le\xi_1$ and $\alpha f_2(\|\widehat{\mathbf{g}}_h-\widetilde{\mathbf{g}}_h\|)\le\xi_2$ for some $\xi_1,\xi_2\le\infty$ and choose K large enough. Then, we may choose

$$c_1=\xi_2+(\alpha+\gamma+\alpha C_1\xi_2)\xi_1^2 \qquad\text{and}\qquad c_2=K\eta-\xi_2\xi-2\xi_2-C_2\xi_2\xi_1^2. \qquad \square$$

REMARK 5.4. Steps (ix) and (x) of the algorithm automate the choice of ρ_k satisfying (5.6).

REMARK 5.5. Under the hypotheses of Theorem 5.3 one can also prove that an appropriately defined conjugate gradient algorithm converges.

6. Computational examples. The optimality system presents a formidable computational challenge even for relatively simple geometries and relatively coarse grids. It involves the forward-in-time Navier–Stokes system coupled to the backward-in-time adjoint system.

One can split the optimality system into its two constituent systems of equations. At any step of the iteration, the Navier–Stokes system can be solved by marching forward in time from the initial condition; at any time, the body force must be determined, through the gradient algorithm, from the adjoint solution obtained in the previous step. At every time step, the nonlinearity of the Navier–Stokes system is treated with a modified Newton–Raphson method. The adjoint equations can then be solved marching backward in time, starting from the final condition, which along with the right-hand side and coefficients, are determined from the Navier–Stokes solution. Thus, both the state and adjoint variables must be available and stored over the entire space-time domain. The iterative procedure, i.e., the forward Navier–Stokes sweeps followed by the backward adjoint sweeps, is repeated until convergence is achieved. The projected gradient algorithm with variable step, described in the previous section, is a slowly convergent method and many iterations are necessary, even for the simple test cases discussed below.

The solution globally consists of eight fields: three for the state of the fluid, three for the adjoint state, and two for the control. As was mentioned above, this solution must be available over the whole time-space domain and cannot be stored in the computer memory, even for simple test cases that use a 20×20 spatial mesh and 100 time steps. Therefore, these fields must be stored out of core and be accessible during the matrix and right-hand side assemblies.

The scheme employed for solving linear systems is the frontal method described in [23]. The implementation of the method we use employs out-of-core storage. Since there are not limitations on the storage of the solver matrix with this method, many engineering problems can be adequate treated. The primary disadvantage of the frontal method is the effort necessary to access and store the matrix.

Of course, our solution strategy and its implementation is a straightforward one. Other more sophisticated strategies and implementations can be devised that require less storage and perhaps less central processing resources. This is a subject that is currently of substantial interest to us and the flow control community.

6.1. Test 1. We consider a unit square domain $(0,1) \times (0,1) \subset \mathbb{R}^2$. We assume that the time interval [0,1] is divided into equal intervals of duration $\Delta t = 1/N$. The finite element spaces are chosen to be continuous piecewise biquadratic polynomials for the velocity and continuous piecewise bilinear polynomials for the pressure, i.e., the Taylor–Hood finite element pair, with respect to a rectangular mesh. The mesh size is denoted by h, and calculations with varying mesh sizes have been performed. In this first test we are interested in the convergence history with respect to the parameters involved, so a simple stationary target velocity $\vec{U}(x,y) = (U(x,y), V(x,y))$ is chosen; e.g.,

$$U = 10\frac{d}{dy}\Big(\phi(0.4,x)\phi(0.4,y)\Big) \quad \text{and} \quad V = -10\frac{d}{dx}\Big(\phi(0.4,x)\phi(0.4,y)\Big)\,, \tag{6.1}$$

where

$$\phi(t,z) = (1-z)^2\big(1 - cos(2\pi t z)\big)\,.$$

Velocity tracking evolution. For the first example we choose the initial velocity

$$u_0(x,y) = -U(x,y) \qquad \text{and} \qquad v_0(x,y) = -V(x,y)$$

so that the initial flow rotates in an opposite sense from the target flow. The flow evolution is given in Figure 6.1. The controlled fluid is on the left, and the desired flow is on the right. (In the figures, all values are normalized by the maximum value.) Figure 6.2 shows the error $\|\widehat{u} - \vec{U}\|$ between the optimally controlled flow $\widehat{u}$ and the

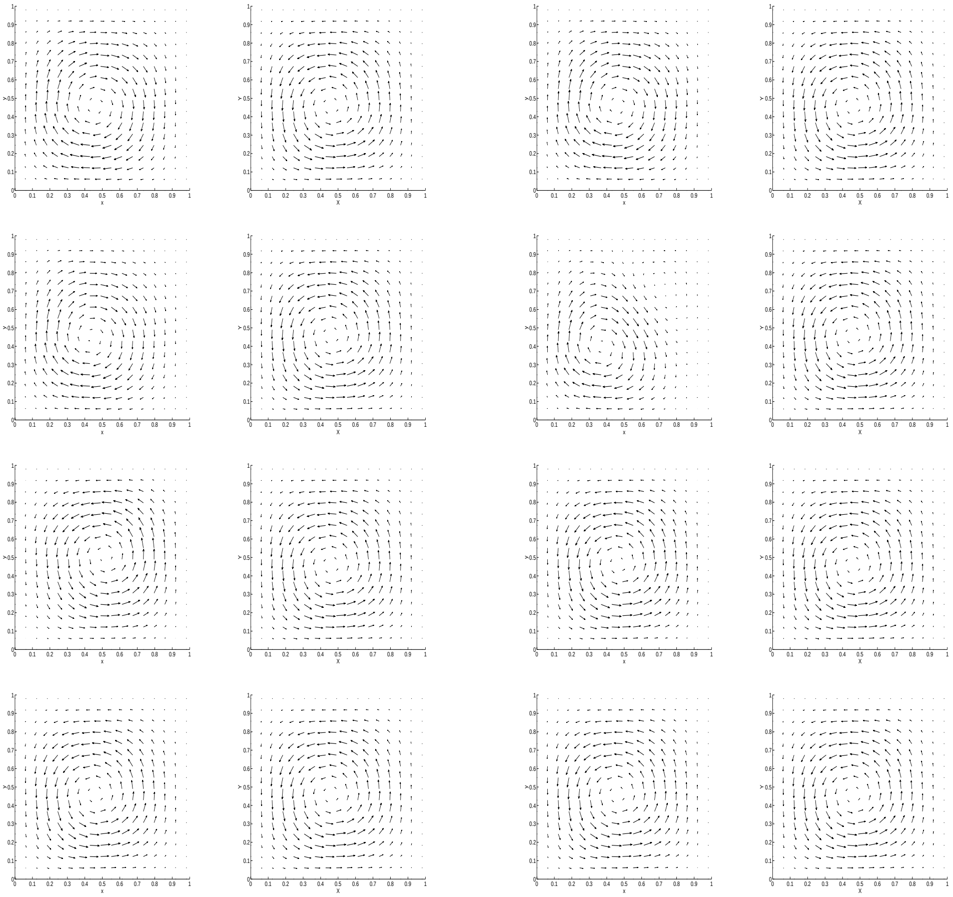


FIG. 6.1. *Test* 1. *Controlled (left) and target (right) flows at* $t = 0, .05, .1, .125, .135, .2, .5, 1$.

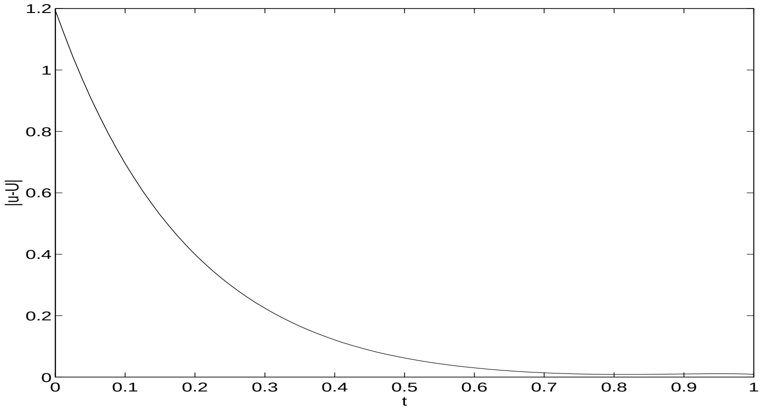


FIG. 6.2. *Test* 1. *Error* $\|\widehat{u} - \vec{U}\|$ *vs. time.*

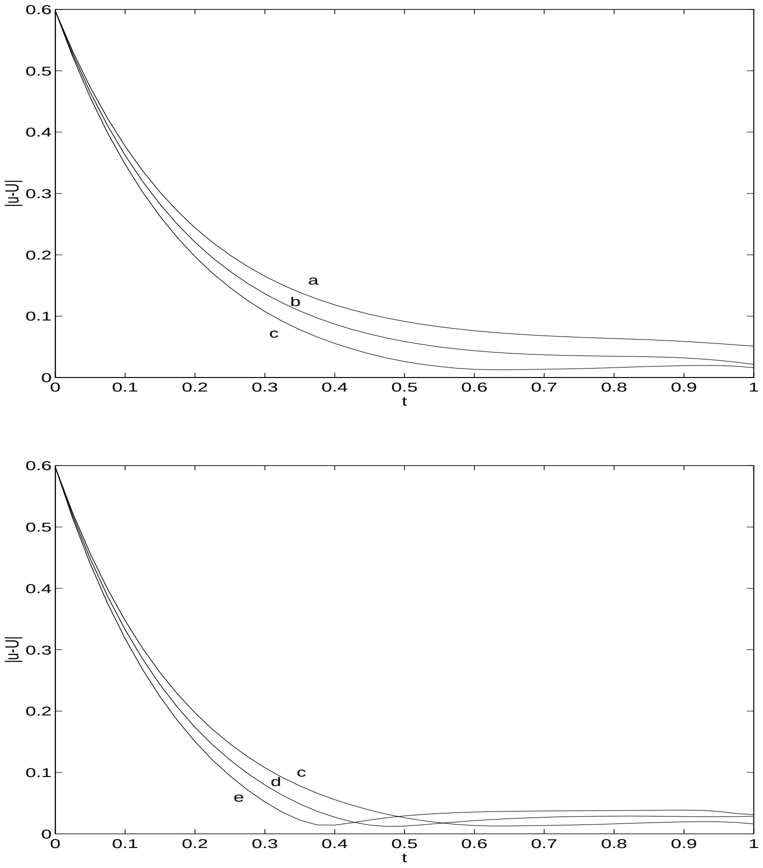


FIG. 6.3. *Test* 1. *Error* $\|\widehat{u} - \vec{U}\|$ *vs. time for different values of the bound* K *for the control;* $K = 3.0$ (a), 3.2 (b), 3.4 (c), 3.6 (d), 3.8 (e).

target flow $\vec{U}$. As we can see, the error does not go to zero very rapidly due to the fact that we have a bounded control. For this calculation, by $t = 0.15$ we achieve a match in shape of the flow and at $t = 0.6$ we achieve a match in magnitude as well. For this calculation, $\Delta t = 0.025$, $h = 1/16$, $\alpha = 1$, $\gamma = 0.5$, and $K = 3.2$.

Velocity tracking with different control norm. We want to analyze what happens if we change K, the a priori bound for the norm of control $\vec{f}$. The values $\Delta t = 0.025$, $h = 1/16$, $\alpha = 1$, and $\gamma = 0.5$ are used in all cases. We now set the initial velocity of the controlled flow to zero and the target flow is still given by (6.1). In Figure 6.3, we give the error $\|\widehat{u} - \vec{U}\|$ between the controlled flow $\widehat{u}$ and the target flow $\vec{U}$ for different value of K. (We are really giving information about the solution of the fully discrete optimal control problem.) We have $K = 3.0$ (a), 3.2 (b), 3.4 (c) and 3.4 (c), 3.6 (d), 3.8 (e), respectively. We observe that for $K \leq 3.4$, as K increases, the controlled flow matches better over the whole time interval. However, for $K > 3.4$, there is a better match only up to some time $t^* < T = 1$, but for $t > t^*$, the performance of the controlled flow gets worse as K increases. Of course, we are minimizing the functional (4.5) that involves a sum over the time steps; it does not follow that the error $\|\widehat{u} - \vec{U}\|$ necessarily decreases with increasing K at every time step. The norm

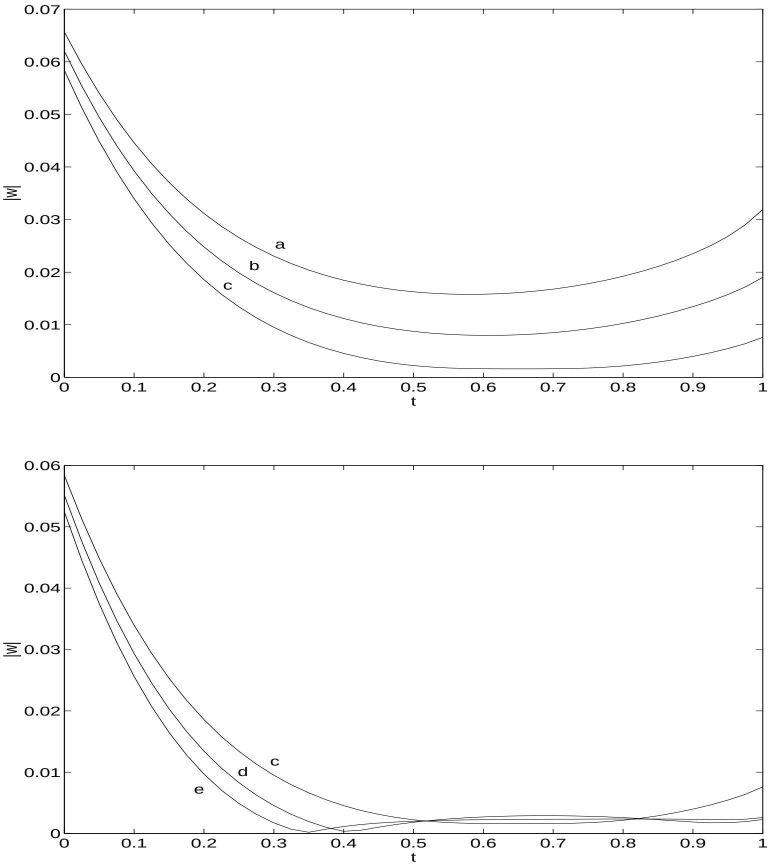


FIG. 6.4. *Test* 1. *Adjoint function norm $\|\vec{w}\|$ vs. time for different values of the bound K for the control;* $K = 3.0$ (a), 3.2 (b), 3.4 (c), 3.6 (d), 3.8 (e).

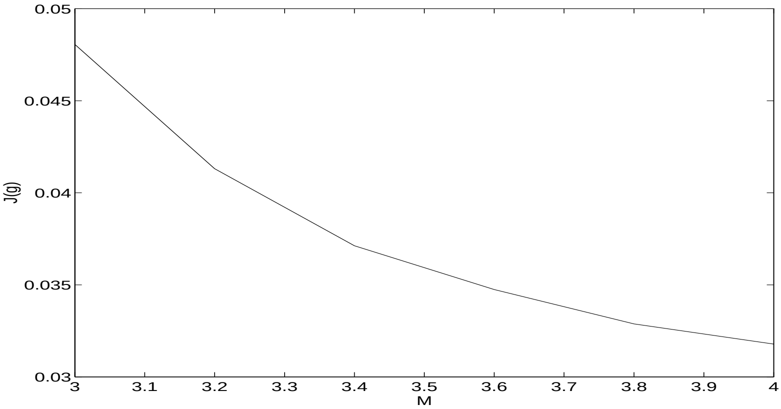


FIG. 6.5. *Test* 1. *Value of the functional vs. the bound $K = M$ for the control.*

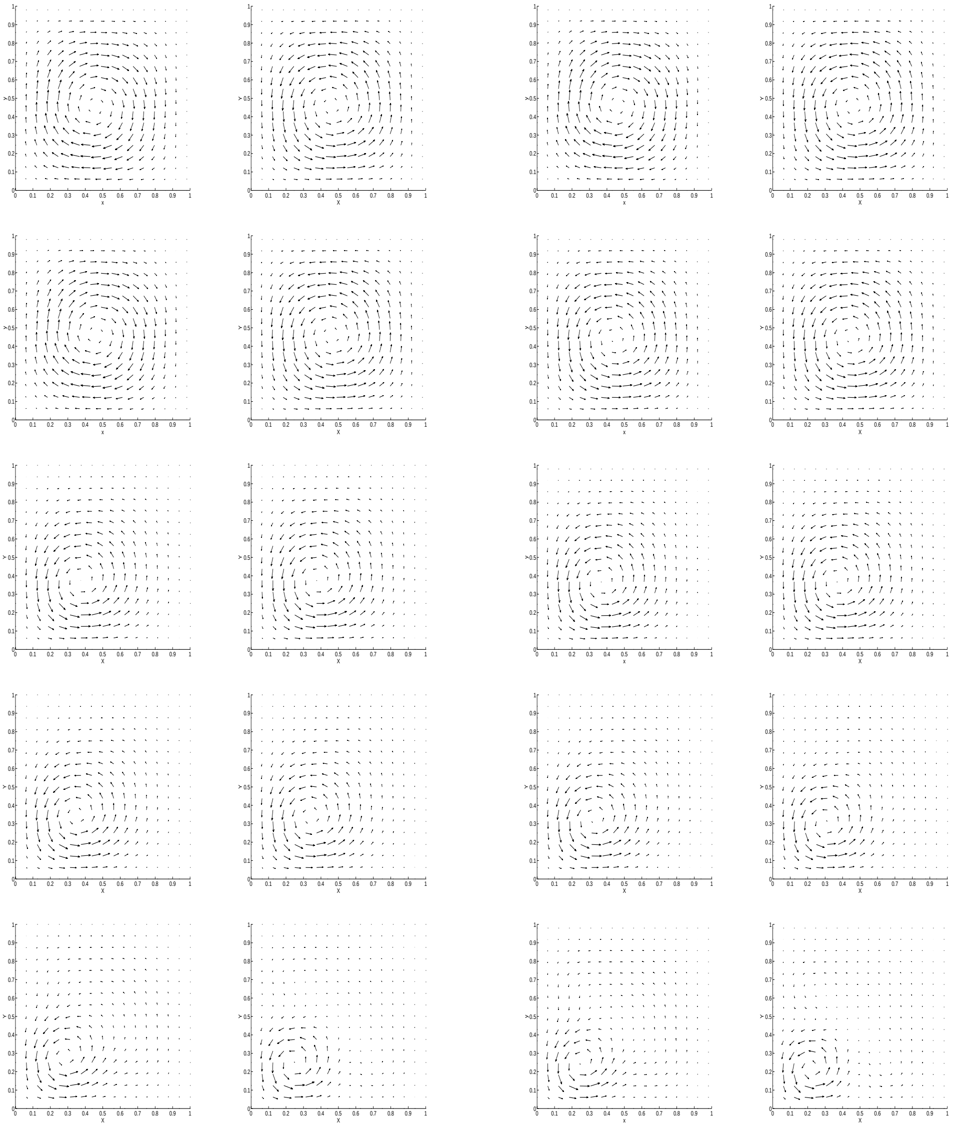


FIG. 6.6. *Test 2. Controlled (left) and target (right) flows at* $t = .1, .2, .3, .4, .5, .6, .7, .8, .9, 1$.

$\|\vec{w}\|$ of the adjoint variable agrees with the intuitive behavior of the error between the controlled and target flows. That norm is given in Figure 6.4 for the different values of K. In Figure 6.5, we display the value of the functional (4.5), evaluated at the optimal solution, vs. K. As is expected, the functional decreases as K increases.

6.2. Test 2. We consider a unit square domain $(0,1) \times (0,1) \subset \mathbb{R}^2$. We assume that the time interval [0,1] is divided in equal intervals of time $\Delta t = 1/N$. The Taylor–Hood finite element pair is used with respect to a rectangular mesh. We report only results for $h = 1/16$; however, calculations with varying mesh sizes have

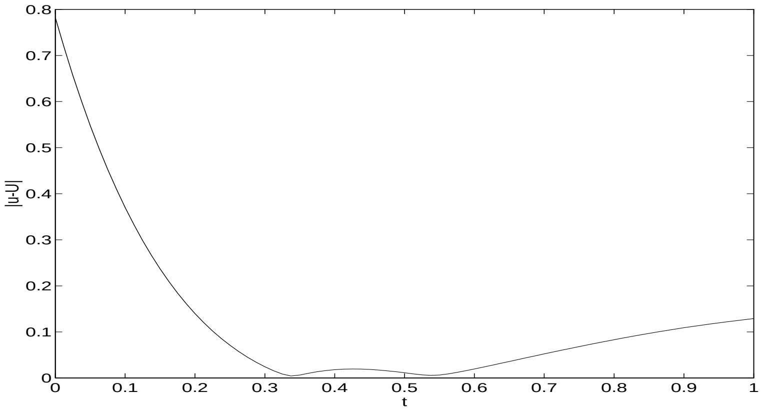


FIG. 6.7. *Test 2. Error* $\|\widehat{u} - \vec{U}\|$ *vs. time.*

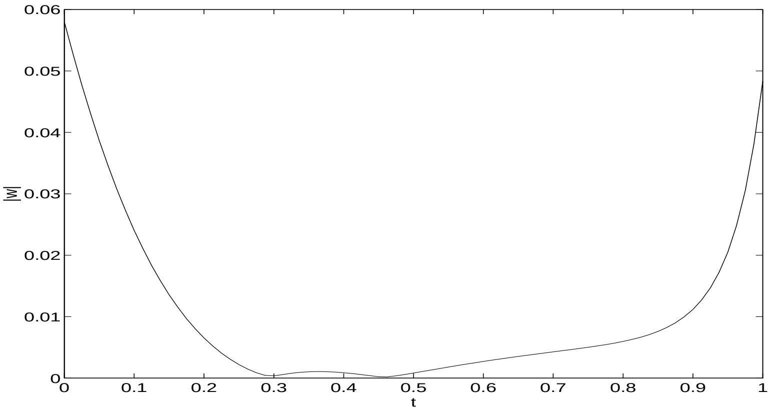


FIG. 6.8. *Test 2. Adjoint function norm* $\|\vec{w}\|$ *vs. time.*

been performed. The target velocity $\vec{U}(x, y)$ is chosen to be

$$U(x, y) = a(1, .4, x, y) + a(2, t, x, y)/(4\pi t + 1),$$
$$V(x, y) = b(1, .4, x, y) + b(2, t, x, y)/(4\pi t + 1)\,,$$

where

$$\phi(k, t, z) = (1 - zt)^2\big(1 - cos(2k\pi tz)\big),$$
$$a(k, t, x, y) = \frac{d}{dy}\Big(\phi(k, t, x)\phi(k, t, y)\Big),$$
$$b(k, t, x, y) = -\frac{d}{dx}\Big(\phi(k, t, x)\phi(k, t, y)\Big)\,.$$

This velocity field is a superposition of two flows, one having a vortex at the center of the domain and another flow with four vortices. Each of these flows prevails at different times of the evolution. The initial velocity for the controlled flow is chosen to be

$$u_0(x,y) = -8U(1/4,x,y) \qquad \text{and} \qquad v_0(x,y) = -8V(1/4,x,y)$$

so that it has an opposite rotational sense and magnitude from that of the target flow. For this computation, $\alpha = 1$, $\gamma = 0.5$, and $K = 1.6$. The evolutionary history is given in Figure 6.6. The controlled fluid is on the right, the desired flow is on the left, and all the pictures are normalized. We can see that by $t = 0.3$, the controlled flow looks very much like the desired flow. Figure 6.7 shows the error $\|\widehat{u} - \vec{U}\|$ between the controlled flow $\widehat{u}$ and the target flow $\vec{U}$. At the beginning, the error rapidly decreases but after this initial interval of time the error increases due to changes in the desired flow which cannot be followed well by a control with the bound we have chosen. The norm of the control is forced to be less then 1.6. Evidently, this does not allow enough power to well-match the time evolution of the desired flow. For the same flow, Figure 6.8 shows the values of the norm of the adjoint variable $\vec{w}$.

REFERENCES

[1] F. Abergel and R. Temam, *On some control problems in fluid mechanics*, Theor. Comput. Fluid Dyn., 1 1990, pp. 303–325.

[2] R. Adams, *Sobolev Spaces*, Pure Appl. Math 65, Academic Press, New York, 1975.

[3] V. Alekseev, V. Tikhomirov, and S. Fomin, *Optimal Control*, Contemporary Soviet Mathematics, Consultants Bureau, New York, 1987.

[4] H. T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, System Control Found. Appl. 1, Birkhauser, Boston, 1989.

[5] M. Berggren, *Control and Simulation of Advection-Diffusion Problems*, Ph.D. thesis, Mechanical Engineering, University of Houston, Houston, TX, 1992.

[6] D. P. Bersekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1995.

[7] D. Buschnell and C. McGinley, *Turbulent control in wall flows*, Annu. Rev. Fluid Mech., 21 (1989), pp. 1–20.

[8] P. Ciarlet, *Introduction to Numerical Linear Algebra and Optimization*, Cambridge University Press, Cambridge, UK, 1989.

[9] P. Constantin and C. Foias, *Navier–Stokes Equations*, University of Chicago Press, Chicago, IL, 1989.

[10] H. Fattorini and S. Sritharan, *Existence of controls for viscous flow problems*, Proc. Roy. Soc. London Ser. A, 439 (1992), pp. 81–102.

[11] H. Fattorini and S. Sritharan, *Necessary and sufficient conditions for optimal controls in viscous flow problems*, Proc. Roy. Soc. Edinburgh Sect. A, 124A (1994), pp. 211–251.

[12] H. Fattorini and S. Sritharan, *Optimal chattering control for viscous flows*, Nonlinear Anal., 25 (1995), pp. 763–797.

[13] A. Fursikov, *On some control problems and results concerning the unique solvability of a mixed boundary value problems for the three-dimensional Navier–Stokes and Euler systems*, Sov. Math. Dokl., 21 (1980), pp. 889–893.

[14] A. Fursikov, *Control problems and theorems concerning the unique solvability of a mixed boundary value problems for the three-dimensional Navier–Stokes and Euler equations*, Math. USSR Sb., 43 (1982), pp. 281–307.

[15] A. Fursikov, *Properties of solutions of some extremal problems connected with the Navier–Stokes system*, Math USSR Sb., 46 (1983), pp. 323–351.

[16] V. Girault and P.-A. Raviart, *Finite Element Method for Navier–Stokes Equations: Theory and Algorithms*, Springer-Verlag, New York, 1986.

[17] R. Glowinski, A. Kearsley, T. Pan, and J. Periaux, *Numerical simulation and optimal shape for viscous flow by a fictitious domain method*, Internat. J. Numer. Methods Fluids, 20 (1995), pp. 695–711.

[18] M. Gunzburger, L. Hou, and T. Svobodny, *Numerical approximation of an optimal control associated with the Navier–Stokes equations*, Appl. Math. Lett., 1 (1989), pp. 29–31.

[19] M. Gunzburger, L. Hou, and T. Svobodny, *Analysis and finite element approximations of optimal control problems for the stationary Navier–Stokes equations with Dirichlet controls*, RAIRO Modél. Anal. Numér., 25 (1991), pp. 711–748.

[20] M. Gunzburger, L. Hou, and T. Svobodny, *Optimal control and optimization of viscous, incompressible flow*, in Incompressible Computational Fluid Dynamics, M.D. Gunzburger and R.A. Nicolaides, eds., Cambridge University Press, Cambridge, New York, 1993, pp. 109–150.

[21] M. Gunzburger and S. Manservisi, *Analysis and approximation of the velocity tracking problem for Navier–Stokes flows with distributed control*, SIAM J. Numer. Anal., to appear.

[22] W. A. Gruver and E. W. Sachs, *Algorithmic Methods in Optimal Control*, Res. Notes Math. 47, Pitman Advanced Publishing Program, London, 1980.

[23] P. Hood, *frontal solution program for unsymmetric matrices*, Internat. J. Numer. Methods Engrg., 10 (1976), pp. 379–400.

[24] L. Hou, S. Ravindran, and Y. Yan, *Numerical solutions of optimal distributed control problems for incompressible flows*, Internat. J. Comput. Fluid Dynam., 8 (1997), pp. 99–114.

[25] L. S. Hou and Y. Yan, *Dynamics for controlled Navier–Stokes systems with distributed controls*, SIAM J. Control Optim., 35 (1997), pp. 654–677.

[26] L. Hou and Y. Yan, *Dynamics and approximations of a velocity tracking problem for the Navier–Stokes flows with piecewise distributed controls*, SIAM J. Control Optim., 35 (1997), pp. 1847–1885.

[27] A. Jameson, *Computational aerodynamics for aircraft design*, Science, 245 (1989), pp. 361–371.

[28] R. D. Joslin, M. D. Gunzburger, R. A. Nicolaides, G. Erlebacher, and M. Y. Hussaini, *Self-contained automated methodology for optimal flow control*, AIAA J., 35 (1997), pp. 816–824.

[29] J.-L. Lions, *Control of Singular Systems*, Bordas, Paris, 1985.

[30] S. Manservisi, *Optimal and Feedback Control of the Time Dependent Navier–Stokes Equations*, Ph.D. thesis, Virginia Tech, Blacksburg, VA, 1997.

[31] K. McManus, T. Poinsot, and S. Candel, *Review of active control of combustion instabilities*, Prog. Energy Comb. Sci., 19 (1993), pp. 1–29.

[32] S. Sritharan, *Dynamic programming of the Navier–Stokes equations*, Systems Control Lett., 16 (1991), pp. 299–307.

[33] R. Temam, *Navier–Stokes Equations*, North-Holland, Amsterdam, 1979.

[34] R. Temam, *Navier–Stokes Equations and Nonlinear Functional Analysis*, 2nd ed., SIAM, Philadelphia, PA, 1995.

[35] V. Tikhomirov, *Fundamental Principles of the Theory of Extremal Problems*, Wiley, Chichester, UK, 1986.

[36] C. R. Vogel, *A constrained least squares regularization method for nonlinear ill-posed problems*, SIAM J. Control Optim., 28 (1990), pp. 34–49.

SIAM J. CONTROL OPTIM.
Vol. 37, No. 6, pp. 1946–1979

APPROXIMATION AND LIMIT RESULTS FOR NONLINEAR FILTERS OVER AN INFINITE TIME INTERVAL*

AMARJIT BUDHIRAJA[†] AND HAROLD J. KUSHNER[‡]

Abstract. This paper is concerned with approximations to nonlinear filtering problems that are of interest over a very long time interval. Since the optimal filter can rarely be constructed, one needs to compute with numerically feasible approximations. The signal model can be a jump-diffusion or just a process that is approximated by a jump-diffusion. The observation noise can be either white or of wide bandwidth. The observations can be taken in either discrete or continuous time. The cost of interest is the pathwise error per unit time over a long time interval. It is shown under quite reasonable conditions on the approximating filter and the signal and noise processes that (as time, bandwidth, process, and filter approximation, etc.) go to their limit in any way at all, the limit of the pathwise average costs per unit time is just what one would get if the approximating processes were replaced by their ideal values and the optimal filter were used. Analogous results are obtained (with appropriate scaling) if the observations are taken in discrete time, and the sampling interval also goes to zero. For these cases, the approximating filter is a numerical approximation to the optimal filter for the presumed limit (signal, observation noise) problem.

Key words. nonlinear filters, approximations to nonlinear filters, infinite time filtering, occupation measures, pathwise average error criteria

AMS subject classifications. 93E11, 60G35

PII. S0363012997328178

1. Introduction.

This paper is concerned with approximations to nonlinear filtering problems that are of interest over a very long time interval. Consider the simplest such problem where the observation is a function of the state of a diffusion process and either or both the function or the process is nonlinear. Then, except in a few cases which on the whole are not of much practical value, it is not feasible to compute the truly optimal filter. Then one must approximate in some way.

Suppose that the approximation is parameterized by a parameter h such that as $h \to 0$, the approximation converges to the true filter in the sense that the computed expectations of any bounded and continuous function converge to the true conditional expectation; equivalently, the computed approximating conditional distribution converges weakly to the true conditional distribution. Now suppose that the filter is of interest over an arbitrarily large interval T and the errors of concern are the *pathwise average* (not the mean value) "prediction" errors per unit time for some rather arbitrary definition of prediction. Pathwise errors are a main item of interest in many such applications since we work with only one long path, and the law of large numbers argument which justifies the use of mean values is inappropriate.

Now there are two parameters: h and T. The convergence of the filter over any fixed finite interval says nothing about the behavior of the pathwise average errors as $h \to 0$ and $T \to \infty$ arbitrarily. We will show that, under quite reasonable conditions,

*Received by the editors October 3, 1997; accepted for publication (in revised form) February 17, 1999; published electronically October 8, 1999. The research of the first author was supported by contracts N00014-96-1-0276 and N00014-96-1-0279 from the Office of Naval Research. The research of the second author was supported by contract DAAH04-96-1-0075 from the Army Research Office and NSF grant ECS 9703895.

http://www.siam.org/journals/sicon/37-6/32817.html

[†]Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556.

[‡]Division of Applied Mathematics, Brown University, Providence, RI 02912 (hjk@dam.brown.edu).

the pathwise errors converge in probability to an optimal deterministic limit in a very natural sense. This limit is what one would get if one used the true optimal filter and not an approximation. In fact, the limit is also what one would get for the true optimal filter if the expectation of the pathwise average replaced the pathwise average. The convergence is independent of how $h \to 0$ or $T \to \infty$. The error function is quite general and can even be an (appropriate) path functional. This is quite a strong result. Note that it is important that we allow $h \to 0$ and $T \to \infty$ in an arbitrary way. If, for example, we required that the time required to get a good approximation by the "ergodic limit" depends on h, for small h, then the approximating filter would not be good for large time, since the better the filter is (i.e., the smaller h is), the more time is required for the pathwise average to be well approximated by the ergodic limit.

Actually, our interest is in more general problems as well. We let the signal process be replaced by a "pre-jump-diffusion," which is only approximated by a jump-diffusion. The discrete time case is also treated. Also, we allow wide bandwidth observation noise. Much (perhaps most) of the theoretical work in nonlinear filtering is in continuous time, and white observation noise is the usual one. But most applications are in discrete time. If the interval between observations becomes small so that one is tempted to use a continuous time model, then one is confronted with the very real presence of wide bandwidth and not white noise. One usually constructs a numerical approximation to a filter which assumes white observation noise and a diffusion model. Then the wide bandwidth observation noise, combined with all of the other approximations, can conceivably lead to serious errors when T is large.

The last part of the paper deals with this general problem. The observations are in discrete time with a small time interval between the observations, the observation noise is wide bandwidth, the signal is only approximated by a sampled diffusion, and the pathwise average errors over a long time interval T are of interest. Under reasonable conditions we show that the desired limit result holds here as well, as all the parameters go to their limits simultaneously (observation interval, observation noise bandwidth, h, the signal converging to a jump-diffusion, T, etc.). This is a true justification of filtering work in continuous time.

Some work on this problem was done in [21], where the pathwise average error was replaced by an expectation of the pathwise average error, and which contains much additional material on the wide bandwidth observation noise problem. In [12], the asymptotics of the filter alone was dealt with, and it was presumed that the filter was the true optimal filter, not an approximation. The approach taken here uses an adaptation of the occupation measure approach which was used successfully on a variety of control and limit problems in [19, 22].

2. Problem formulation and occupation measures.

Let h parameterize the approximating filter. Until section 5, we work with observations in continuous time. Let $X(\cdot)$ denote the basic signal process, which is the solution to an Itô equation. The process can be a jump-diffusion with time independent coefficient functions, with only notational changes required in the development. But for notational simplicity, we work with a simple diffusion, which takes values in a compact set in $\mathbb{R}^r$, Euclidean r-space. Where continuity is used in the current proof, for the jump-diffusion case it is replaced by the fact that we have continuity at each time with probability one, and uniform (in t) right continuity in probability. Similarly, under appropriate conditions on the boundary and reflection directions, the results hold for a reflected jump-diffusion. Our basic condition on the signal process is the following.

ASSUMPTION 2.1. *$X(\cdot)$ takes values in a compact subset G of $\mathbb{R}^r$ for all $X(0) \in G$ and satisfies the Itô equation*

$$dX = p(X)dt + \sigma(X)dW, \tag{2.1}$$

where $W(\cdot)$ is a standard vector-valued Wiener process and $p(\cdot)$ and $\sigma(\cdot)$ are continuous. The solution of the above Itô equation is unique in the weak sense for each initial condition.

The uniqueness assumption imposes rather mild requirements on $p(\cdot)$ and $\sigma(\cdot)$, while the compactness of the state space will hold if, say, $\sigma(\cdot)$ vanishes outside a compact set and 0 is a global attractor for the deterministic equation obtained by substituting $\sigma = 0$ in (2.1). Alternatively, compactness of the state space can be achieved by constraining a diffusion in a bounded region by appropriately reflecting it at the boundary. Then the model is represented as the solution to the Skorohod problem, and the techniques of the current paper can be used to cover this case as well. In many applications, the state spaces are inherently compact; for example, where the state is the phase of a signal. Additionally, unbounded signal models are often used for mathematical convenience, while the true process does take values in a compact set.

The observation process is $Y(\cdot)$, defined by

$$Y(t) = \int_0^t g(X(s))ds + B(t), \tag{2.2}$$

where $g(\cdot)$ is a continuous vector-valued function and $B(\cdot)$ is a standard vector-valued Wiener process, independent of $W(\cdot)$ and $X(0)$.

Unless otherwise stated all statements of the form, "with probability one (w.p.1)," "almost surely (a.s.)", "convergence in probability," etc., will correspond to the basic probability space which supports $X(\cdot), Y(\cdot), W(\cdot), B(\cdot)$.

The optimal filter process. We will use the representation of the optimal filter as it was originally developed in [15]. This is most convenient for our purposes and it is completely equivalent to the forms used later, as in [8, 11, 23]. Let $\tilde{X}(\cdot)$ be a process satisfying (2.1), and one which (loosely speaking) is conditionally independent of $(X(\cdot), W(\cdot), B(\cdot))$ given its initial condition. We formalize this as follows: $\tilde{X}(\cdot)$ is a process satisfying (2.1) such that there exists a (possibly random) probability measure Π^* on $\mathbb{R}^r$ with the properties that conditioned on Π^*, $\tilde{X}(\cdot)$ is independent of $(X(\cdot), W(\cdot), B(\cdot))$, and the conditional distribution of $\tilde{X}(0)$ given Π^* is Π^*. We will call Π^* the "random initial distribution" of $\tilde{X}(\cdot)$ (i.e., the distribution of $\tilde{X}(0)$). It will vary depending on the need and will be specified when needed.

Let $Y_{a,b}, a < b$, denote the values of $Y(\cdot) - Y(a)$ in the time interval $[a, b]$. Until further notice, let $\Pi(0)$ denote the distribution of $X(0)$ and $\Pi(t)$ the distribution of $X(t)$, given the data $Y_{0,t}$ and $\Pi(0)$. Let $\tilde{X}_{0,a}$ denote the values of $\tilde{X}(\cdot)$ in the time interval $[0, a]$. Define the function

$$R(\tilde{X}_{0,t}, Y_{0,t}) = \exp\left[\int_0^t g'(\tilde{X}(s))dY(s) - \frac{1}{2}\int_0^t \left|g(\tilde{X}(s))\right|^2 ds\right]. \tag{2.3}$$

Let $E_Z f$ denote the expectation of a function f given the data Z. Then the optimal filter $\Pi(\cdot)$ satisfies the following relation (which defines the inner product notation). For each bounded and measurable real-valued function $\phi(\cdot)$

$$\int \phi(x)\Pi(t, dx) \equiv \langle \Pi(t), \phi \rangle = \frac{E_{\{\Pi(0), Y_{0,t}\}}\left[\phi(\tilde{X}(t))R(\tilde{X}_{0,t}, Y_{0,t})\right]}{E_{\{\Pi(0), Y_{0,t}\}}R(\tilde{X}_{0,t}, Y_{0,t})}, \tag{2.4}$$

where $\Pi(0)$ is the distribution of the initial condition $\tilde{X}(0)$.

Owing to the Markov property of $X(\cdot)$, the optimal filter defined by (2.4) satisfies the following semigroup relation: For $0 < s < t$

$$\langle \Pi(t), \phi \rangle = \frac{E_{\{\Pi(t-s), Y_{t-s,t}\}} \left[\phi(\tilde{X}(s)) R(\tilde{X}_{0,s}, Y_{t-s,t}) \right]}{E_{\{\Pi(t-s), Y_{t-s,t}\}} R(\tilde{X}_{0,s}, Y_{t-s,t})}. \tag{2.5}$$

In (2.5), $\Pi(t-s)$ is the (possibly random) distribution of the initial condition $\tilde{X}(0)$. Throughout the paper, we use the notation $E_{\{\Pi(a), Y_{a,b}\}} F(\tilde{X}_{0,s}, Y_{a,b})$ for the conditional expectation, given the data $Y_{a,b}$ and where the random initial distribution for $\tilde{X}(\cdot)$ is $\Pi(a)$. The analogous notation will be used when approximations to $\tilde{X}(\cdot)$ are used.

Since $\Pi(t)$ is nearly always hard (if not impossible) to compute for nonlinear problems, in applications one uses various approximations. Let $\Pi^h(\cdot)$ denote the actual measure-valued process, to be specified more precisely below (see (2.7)), which is the "numerical" filter used in the application. Thus, $\Pi^h(t)$ (with values $\Pi^h(t, dx)$) is the approximation at time t.

As discussed earlier, $\Pi(0)$ and $\Pi^h(0)$, which represent the initial condition of the auxiliary process used in the filter formula, will be allowed to be random (i.e., they can be measure-valued random variables, although always nonanticipative with respect to the Wiener processes. In this case, we can still write $P_{\Pi(0)}\{X(0) \in A\} = \Pi(0, A)$, random or not.

Defining $\Pi(0)$ arbitrarily. Equation (2.4) defines a measure-valued process $\Pi(\cdot)$, with initial condition $\Pi(0)$. This process $\Pi(\cdot)$ is well defined even if $\Pi(0)$ is not the distribution of $X(0)$, although it would not then be the optimal filter. Keep in mind the following important fact: *In what follows, we allow* $\Pi(0)$ *to be arbitrary but with support in G, and we will note it explicitly if it is the distribution of* $X(0)$. Then, the semigroup property in equation (2.5) shows that the pair $(X(\cdot), \Pi(\cdot))$ is Markov.

An approximating auxiliary process and approximate filter. A common way to approximate the optimal filter is to find a process which suitably approximates $X(\cdot)$ but for which the optimal filter is feasible. We then construct the optimal filter for the approximating process but use the actual observations (2.2). This was the motivation for the Markov chain approximation filter in [16, 20]. The approximating filter $\Pi^h(\cdot)$, which will be defined later in this section, will be constructed by replacing the auxiliary process $\tilde{X}(\cdot)$ in (2.4) by $\tilde{X}^h(\cdot)$, where $\tilde{X}^h(\cdot)$ approximates the auxiliary process $\tilde{X}(\cdot)$ but is "simpler" than it. Thus, this filter is built under the assumption that the true signal process is $\tilde{X}^h(\cdot)$, not $X(\cdot)$. The process $\tilde{X}^h(\cdot)$ might be Markov; for example, a continuous time Markov chain on a finite state space. More commonly, it is an interpolation of a discrete parameter process; i.e, there is $\delta_h > 0$ and which goes to zero as $h \to 0$ such that $\tilde{X}^h(\cdot)$ is constant on the intervals $[n\delta_h, n\delta_h + \delta_h)$ and $\tilde{X}^h(n\delta_h), n = 0, \ldots,$ is Markov. When the signal process is defined in continuous time, we always assume that $\tilde{X}^h(\cdot)$ is of one of these two forms. Furthermore, we always suppose (without loss of generality) that $\tilde{X}^h(t)$ takes values in G.

To quantify the sense in which $\tilde{X}^h(\cdot)$ approximates $\tilde{X}(\cdot)$, we will assume the following basic consistency condition.

ASSUMPTION 2.2. *For any sequence $\{\Pi^h\}$ of probability measures converging weakly to some probability measure Π, $\tilde{X}^h(\cdot)$ with the initial distribution Π^h converges weakly to $\tilde{X}(\cdot)$ with the initial distribution Π.*

This is a very weak assumption on the approximating filter. It holds, for example, when $\tilde{X}^h(\cdot)$ is a finite state Markov chain, which is "asymptotically consistent" with

$X(\cdot)$, as defined in [16, 20]. Indeed, the convergence (as $h \to 0$ and for an arbitrary finite time) of filters based on this approximating model has been proved in [16, 20], and the approximation performs well if the state dimension is not higher than four. Another possibility is to use an appropriate discrete time approximation to $\tilde{X}(\cdot)$. One might discretize time, but not space, provided that the computation is feasible. In either case, Assumption 2.2 will be satisfied if the approximating process satisfies the minimal consistency conditions in [16, 20].

Similar to the construction of $\tilde{X}(\cdot)$, we will take $\tilde{X}^h(\cdot)$ to be independent of $(X(\cdot), W(\cdot), B(\cdot))$, given the initial condition. The distribution of the initial condition $\tilde{X}^h(0)$ will be a measure-valued random variable, analogous to what was done in (2.4) and (2.5), when $\tilde{X}(\cdot)$ was used. Define

$$R(\tilde{X}^h_{0,t}, Y_{0,t}) = \exp\left[\int_0^t g'(\tilde{X}^h(s))dY(s) - \frac{1}{2}\int_0^t \left|g(\tilde{X}^h(s))\right|^2 ds\right]. \tag{2.6}$$

For Markov $\tilde{X}^h(\cdot)$, the approximating filter $\Pi^h(\cdot)$ is defined by

$$\langle \Pi^h(t), \phi\rangle = \frac{E_{\{\Pi^h(0),Y_{0,t}\}}\left[\phi(\tilde{X}^h(t))R(\tilde{X}^h_{0,t}, Y_{0,t})\right]}{E_{\{\Pi^h(0),Y_{0,t}\}}R(\tilde{X}^h_{0,t}, Y_{0,t})} \tag{2.7}$$

and $\Pi^h(\cdot)$ satisfies the semigroup equation

$$\langle \Pi^h(t+s), \phi\rangle = \frac{E_{\{\Pi^h(t),Y_{t,t+s}\}}\left[\phi(\tilde{X}^h(s))R(\tilde{X}^h_{0,s}, Y_{t,t+s})\right]}{E_{\{\Pi^h(t),Y_{t,t+s}\}}R(\tilde{X}^h_{0,s}, Y_{t,t+s})}, \quad s > 0,\ t \geq 0. \tag{2.8}$$

According to our standard notation, the initial distribution of $\tilde{X}^h(\cdot)$ in (2.7) is $\Pi^h(0)$ and in (2.8) it is $\Pi^h(t)$.

When $\tilde{X}^h(\cdot)$ is piecewise constant with $\{\tilde{X}^h(n\delta_h); n \geq 1\}$ being Markov, then the approximating filter is defined by (2.7) and (2.8), but where t and s are integral multiples of δ_h and $\Pi^h(\cdot)$ is constant on the intervals $[n\delta_h, n\delta_h + \delta_h)$. Thus, the evolution of $\Pi^h(\cdot)$ can be written in recursive form in general.

From the Feller–Markov property of $X(\cdot)$, it follows easily that the pair $(X(\cdot), \Pi(\cdot))$ is Feller–Markov. Since $X(t)$ takes values in a compact set G, and the samples of $\Pi(t)$ are measures with support in G, we have at least one invariant measure for the Markov family determined by $(X(\cdot), \Pi(\cdot))$. When initialized at this measure, the process $(X(\cdot), \Pi(\cdot))$ is stationary; i.e., the distribution of

$$(X(t+\cdot), \Pi(t+\cdot))$$

does not depend on t. Let $\bar{Q}(\cdot)$ denote the measure of the joint process

$$\Psi(\cdot) = (X(\cdot), \Pi(\cdot), Y(\cdot), B(\cdot), W(\cdot)),$$

where $(X(\cdot), \Pi(\cdot))$ is stationary. Let $\bar{Q}_f(\cdot)$ denote the measure of the stationary joint process $(X(\cdot), \Pi(\cdot))$. We will make the following fundamental assumption throughout this paper.

ASSUMPTION 2.3. *There is a unique invariant measure for* $(X(\cdot), \Pi(\cdot))$.

In section 7 we provide conditions for the uniqueness of the invariant measure to hold and provide some examples for which these conditions can be verified. The critical importance of the uniqueness of the stationary joint process was first raised in [21].

The stationary process and associated measure will be given a filtering interpretation later in this section.

In the following, lower case letters $x(\cdot), \pi(\cdot)$, etc., are used for the canonical sample paths. Letters such as $x, y, \ldots$ are used to denote vectors such as $x(t), y(t)$, etc. Define $\psi_f(\cdot) = (x(\cdot), \pi(\cdot))$, $\Psi_f(\cdot) = (X(\cdot), \Pi(\cdot))$, and $\psi(\cdot) = (x(\cdot), \pi(\cdot), y(\cdot), b(\cdot), w(\cdot))$, although we will not always need the component $w(\cdot)$.

For each $t \geq 0$, define the process

$$\Psi^h(t, \cdot) = \left(X(t + \cdot), \Pi^h(t + \cdot), Y(t + \cdot) - Y(t), B(t + \cdot) - B(t), W(t + \cdot) - W(t)\right).$$

Define $\Psi_f^h(t, \cdot)$ similarly. In sections 4 and 6 in which we consider more general approximate filtering problems, we will consider modifications to the $\Psi^h(t, \cdot)$ process by replacing $X(\cdot), Y(\cdot), B(\cdot)$ with h-parametrized processes $X^h(\cdot), Y^h(\cdot), B^h(\cdot)$.

The path spaces. The path space $D[\mathbb{R}^k; 0, \infty)$ of $\mathbb{R}^k$-valued functions which are right continuous and have left hand limits (CADLAG), endowed with the Skorohod topology [3, 9], will be used for the vector-valued processes such as $X(\cdot), Y(\cdot), B(\cdot), \tilde{X}(\cdot)$, etc., for the appropriate value of k. Let $\mathcal{M}(G)$ denote the space of measures on G with the weak topology. Let $m_n(\cdot)$ and $m(\cdot)$ be in $\mathcal{M}(G)$. Recall that $m_n(\cdot)$ converges weakly to $m(\cdot)$ if for each bounded and continuous function $\phi(\cdot)$ on G, $\langle m_n, \phi\rangle \to \langle m, \phi\rangle$. Equivalently, for a set of continuous functions $\{\phi_i(\cdot)\}$ which are dense (in the topology of uniform convergence) in the set of bounded and continuous functions on G, we have the metric

$$d(m_n, m) = \sum_i 2^{-i} |\langle m_n - m, \phi_i\rangle| \to 0.$$

Owing to the fact that the values of $X(t)$ and $\tilde{X}^h(t)$ are in G, the samples of the $\Pi(t)$ and its approximations $\Pi^h(\cdot)$ take values in $\mathcal{M}(G)$. The path $\Pi(\cdot)$ and its approximations will take values in the CADLAG space $D[\mathcal{M}(G); 0, \infty)$, also with the Skorohod topology used. The measure-valued random variables $Q^{h,T}(\cdot)$ defined by (2.9) below take values in the space of measures on the product path space

$$\mathcal{M}\left(D[\mathbb{R}^k; 0, \infty) \times D[\mathcal{M}(G); 0, \infty)\right)$$

for the appropriate value of k (which is the sum of the dimensions of x, y, b, w).

A "prediction error" function. We start with a special case of the error or performance function in order to fix ideas. Let $\phi(\cdot)$ be a bounded, continuous, and real-valued function of x and consider the pathwise average error per unit time

$$G^{h,T}(\phi) \equiv \frac{1}{T} \int_0^T \left[\langle \Pi^h(t), \phi\rangle - \phi(X(t))\right]^2 dt. \tag{2.9}$$

We will show that, under quite broad conditions on the approximate filter,

$$G^{h,T}(\phi) \to \int \left[\langle \pi(0), \phi\rangle - \phi(x(0))\right]^2 \bar{Q}_f(d\psi_f(\cdot)) \tag{2.10}$$

in the sense of probability, as $h \to 0$ and $T \to \infty$, in any way at all. Later in this section, it will be seen that the right side of (2.10) is what one would also get as the limit if the true optimal filter were used and not the approximation. In this sense there is pathwise asymptotic optimality. The result will actually be much more general. The errors will be written in terms of the following occupation measure.

The occupation measures. For a random variable Z and set A, let $I_A(Z)$ denote the indicator function of the event that $Z \in A$. Let C and C' be measurable sets in the product path spaces of $\Psi^h(t,\cdot)$ and $\Psi^h_f(t,\cdot)$, resp. Define the occupation measures $Q^{h,T}(\cdot)$ and $Q^{h,T}_f(\cdot)$ as

$$Q^{h,T}(C) = \frac{1}{T}\int_0^T I_C(\Psi^h(t,\cdot))dt \tag{2.11}$$

and

$$Q^{h,T}_f(C') = \frac{1}{T}\int_0^T I_{C'}(\Psi^h_f(t,\cdot))dt. \tag{2.12}$$

We can write (2.9), in terms of $Q^{h,T}_f(\cdot)$, as

$$G^{h,T}(\phi) = \int [\langle \pi(0), \phi\rangle - \phi(x(0))]^2\, Q^{h,T}_f(d\psi_f(\cdot)). \tag{2.13}$$

A more general error or performance function. It will be shown that for any (bounded, measurable, and continuous w.p.1 with respect to the measure $\bar{Q}_f(\cdot)$) real-valued function $F(\cdot)$

$$\frac{1}{T}\int_0^T F(\Psi^h_f(t))dt = \int F(\psi_f(\cdot))Q^{h,T}_f(d\psi_f(\cdot)) \to \int F(\psi_f(\cdot))\bar{Q}_f(d\psi(\cdot)) \tag{2.14}$$

in probability. This says that sample mean errors of many types (see the example in the next paragraph) will converge to the stationary value, which is the same value that one would get if the true optimal filter were used, and the pathwise occupation measure replaced by its expected value. The convergence is in the sense of probability, and holds as $T \to \infty$ and $h \to 0$ in any way at all. The arbitrariness of the way that $T \to \infty$ and $h \to 0$ is crucial in applications. It is important that the approximation is good for all small h, not depending on T, if T is large enough. Results such as (2.14) are the main contributions of this paper.

The function $F(\cdot)$ can be quite general. In (2.10), $F(\psi_f(\cdot))$ depends only on $\psi_f(0)$, since it equals $|\langle \pi(0), \phi\rangle - \phi(x(0))|^2$. But, $F(\cdot)$ might have a more complex dependence on $\psi_f(\cdot)$. For example, consider the case where we are interested in

$$\frac{1}{T}\int_0^T \left[\max_{s\in[t-1,t]} \left|\phi(X(s)) - \langle \Pi^h(s), \phi\rangle\right|\right] dt.$$

Here

$$F(\psi_f(\cdot)) = \max_{s\le 1} |\phi(x(s)) - \langle \pi(s), \phi\rangle|.$$

Comment on the stationary process. Let us deviate a little and discuss the meaning of a stationary process $(X(\cdot), \Pi(\cdot))$. Recall that we assume that the measure $\bar{Q}_f(\cdot)$ of the *stationary pair* $(X(\cdot), \Pi(\cdot))$ is unique. Now, consider the *correctly initialized* filter, i.e., the true optimal filter, which we denote by $\Pi_0(\cdot)$, i.e., $\Pi_0(B) = P\{X(0) \in B\}$ for each Borel set $B \in \mathbb{R}^r$. Define the mean occupation measure $Q^T_f(\cdot)$ by

$$Q^T_f(C) = \frac{1}{T}\int_0^T P\left((X(t+\cdot), \Pi_0(t+\cdot)) \in C\right) dt, \tag{2.15}$$

for measurable sets C in the product path space. Note that $Q_f^T(\cdot)$ is a measure but it is not random. It follows from the proofs of Theorems 3.1 and 3.2 that $\{Q_f^T(\cdot); T < \infty\}$ is tight and converges to $\bar{Q}_f(\cdot)$, the measure of the unique stationary process. Thus, the limit in (2.14) is the same as if the optimal filter were used (with probability one) and the pathwise occupation measure were replaced by the mean occupation measure.

Note that the integrand in (2.14) is the measure of the (system, filter) pair on the time interval $[t, \infty)$, and the integral is equivalent to choosing t at random in $[0, T]$. This gives a loose interpretation of $\bar{Q}_f(\cdot)$ in terms of a random initial time.

3. Tightness and weak convergence.

The following lemma will be used in the tightness arguments to follow.

LEMMA 3.1 (see [14, Theorem 2.7b]). *Let $Z_{n,s}(\cdot)$, $n = 1, 2, \ldots; s > 0$ be a family of processes with paths in the Skorohod space $D[S_0; 0, \infty)$, where S_0 is a complete and separable metric space with metric $\gamma(\cdot)$. For each $\delta > 0$ and each t in a dense set, let there be a compact set $S_{\delta,t} \subset S_0$ such that*

$$\sup_{n,s} P\{Z_{n,s}(t) \notin S_{\delta,t}\} \leq \delta. \tag{3.1}$$

Let $\mathcal{F}_t^{n,s}$ denote the minimal σ-algebra which measures $\{Z_{n,s}(u), u \leq t\}$, and $\mathcal{T}_{n,s}(T)$ the set of $\mathcal{F}_t^{n,s}$-stopping times which are less than $T > 0$. Suppose that for each T

$$\lim_{\delta \to 0} \limsup_n \sup_s \sup_{\tau \in \mathcal{T}_n(T)} E\left[\gamma\left(Z_{n,s}(\tau + \delta), Z_{n,s}(\tau)\right) \wedge 1\right] = 0. \tag{3.2}$$

Then to every $\epsilon > 0$ there exists a compact set K_ϵ such that

$$\limsup_n \sup_s P(Z_{n,s}(\cdot) \in K_\epsilon) > 1 - \epsilon.$$

We will apply the above theorem for the doubly parametrized process $\{\Pi^h(t + \cdot); h > 0, t > 0\}$. Condition (3.1) will hold trivially due to the compactness of the space and we will concentrate on verifying (3.2).

THEOREM 3.1. *Assume that Assumptions 2.1, 2.2, and 2.3 hold. Then*

$$\{X(t + \cdot); t \geq 0\} \text{ is tight,} \tag{3.3a}$$

$$\left\{\tilde{X}^h(\cdot); h, \text{ all possible initial conditions}\right\} \text{ is tight.} \tag{3.3b}$$

To every $\epsilon > 0$ there exists a compact set K_ϵ such that

$$\limsup_{h \to 0} \sup_t P(\Pi^h(t + \cdot) \in K_\epsilon) > 1 - \epsilon. \tag{3.3c}$$

Also for every sequence $h_k \to 0$, $T_k \to \infty$,

$$\left\{Q^{h_k, T_k}(\cdot); k \geq 1\right\} \text{ is tight.} \tag{3.4}$$

Let $Q(\cdot)$ denote a weak sense limit of the set in (3.4), as $k \to \infty$. Let ω be the canonical variable on the probability space on which $Q(\cdot)$ is defined, and denote the samples by $Q^\omega(\cdot)$. Then, for each ω, $Q^\omega(\cdot)$ is a measure on the product path space $D[\mathbb{R}^k; 0, \infty) \times D[\mathcal{M}(G); 0, \infty)$ for appropriate k. It induces a process

$$\Psi^\omega(\cdot) = (X^\omega(\cdot), \Pi^\omega(\cdot), Y^\omega(\cdot), B^\omega(\cdot), W^\omega(\cdot)). \tag{3.5}$$

For almost all ω, $(X^\omega(\cdot), \Pi^\omega(\cdot))$ is stationary.

Proof. The set $\{X(t+\cdot), t \geq 0\}$ is tight, by the fact that the set of "initial conditions" $\{X(t), t > 0\}$ are confined to a compact set and the properties of the diffusion (2.1). The set in (3.3b) is tight by the assumptions on $\tilde{X}^h(\cdot)$ concerning weak convergence, which (using the fact that the possible values of $\tilde{X}^h(0)$ are confined to some compact set) implies that every subsequence of the set in (3.3b) has a further subsequence which converges weakly.

To prove (3.3c) we will apply Lemma 3.1. For a real number $\rho > 0$ and each fixed h and t, let $\mathcal{T}^{h,t}(\rho)$ denote the set of stopping times with respect to the filtration $\sigma\{\Pi^h(t+s) : s \leq u\}_{u \geq 0}$, bounded by ρ. Then, by Lemma 3.1, it is sufficient to show that

$$\lim_{\delta\to 0} \sup_{h\to 0} \sup_{t} \sup_{\tau\in\mathcal{T}^{h,t}(\rho)} E\left|\langle \Pi^h(t+\tau+\delta), \phi\rangle - \langle \Pi^h(t+\tau), \phi\rangle\right| = 0 \tag{3.6}$$

for each continuous, bounded, and real-valued $\phi(\cdot)$. Define the function

$$K(h,t,\tau,\delta) = \frac{E_{\{\Pi^h(t+\tau), Y_{t+\tau,t+\tau+\delta}\}}\left[\phi(\tilde{X}^h(\delta)) R(\tilde{X}^h_{0,\delta}, Y_{t+\tau,t+\tau+\delta})\right]}{E_{\{\Pi^h(t+\tau), Y_{t+\tau,t+\tau+\delta}\}} R(\tilde{X}^h_{0,\delta}, Y_{t+\tau,t+\tau+\delta})}.$$

Owing to the use of the weak topology, the boundedness of $\phi(\cdot)$, and the semigroup property (2.18), proving that (3.6) is equivalent to proving that

$$\lim_{\delta\to 0} \sup_{h\to 0} \sup_{t} \sup_{\tau\in\mathcal{T}^{h,t}(\rho)} E\left|K(h,t,\tau,\delta) - \langle \Pi^h(t+\tau), \phi\rangle\right| = 0.$$

But showing this is straightforward, owing to the convergence to unity of the exponential function $R(\tilde{X}^h_{0,\delta}, Y_{t+\tau,t+\tau+\delta})$ as $\delta \to 0$. In particular, it is sufficient to show that (where d_s denoted the differential with respect to the s variable)

$$\begin{aligned} &E_{\{\Pi^h(t+\tau), Y_{t+\tau,t+\tau+\delta}\}} \phi(\tilde{X}^h(\delta)) \\ &\quad \times \left\{ \exp\left[\int_0^\delta g'(\tilde{X}^h(s)) d_s Y(t+\tau+s) - \frac{1}{2}\int_0^\delta \left| g(\tilde{X}^h(s))\right|^2 ds \right] - 1 \right\} \end{aligned}$$

goes to zero in mean as $\delta \to 0$, uniformly in (h,t,τ), for each bounded and continuous function $\phi(\cdot)$. But this follows from the boundedness of $g(\cdot)$. This proves (3.3c).

Since

$$\{B(t+\cdot) - B(t), W(t+\cdot) - W(t), Y(t+\cdot) - Y(t), t > 0\}$$

is always tight, (3.3c) implies that to every $\epsilon > 0$ there exists a compact set K_ϵ such that

$$\lim_{h} \sup_{t} \sup P(\Psi^h(t+\cdot) \in K_\epsilon) > 1 - \epsilon.$$

Now standard arguments show that the family in (3.4) is tight; for more detail see [19, Chapter 1, section 6] or [22, section 5], where there are general results on getting the tightness of occupation measures from the tightness of the set of time shifted processes.

The proofs in [22, Theorem 6.3, part 2] or in [19, Chapter 5, Theorem 2.2] can be adapted to show the stationarity. We work with the marginals $Q_f^{h,T}(\cdot)$ and a subsequence (h, T) such that $\{Q_f^{h,T}(\cdot)\}$ converges weakly. Let H be a Borel set in the product $(x(\cdot), \pi(\cdot))$ path space. For $c > 0$, define the left shift H_c by $H_c = \{\psi_f(\cdot) : \psi_f(c, \cdot) \in H\}$. Then

$$Q_f^{h,T}(H_c) = \frac{1}{T}\int_0^T I_{H_c}(\Psi_f^h(t, \cdot))dt = \frac{1}{T}\int_0^T I_H(\Psi_f^h(t + c, \cdot))dt,$$

$$Q_f^{h,T}(H_c) - Q_f^{h,T}(H) = \frac{1}{T}\int_T^{T+c} I_H(\Psi_f^h(t, \cdot))dt - \frac{1}{T}\int_0^c I_H(\Psi_f^h(t, \cdot))dt. \tag{3.7}$$

The difference goes to zero as $T \to \infty$ for all ω, c, h, and H. This and the weak convergence yields the stationarity of the limit processes $(X^\omega(\cdot), \Pi^\omega(\cdot))$ for almost all ω. □

We now need to identify the processes associated with the limits of the occupation measures and prove the main result (2.10).

Theorem 3.2. *Assume that Assumptions* 2.1, 2.2, *and* 2.3 *hold. Then, for almost all* ω, *the following hold.*

(i) $(B^\omega(\cdot), W^\omega(\cdot))$ *are standard Wiener processes, with respect to which* $(X^\omega(\cdot)$, $\Pi^\omega(\cdot)$, $Y^\omega(\cdot))$ *are nonanticipative.*

(ii)

$$dY^\omega = g(X^\omega)dt + dB^\omega, \tag{3.8}$$

(iii)

$$dX^\omega = p(X^\omega)dt + \sigma(X^\omega)dW^\omega. \tag{3.9}$$

(iv) *For each bounded and measurable real-valued function* $\phi(\cdot)$

$$\langle \Pi^\omega(t), \phi\rangle = \frac{E_{\{\Pi^\omega(0), Y_{0,t}^\omega\}}\left[\phi(\tilde{X}(t))R(\tilde{X}_{0,t}, Y_{0,t}^\omega)\right]}{E_{\{\Pi^\omega(0), Y_{0,t}^\omega\}}R(\tilde{X}_{0,t}, Y_{0,t}^\omega)}. \tag{3.10}$$

(v) $(X^\omega(\cdot), \Pi^\omega(\cdot))$ *is the unique stationary process, and hence its distribution does not depend on* ω.

Finally, for any bounded and measurable real-valued function $F(\cdot)$ *which is continuous almost everywhere with respect to the unique measure* $\bar{Q}_f(\cdot)$ *of the stationary process,*

$$\frac{1}{T}\int_0^T F(\Psi_f^h(t + \cdot))dt \to \int F(\psi_f(\cdot))\bar{Q}_f(d\psi_f(\cdot)) \tag{3.11}$$

in probability as $h \to 0$ *and* $T \to \infty$ *in any way at all.*

Proof. Let $h_k \to 0, T_k \to \infty$. Fix a weakly convergent subsequence of $\{Q^{h_k,T_k}(\cdot)\}_{k\geq 1}$, and index it by h_k, T_k also, abusing the notation. It will turn out that all convergent subsequences have the same limit.

The processes $B^\omega(\cdot), W^\omega(\cdot)$. Let us demonstrate the Wiener and mutual independence properties (for almost all ω). Let $f_i(\cdot)$ and $q(\cdot)$ be real-valued and continuous functions of their arguments and with compact support, with the $f_i(\cdot)$ being twice continuously differentiable. Let $\phi_j(\cdot)$ be an arbitrary finite collection of bounded and continuous real-valued functions. Let ρ and τ be arbitrary nonnegative numbers, and let $u_i \leq \rho$ be (a finite collection of) arbitrary nonnegative real numbers. Define the function

$$\begin{aligned} F(\psi(\cdot)) &= q(x(u_i), \langle \pi(u_i), \phi_j \rangle, b(u_i), w(u_i); i, j) \\ &\times \left[f_1(b(\rho+\tau)) f_2(w(\rho+\tau)) - f_1(b(\rho)) f_2(w(\rho)) - \int_\rho^{\rho+\tau} A_0 f_1(b(s)) f_2(w(s))) ds \right], \end{aligned}$$

where $A_0 = (1/2)\left[\sum_i \partial^2/\partial b_i^2 + \sum_i \partial^2/\partial w_i^2\right]$. The b_i (resp., w_i) are the scalar components of b (resp., w).

We will show that

$$\int Q^\omega(d\psi(\cdot))F(\psi(\cdot)) = 0 \tag{3.12}$$

for almost all ω. The arbitrariness of the functions and of $\rho, \tau, u_i \leq \rho$ and (3.12) imply that $W^\omega(\cdot)$ and $B^\omega(\cdot)$ are standard vector-valued and independent Wiener processes, and martingales with respect to the filtration generated by $\{\Psi^h(s), s \leq t\}$, hence the nonanticipativity.

To get (3.12), we will show that the mean square value

$$E\left[\int Q^{h_k,T_k}(d\psi(\cdot))F(\psi(\cdot))\right]^2 \tag{3.13}$$

goes to zero as h, T go to their limits. This and the weak convergence imply (3.12) since they imply that $E\left[\int Q(d\psi(\cdot))F(\psi(\cdot))\right]^2 = 0$. (An analogous calculation for a queueing problem is in [22].) By the definition of the occupation measure, the expression

$$\int Q^{h_k,T_k}(d\psi(\cdot))F(\psi(\cdot)) \tag{3.14a}$$

equals

$$\frac{1}{T_k}\int_0^{T_k} F_1(t)dt, \tag{3.14b}$$

where $F_1(t)$ equals $F_2(t)F_3(t)$, and where

$$F_2(t) = q\left(X(t+u_i), \langle \Pi^{h_k}(t+u_i), \phi_j \rangle, B(t+u_i) - B(t), W(t+u_i) - W(t); i, j\right),$$

$$\begin{aligned} F_3(t) = \Bigg[& f_1(B(t+\rho+\tau) - B(t)) f_2(W(t+\rho+\tau) - W(t)) \\ & - f_1(B(t+\rho) - B(t)) f_2(W(t+\rho) - W(t)) \\ & - \int_{t+\rho}^{t+\rho+\tau} A_0 f_1(B(t+s) - B(t)) f_2(W(t+s) - W(t))) ds \Bigg]. \end{aligned}$$

The mean square value of (3.14b) equals

$$\frac{1}{T_k^2}\int_0^{T_k}\int_0^{T_k} EF_2(t)F_2(v)F_3(t)F_3(v)dtdv. \tag{3.15}$$

Now, using the fact that (the martingale property)

$$E\left[F_3(v)|W(u),B(u),u\le t+\rho\right]=0 \quad \text{for } v\ge t, \tag{3.16}$$

we see that (3.15) is of the order of $O(1/T_k)$, which yields the desired result.

Identifying the limits $X^\omega(\cdot), Y^\omega(\cdot)$. For a vector z, define $|z|_1^2=|z|^2\wedge 1$. To identify the $Y^\omega(\cdot)$ process, we will prove that

$$E\int Q^\omega(d\psi(\cdot))\left|y(s)-\int_0^s g(x(u))du-b(s)\right|_1^2=0 \tag{3.17}$$

for each s. This is equivalent to

$$EE^\omega\left|Y^\omega(s)-\int_0^s g(X^\omega(u))du-B^\omega(s)\right|_1^2=0,$$

which implies (3.8) for almost all ω.

Equation (3.17) is shown by using the definition of the occupation measure and the weak convergence, i.e., by using the fact that (2.2) implies

$$\begin{aligned}&E\int Q^{h_k,T_k}(d\psi(\cdot))\left|y(s)-\int_0^s g(x(u))du-b(s)\right|_1^2\\&\quad=E\frac{1}{T_k}\int_0^{T_k}\left|Y(t+s)-Y(t)-\int_t^{t+s}g(X(u))du-(B(t+s)-B(t))\right|_1^2dt=0.\end{aligned}$$

Now use the weak convergence and the fact that the integrand in the first integral is a bounded and continuous function of $\psi(\cdot)$.

We deal with $X(\cdot)$ by a process approximation method. Recall the definitions of $p(\cdot)$ and $\sigma(\cdot)$ from (2.1). For each $\epsilon>0$ and $t\ge 0$, there is $\Delta_0>0$ such that for $\Delta\le\Delta_0$ we have (uniformly in $u\ge 0$)

$$\begin{aligned}E\Bigg|&X(u+t)-X(u)-\int_u^{u+t}p(X(s))ds\\&-\sum_{i:i\Delta<t}\sigma(X(u+i\Delta))\left[W(u+i\Delta+\Delta)-W(u+i\Delta)\right]\Bigg|^2=O(\epsilon).\end{aligned} \tag{3.18}$$

We have shown above that $(W^\omega(\cdot),B^\omega(\cdot))$ are Wiener processes, and that $(X^\omega(\cdot),\Pi^\omega(\cdot))$ is nonanticipative with respect to them. Suppose that we can show that for each $\epsilon>0$ and $t\ge 0$ there is $\Delta_\epsilon>0$ such that for $\Delta\le\Delta_\epsilon$,

$$\begin{aligned}E\int\Bigg[&x(t)-x(0)-\int_0^t p(x(s))ds\\&-\sum_{i:i\Delta<t}\sigma(x(i\Delta))[w(i\Delta+\Delta)-w(i\Delta)]\Bigg]_1^2Q^\omega(d\psi(\cdot))=O(\epsilon).\end{aligned} \tag{3.19}$$

Then, analogous to what was done for $Y^\omega(\cdot)$ above, we have, for almost all ω,

$$X^\omega(t)=X^\omega(0)+\int_0^t p(X^\omega(s))ds+\sum_{i:i\Delta<t}\sigma(X^\omega(i\Delta))[W^\omega(i\Delta+\Delta)-W^\omega(i\Delta)]$$

modulo $O(\epsilon)$ in mean square. This and the arbitrariness of ϵ and Δ imply (3.9). The reason for using the finite sum approximation to the stochastic integral is that now the integrand in (3.19) is a bounded and continuous function of $\psi(\cdot)$, w.p.1 for each ω, since the $Q^\omega(\cdot)$ must induce continuous processes. (If $X(\cdot)$ is jump-diffusion, the continuity w.p.1 still holds since the probability that a jump occurs at any given (nonrandom) time point is zero.) Thus, the weak convergence can be used.

Consider now

$$E\int \left| x(t) - x(0) - \int_0^t p(x(s))ds - \sum_{i:i\Delta<t} \sigma(x(i\Delta))[w(i\Delta+\Delta) - w(i\Delta)] \right|_1^2 Q^{h_k,T_k}(d\psi(\cdot)).$$

The h index is irrelevant in this calculation. By the definition of the occupation measure, the above expression is equivalent to

$$E\frac{1}{T_k}\int_0^{T_k} \left| X(t+u) - X(u) - \int_u^{t+u} p(X(s))ds \right.$$
$$\left. - \sum_{i:i\Delta<t} \sigma(X(u+i\Delta))[W(u+i\Delta+\Delta) - W(u+i\Delta)] \right|_1^2 du.$$

But, (3.18) implies that, for small enough $\Delta > 0$, the expectation of this is $O(\epsilon)$ uniformly in T. This and the weak convergence of $Q^{h,T}(\cdot)$ to $Q(\cdot)$ imply (3.19).

The $\Pi^\omega(\cdot)$ component. Now we turn our attention to $\Pi^h(\cdot)$ and the sample value $\Pi^\omega(\cdot)$ component of the weak sense limit. For small $\Delta > 0$, define $R^\Delta(\tilde{X}^h_{0,s}, Y_{0,s})$ by approximating the stochastic integral $\int_0^s g'(\tilde{X}^h(u))dY(u)$ in the exponent by the sum

$$\sum_{i:i\Delta<s} g'(\tilde{X}^h(i\Delta))\,[Y(i\Delta+\Delta) - Y(i\Delta)]$$

analogously to what was done in (3.18). For arbitrary t define $R^\Delta(\tilde{X}^h_{0,s}, Y_{t,t+s})$ analogously. For each h, t, define the difference $\hat{Z}^{h,\Delta}(t,\cdot)$ by

$$\hat{Z}^{h,\Delta}(t,s) = \frac{E_{\{\Pi^h(t),Y_{t,t+s}\}}\left[\phi(\tilde{X}^h(s))R(\tilde{X}^h_{0,s}, Y_{t,t+s})\right]}{E_{\{\Pi^h(t),Y_{t,t+s}\}}R(\tilde{X}^h_{0,s}, Y_{t,t+s})}$$
$$- \frac{E_{\{\Pi^h(t),Y_{t,t+s}\}}\left[\phi(\tilde{X}^h(s))R^\Delta(\tilde{X}^h_{0,s}, Y_{t,t+s})\right]}{E_{\{\Pi^h(t),Y_{t,t+s}\}}R^\Delta(\tilde{X}^h_{0,s}, Y_{t,t+s})}.$$

We note that $\{\tilde{X}^h(\cdot)\}$ is equicontinuous in probability in the following sense. For each positive ρ and t, there is $h_0 > 0$ and $\alpha > 0$ such that

$$\sup_{h\le h_0} P\left\{ \sup_{u\le\alpha,\ s\le t} \left|\tilde{X}^h(s+u) - \tilde{X}^h(s)\right| \ge \rho \right\} \le \rho.$$

By the properties of the approximation of the stochastic integral by a sum, for each $\epsilon' > 0$ and $\epsilon'' > 0$, there is Δ_0 such that for $\Delta < \Delta_0$, and small $h > 0$, we have

$$\sup_{\Pi^h(t),t} P\left\{ \left|\hat{Z}^{h,\Delta}(t,s)\right| \ge \epsilon' \right\} \le \epsilon''. \tag{3.20}$$

This is due to the just-cited equicontinuity in probability, keeping in mind that the state space is compact, and where the sup is over all possible values. Next we verify that (3.10) holds. We would like to show that, for each s,

$$E\int Q^{\omega}(d\psi(\cdot))\left|\langle\pi(s),\phi\rangle-\frac{E_{\{\pi(0),y_{0,s}\}}\left[\phi(\tilde{X}(s))R(\tilde{X}_{0,s},y_{0,s})\right]}{E_{\{\pi(0),y_{0,s}\}}R(\tilde{X}_{0,s},y_{0,s})}\right|_1^2=0. \tag{3.21}$$

But the integrand function is not well defined, since $y(\cdot)$ is an arbitrary path function and the stochastic integral in $R(\cdot)$ in (3.21) is not defined. The problem is resolved by working with an approximation.

It will actually be shown that

$$E\int Q^{\omega}(d\psi(\cdot))\left|\langle\pi(s),\phi\rangle-\frac{E_{\{\pi(0),y_{0,s}\}}\left[\phi(\tilde{X}(s))R^{\Delta}(\tilde{X}_{0,s},y_{0,s})\right]}{E_{\{\pi(0),y_{0,s}\}}R^{\Delta}(\tilde{X}_{0,s},y_{0,s})}\right|_1^2 \tag{3.22}$$

goes to zero as $\Delta\to 0$, which implies (3.10). Here the $\tilde{X}(\cdot)$ is a process with initial distribution $\pi(0)$ and the law of evolution of $X(\cdot)$. Note that the integrand function in (3.22) does not depend on this process, only on its initial condition. Also it is bounded and continuous w.p.1 (with respect to $Q^{\omega}(\cdot)$) for almost all ω.

By the weak convergence, (3.22) is the limit of

$$E\int Q^{h_k,T_k}(d\psi(\cdot))\left|\langle\pi(s),\phi\rangle-\frac{E_{\{\pi(0),y_{0,s}\}}\left[\phi(\tilde{X}^{h_k}(s))R^{\Delta}(\tilde{X}^{h_k}_{0,s},y_{0,s})\right]}{E_{\{\pi(0),y_{0,s}\}}R^{\Delta}(\tilde{X}^{h_k}_{0,s},y_{0,s})}\right|_1^2.$$

Thus, we need only show that, for each $s>0$ and each bounded and continuous function $\phi(\cdot)$, this expression is small for small $\Delta>0, h_k>0$ and large $T_k<\infty$. By the definition of the measure $Q^{h,T}(\cdot)$, this expression equals

$$E\frac{1}{T_k}\int_0^{T_k}\left|\hat{Z}^{h_k,\Delta}(t,s)\right|_1^2 dt. \tag{3.23}$$

Now, the fact that (3.23) is small for small h_k,Δ follows from (3.20).

Theorem 3.1 yields the stationarity of $(X^{\omega}(\cdot),\Pi^{\omega}(\cdot))$ for almost all ω. The measure $\bar{Q}_f(\cdot)$ of the stationary $(X^{\omega}(\cdot),\Pi^{\omega}(\cdot))$ process is unique by assumption. Thus $Q_f^{h_k,T_k}(\cdot)$ converges weakly to $\bar{Q}_f(\cdot)$ along the chosen subsequence. The uniqueness also implies that the subsequence is irrelevant and that $Q_f^{h,T}(\cdot)$ converges weakly to $\bar{Q}_f(\cdot)$ along any (h,T) subsequence. This implies (3.11). $\square$

4. Non-Markov signal processes. In actual physical applications, the signal process would not generally be a diffusion or jump-diffusion, but rather some non-Markov process which is only approximated by a diffusion or jump-diffusion. For example, the physical process might be a dynamical system which is driven by a wide bandwidth noise process and not a Wiener process. To quantify the connection, we need to parameterize the physical process, and suppose that it converges weakly to some "ideal" signal process, a diffusion or jump-diffusion, as the approximation parameter h goes to its limit. We would construct an approximation to the optimal filter for the ideal signal process, but use the physical observations. Thus, we are concerned with the convergence of the pathwise average errors as the time interval goes to its infinite limit; the physical process is better and better approximated by its ideal limit and the filter approximation converges to the optimal filter for that ideal limit. Thus, we have a physical process $X^h(\cdot)$, with values on the same compact range space G and which is approximated by the diffusion $X(\cdot)$, defined by (2.1), and a filter $\Pi^h(\cdot)$ which is a numerical approximation to the optimal filter for (2.1) with observations (2.2), but where the actual physical observations defined by

$$dY^h = g(X^h(t))dt + dB \tag{4.1}$$

are used. We could use another parameter rather than h to index the "prediffusion" process and then let this go to its limit together with h, T, but there is no loss of generality in using h to index this process as well.

The filter to be built is as in section 2. This would be a filter constructed under the assumption that the system is (2.1) and observations (2.2). But, with (4.1) actually used. The approximating filter uses the auxiliary process $\tilde{X}^h(\cdot)$, which satisfies (A2.2). Thus the approximating filter takes the form (2.7), but with $Y^h(\cdot)$ replacing $Y(\cdot)$. The process $X^h(\cdot)$ is not to be confused with $\tilde{X}^h(\cdot)$. The former is the actual physical signal process. The latter continues to play the role that it had in the previous sections; i.e., it is an approximation to the ideal limit, which is used only to get a numerical approximation to the optimal filter for (2.1) and (2.2). The following is the basic assumption on the physical process $X^h(\cdot)$.

ASSUMPTION 4.1. *$X^h(t) \in G$ for all $t, h > 0$. The family $\{X^h(t+\cdot); h > 0, t \geq 0\}$ is tight and whenever for some subsequence $(h_k, t_k)_{k\geq 1}$, where $h_k \to 0$ as $k \to \infty$, the distribution of $X^{h_k}(t_k)$ converges weakly to some measure* Π *and the sequence $\{X^{h_k}(t_k + \cdot)\}_{k\geq 1}$ is weakly convergent, then the limit of the latter is $X(\cdot)$ with initial distribution* Π.

If the filter is to operate over some fixed finite time interval, then convergence results are relatively easy to obtain, since the criteria of interest in that case are mean values. Many such results are in [21]. In the definition of $\Psi^h(\cdot)$, drop the W-component and replace $X(\cdot), Y(\cdot)$, by $X^h(\cdot), Y^h(\cdot)$, resp. Redefine the $\Psi(\cdot), \psi(\cdot)$, $Q^{h,T}(\cdot)$ analogously. In Theorem 4.1 we obtain the analog of Theorems 3.1 and 3.2 for this approximate filtering problem. Theorem 3.1 continues to hold, with the same proof. In the proof of the analog of Theorem 3.2, there is no new problem with the identification of the limit processes $Y^\omega(\cdot)$ and $\Pi^\omega(\cdot)$. But there is a problem in the identification of the limit $X^\omega(\cdot)$. The assumption of weak convergence on $X^h(t + \cdot)$ made above is not enough. The problem is that the $Q^{h,T}(\cdot)$ are occupation measures, whose values depend on the *samples* of the paths. Owing to this we need to make a (quite unrestrictive) assumption so that a martingale type method can be used to identify $X^\omega(\cdot)$. Theorem 3.2 can be viewed as an ergodic theorem or weak law of large numbers, and it works partly due to the ergodic properties of $X(\cdot)$, as reflected in the uniqueness of the stationary process. But the assumed weak convergence of $X^h(t+\cdot)$ is not enough in itself to get such a weak law of large numbers. It says little about "long range dependence." This is the main issue that we will need to deal with.

A note on the martingale problem method. In order to motivate the assumptions to be used, we first recall the classical martingale problem formulation of the existence of a solution to (2.1). Let $\phi_j(\cdot)$, $q(\cdot)$ and $\rho, \tau, u_i \leq \rho$, satisfy the conditions above (3.12) and let $f(\cdot)$ be real-valued with compact support and with partial derivatives up to third order being continuous. Let A denote the differential generator of $X(\cdot)$. To identify the process $X^\omega(\cdot)$ as that satisfying (3.9) for an appropriate Wiener process

$W^\omega(\cdot)$, it is sufficient to show that $X(\cdot)$ solves the martingale problem of Stroock and Varadhan; namely, for all such $f(\cdot), q(\cdot)$, etc.,

$$\begin{aligned}\int Q^\omega(d\psi(\cdot))q(x(u_i), \langle \pi(u_i), \phi_j\rangle, b(u_i); i, j)\\ \times \left[f(x(\rho+\tau)) - f(x(\rho)) - \int_\rho^{\rho+\tau} Af(x(s))ds\right] = 0.\end{aligned} \tag{4.2}$$

To show (4.2), it is natural to try to evaluate

$$\begin{aligned}E\Bigg[\int Q^{h,T}(d\psi(\cdot))q(x(u_i), \langle \pi(u_i, \phi_j\rangle, b(u_i); i, j,)\\ \times\left[f(x(\rho+\tau)) - f(x(\rho)) - \int_\rho^{\rho+\tau} Af(x(s))ds\right]\Bigg]_1^2\end{aligned} \tag{4.3}$$

and show that it goes to zero as h, T go to their limits, analogous to the procedure used in connection with (3.12) and (3.13). This and the weak convergence imply (4.2); equivalently, there exists $X^\omega(\cdot)$ and $W^\omega(\cdot)$, with $X^\omega(\cdot)$ nonanticipative, satisfying (3.9) for almost all ω.

It is hard to show (4.3) directly. We take an approach which has been one of the most powerful tools to date for functional limit theorems for convergence to diffusion type processes, namely, the perturbed test function method [5, 13, 17, 18, 19, 26]. This method uses a perturbation $f^h(\cdot)$ to the test function $f(\cdot)$ in (4.2). To simplify the development, and preserve the generality of the results, we will simply assume that a suitable perturbation exists, but methods for constructing it in many important cases are in the references.

The perturbed test function method. Given a test function $f(\cdot)$, which is three times continuously differentiable and with compact support, we seek a process $f^h(\cdot)$ which is close to $f(X^h(\cdot))$ and an extension $\hat{A}^h$ of the operator A such that, loosely speaking,

$$f^h(\cdot) - f(X^h(\cdot)) \to 0,$$

$$\hat{A}^h f^h(\cdot) - Af(X^h(\cdot)) \to 0.$$

With such perturbed test functions $f^\gamma(\cdot)$ available, simple adaptations of the martingale method can be used to identify $X^\omega(\cdot)$.

The operator $\hat{A}^h$. Let $\mathcal{F}_t^h$ be a filtration on the probability space, where $\mathcal{F}_t^h$ measures at least $\{\Psi^h(s), s \le t\}$, and where $\Psi^h(s) = (X^h(s), \Pi^h(s), Y^h(s), B(s))$. Let $u(\cdot)$ and $v(\cdot)$ denote measurable processes which are $\mathcal{F}_t^h$-adapted and progressively measurable and satisfy the following conditions:

$$\sup_t E|u(t)| < \infty, \quad \sup_t E|v(t)| < \infty,$$

$$\sup_{0<\delta,t} \frac{E|E[u(t+\delta)|\mathcal{F}_t^h] - u(t)|}{\delta} < \infty, \tag{4.4}$$

$$\lim_{\delta\to 0} E\left|\frac{E[u(t+\delta)|\mathcal{F}_t^h] - u(t)}{\delta} - v(t)\right| = 0 \text{ almost all } t.$$

Then we say that $u(\cdot) \in \mathcal{D}(\hat{A}^h)$, the *domain of the operator* $\hat{A}^h$, and write $\hat{A}^h u = v$. This operator was introduced originally in [13, 27]. The most important property of the operator $\hat{A}^h$ is given by Lemma 4.1.

The usefulness of the operator $\hat{A}^h$ for proving weak convergence depends on the choice of the conditioning σ-algebra $\mathcal{F}_t^h$. Suppose that $X^h(\cdot)$ is driven by a wide bandwidth noise process $\xi^h(\cdot)$. Then, we would normally set $\mathcal{F}_t^h$ to be the minimal σ-algebra which measures $\{\Psi^h(s), \xi^h(s),\ s \le t\}$. See the references for more detail.

LEMMA 4.1 (see [13; 18, section 3.2; 19, section 7.2]). *Let $u(\cdot) \in \mathcal{D}(\hat{A}^h)$. Then the process defined by*

$$M_u(t) = u(t) - \int_0^t \hat{A}^h u(s) ds \tag{4.5a}$$

*is an $\mathcal{F}_t^h$-martingale and, w.p.*1,

$$E[u(t+s)|\mathcal{F}_t^h] = u(t) + \int_t^{t+s} E[\hat{A}^h u(r)|\mathcal{F}_t^h] dr. \tag{4.5b}$$

Equation (4.5) *continues to hold if the t and $t+s$ are replaced by any bounded stopping times τ_1 and τ_2 with $\tau_1 \le \tau_2$, and which take countably many values. If $u(\cdot)$ is right continuous, then the τ_i can be any bounded stopping times with $\tau_1 \le \tau_2$.*

In order to apply the perturbed test function method we need that the domain of the operator $\hat{A}^h$ is sufficiently rich and as $h \to 0$ the operator $\hat{A}^h$ is consistent with A. This is made precise in the following assumption.

ASSUMPTION 4.2. *For every continuous real-valued function $f(\cdot)$, with compact support and whose partial derivatives up to third order are continuous, there exists $f^h(\cdot) \in \mathcal{D}(\hat{A}^h)$ which is right continuous w.p.*1 *and which satisfies* (4.6) *and* (4.7) *for each $T_1 > 0$:*

$$\lim_h \sup_t E|f^h(t) - f(X^h(t))| = 0, \tag{4.6}$$

$$\lim_h \sup_t \sup_{\tau \le T_1} E \left| \int_t^{t+\tau} \left[\hat{A}^h f^h(s) - Af(X^h(s)) \right] ds \right| = 0. \tag{4.7}$$

General methods of construction of $f^h(\cdot)$ are in [18] under quite broad conditions on the processes involved, and the reader is referred to that reference for more information. See also [19]. The key result for applications is the following lemma.

LEMMA 4.2 (see [18]). *Let the sequence $\{X^h(\cdot), h > 0\}$ converge weakly in $D[\mathbb{R}^r; 0, \infty)$ to a process $X(\cdot)$. Assume that Assumption* 4.2 *holds. Then, $X(\cdot)$ solves the martingale problem for operator A.*

The following theorem essentially says that under the above assumptions Theorems 3.1 and 3.2 continue to hold for this approximate filtering problem.

THEOREM 4.1. *Assume that Assumptions* 2.1, 2.2, 2.3, 4.1, *and* 4.2 *hold. Then* (3.3a) *holds with $X(\cdot)$ replaced by $X^h(\cdot)$. Also,* (3.3b), (3.3c), *and* (3.4) *continue to hold with the modified definitions of $\Pi^h(\cdot)$ and $Q^{h,T}(\cdot)$ stated at the beginning of this section.*

Let Q be a weak limit as in Theorem 3.1*; then Q^ω induces a process*

$$\Psi^\omega(\cdot) = (X^\omega(\cdot), \Pi^\omega(\cdot), Y^\omega(\cdot), B^\omega(\cdot)).$$

For almost all ω the following hold.

(i) *The pair* $(X^{\omega}(\cdot), \Pi^{\omega}(\cdot))$ *is stationary.*

(ii) *There exists a Wiener process* W^{ω} *for which statements* (i) *through* (v) *in Theorem* 3.2 *hold.*

Finally the last conclusion of Theorem 3.2 *also holds; i.e.,* (3.11) *is true.*

Comment on the proof. The only difference from the proofs of Theorems 3.1 and 3.2 concern the proof of (4.2). As noted, this is to be done by showing that (4.3) goes to zero as h, T go to their limits. Thus, we confine our remarks to an evaluation of (4.3).

Let $h \to 0, T \to \infty$ index a weakly convergent subsequence. The expression

$$\begin{aligned} &\int Q^{h,T}(d\psi(\cdot)) q(x(u_i), \langle \pi(u_i), \phi_j \rangle, b(u_i); i, j) \\ &\qquad \times \left[f(x(\rho+\tau)) - f(x(\rho)) - \int_{\rho}^{\rho+\tau} Af(x(s))ds \right] \end{aligned} \tag{4.8}$$

equals

$$\begin{aligned} &\frac{1}{T}\int_0^T q(X^h(t+u_i), \langle \Pi^h(t+u_i), \phi_j \rangle, B(t+u_i) - B(t); i, j) \\ &\qquad \times \left[f(X^h(t+\tau+\rho)) - f(X^h(t+\rho)) - \int_{t+\rho}^{t+\rho+\tau} Af(X^h(s))ds \right] dt. \end{aligned} \tag{4.9}$$

Define

$$M_f^h(t) = f^h(t) - \int_0^t \hat{A}^h f^h(s) ds.$$

Rewrite the bracketed term in (4.9) as

$$\left[M_f^h(t+\tau+\rho) - M_f^h(t+\rho) \right] + [V^h(t+\tau+\rho) - V^h(t+\rho)],$$

where

$$V^h(t) = [f(X^h(t)) - f^h(t)] - \int_0^t [Af(X^h(s)) - \hat{A}^h f^h(s)] ds.$$

By (4.6) and (4.7), $\lim_h \sup_t E|V^h(t+\tau+\rho) - V^h(t+\rho)| = 0$. The weak sense limits of (4.9) are then the same as those of

$$\frac{1}{T}\int_0^T q(X^h(t+u_i), \langle \Pi^h(t+u_i), \phi_j \rangle, B(t+u_i) - B(t); i, j) \left[M_f^h(t+\tau+\rho) - M_f^h(t+\tau) \right] dt. \tag{4.10}$$

Now, square (4.10), take expectations and use the martingale properties of $M_f^h(\cdot)$ to get that the mean square value is $O(1/T)$. This and the weak convergence imply (4.2). □

5. Discrete time observations.

5.1. A Markov signal process. Now, let all processes be defined in discrete time. The idea is fully analogous to what was done in the continuous time case. The following is the basic assumption on the signal process.

ASSUMPTION 5.1. *The signal process $X(\cdot) = \{X(n), n < \infty\}$ is Feller–Markov and takes values in a compact subset G of $\mathbb{R}^r$.*

The observations are defined by $Y(0) = 0$ and

$$Y(n) - Y(n-1) = g(X(n)) + \xi(n),\ n = 1, \ldots, \tag{5.1}$$

where $\{\xi(n)\}$ are mutually independent $(0, I)$ Gaussian random variables which are independent of $X(\cdot)$, and $g(\cdot)$ is continuous.

The Bayes' rule formula for the true conditional distribution of $X(n)$ given $Y_{0,n}$ can be represented in terms of an auxiliary process $\tilde{X}(\cdot)$ as in section 2, where $\tilde{X}(\cdot)$ has the probability law of $X(\cdot)$ but, conditioned on its (possibly random) initial condition, is independent of all the other processes. Analogous to the notation used for the continuous time problem, let $Y_{0,n}$ denote the set $\{Y(i); i \leq n\}$. Define

$$R(\tilde{X}_{0,n}, Y_{0,n}) = \exp\left[\sum_{i=1}^{n} g'(\tilde{X}(i))[Y(i) - Y(i-1)] - \frac{1}{2}\sum_{i=1}^{n}\left|g(\tilde{X}(i))\right|^2\right].$$

Then as in section 2, the optimal filter $\Pi(\cdot)$ can be defined by its moments:

$$\langle \Pi(n), \phi\rangle = \frac{E_{\{\Pi(0),Y_{0,n}\}}\left[\phi(\tilde{X}(n))R(\tilde{X}_{0,n}, Y_{0,n})\right]}{E_{\{\Pi(0),Y_{0,n}\}}R(\tilde{X}_{0,n}, Y_{0,n})}, \tag{5.2}$$

where $\Pi(0)$ is the distribution of $X(0)$ and $\tilde{X}(0)$. As before $(X(\cdot), \Pi(\cdot))$ is Feller–Markov. Henceforth, as in the previous sections, when discussing $\Pi(\cdot)$, we allow $\Pi(0)$ to be arbitrary, and not necessarily the distribution of $X(0)$.

Analogous to the case in section 2, it is generally necessary to approximate the optimal filter in some way. For example, the sequence $X(\cdot)$ might be samples of a diffusion process taken at discrete instants, and the transition function must then be computed approximately. In order to approximate the transition function we might try to solve the corresponding Fokker–Planck equation on $[0, 1]$ by the Markov chain approximation method and also appropriately approximate the initial condition. Or we might try to solve it by some other numerical method (e.g., finite elements, spectral method, etc.) which yields approximations to the one step transition density (assuming that there is a density) which converge in L_2 to the true transition density as the approximation parameter goes to its limit. If this approximation is not nonnegative, we can take its positive part and renormalize to get a transition probability. Finally, even if the original processes were given in discrete time, i.e., no approximations of a p.d.e. were needed, it might be too hard to perform the integrations with the true transition function and we might have to use an appropriate approximation instead. These considerations lead to a formulation analogous to what led to the approximating form (2.7), which we now develop.

Analogous to for the continuous time case, the approximating filter can often be represented in terms of a discrete parameter auxiliary Markov process $\tilde{X}^h(\cdot)$, with values in a compact subset of $\mathbb{R}^r$, and which is independent of the other processes, given its initial condition. As in the continuous time case we will assume the following condition on the auxiliary process.

ASSUMPTION 5.2. *If $\tilde{X}^h(0)$ converges weakly (for any subsequence of values of $h \to 0$) with limit distribution $\Pi(0)$, then the sequence $\tilde{X}^h(\cdot)$ converges weakly to $X(\cdot)$ with initial distribution $\Pi(0)$.*

Define, for $n = 1, \ldots,$

$$R(\tilde{X}^h_{0,n}, Y_{0,n}) = \exp\left[\sum_{i=1}^{n} g'(\tilde{X}^h(i))[Y(i) - Y(i-1)] - \frac{1}{2}\sum_{i=1}^{n}\left|g(\tilde{X}^h(i))\right|^2\right].$$

Again, define the approximating filter $\Pi^h(\cdot)$ by its moments as follows:

$$\langle \Pi^h(n), \phi\rangle = \frac{E_{\{\Pi^h(0), Y_{0,n}\}}\left[\phi(\tilde{X}^h(n))R(\tilde{X}^h_{0,n}, Y_{0,n})\right]}{E_{\{\Pi^h(0), Y_{0,n}\}}R(\tilde{X}^h_{0,n}, Y_{0,n})}. \tag{5.3}$$

Equation (5.3) is just the Bayes' rule formula for the filter for the signal process $\tilde{X}^h(\cdot)$, but with the actual observations $Y(n) - Y(n-1)$ used.

The analog of the semigroup relation (2.8) holds, but where time is discrete. The semigroup property stems from the fact that the filter is that for the discrete parameter Markov process $\tilde{X}^h(\cdot)$, but with the actual physical observations used.

Define the sequences $B(n) = \sum_{i=1}^{n} \xi^i$ and

$$\Psi^h(n, \cdot) = \left\{X(n + \cdot), \Pi^h(n + \cdot), B(n + \cdot) - B(n), Y(n + \cdot) - Y(n)\right\},$$

$$\Psi^h_f(n, \cdot) = \{X(n + \cdot), \Pi^h(n + \cdot)\}$$

. Define $\psi(\cdot)$ and $\psi_f(\cdot)$ analogously.

The Skorohod topology is replaced by a "sequence" topology as follows: The $\Pi^h(n)$ still take values in $\mathcal{M}(G)$ and the weak topology is still used on this space. Let $d^\pi(\cdot)$ and $d^k(\cdot)$ denote the metrics on the space of measures and on Euclidean-k space, resp., where k is the sum of the dimensions of $Y(n), B(n)$, and $X(n)$. Let $d_0(\cdot)$ denote the product metric. Then the metric on the product path (sequence) space is

$$d(a(\cdot), b(\cdot)) = \sum_{n=0}^{\infty} 2^{-n}\left[d_0(a(n), b(n)) \wedge 1\right].$$

Define the occupation measure $Q^{h,N}(\cdot)$ by: For a Borel set C in the product sequence space,

$$Q^{h,N}(C) = \frac{1}{N}\sum_{n=1}^{N} I_C(\Psi^h(n, \cdot)). \tag{5.4}$$

Finally we impose the following basic condition.

ASSUMPTION 5.3. *There is a unique stationary measure $\bar{Q}_f(\cdot)$ of the $\Psi_f(\cdot) = (X(\cdot), \Pi(\cdot))$ process.*

Let $F(\cdot)$ be a real-valued bounded and continuous (w.p.1 with respect to $\bar{Q}_f(\cdot)$) function of $\psi_f(\cdot)$. Then, analogous to (2.10) and (2.14), we are concerned with the convergence (in probability)

$$\frac{1}{N}\sum_{n=1}^{N} F(\Psi^h_f(n + \cdot)) \to \int F(\psi_f(\cdot))\bar{Q}_f(d\psi_f(\cdot)), \tag{5.5}$$

where $h \to 0$ and $N \to \infty$ in any way at all.

The analogs of Theorems 3.1 and 3.2 hold, and the proof is close to those proofs. But owing to the discrete time some of the details are simpler, and others slightly different.

THEOREM 5.1. *Assume that Assumptions* 5.1, 5.2, *and* 5.3 *hold. Then* $\{Q^{h,N}(\cdot); h > 0, N > 0\}$ *is tight. Let* $Q(\cdot)$ *denote a weak sense limit, always as* $h \to 0$ *and* $N \to \infty$. *Let* ω *be the canonical variable on the probability space on which* $Q(\cdot)$ *is defined, and denote the sample values by* $Q^{\omega}(\cdot)$. *Then, for each* ω, $Q^{\omega}(\cdot)$ *is a measure on the product path* (*sequence*) *space. It induces a process*

$$\Psi^{\omega}(\cdot) = (X^{\omega}(\cdot), \Pi^{\omega}(\cdot), Y^{\omega}(\cdot), B^{\omega}(\cdot)) .$$

For almost all ω, *the following hold:*

(i) $(X^{\omega}(\cdot), \Pi^{\omega}(\cdot))$ *is stationary.*

(ii) $B^{\omega}(\cdot)$ *is the sum if mutually independent* $N(0, I)$ *random variables* $\{\xi^{\omega}(n)\}$ *which are independent of* $X^{\omega}(\cdot)$.

(iii)

$$Y^{\omega}(n) - Y^{\omega}(n-1) = g(X^{\omega}(n)) + \xi^{\omega}(n).$$

(iv) $X^{\omega}(\cdot)$ *has the transition function of* $X(\cdot)$.

Finally, for each integer n *and each bounded and measurable real-valued function* $\phi(\cdot)$,

$$\langle \Pi^{\omega}(n), \phi \rangle = \frac{E_{\{\Pi^{\omega}(0), Y^{\omega}_{0,n}\}} \left[\phi(\tilde{X}(n)) R(\tilde{X}_{0,n}, Y^{\omega}_{0,n}) \right]}{E_{\{\Pi^{\omega}(0), Y^{\omega}_{0,n}\}} R(\tilde{X}_{0,n}, Y^{\omega}_{0,n})}.$$

Remark on the proof. The details are essentially the same as those of the proofs of Theorems 3.1 and 3.2 and are omitted. We note only two differences. First, owing to the use of discrete time, the martingale method cannot be used to characterize $B^{\omega}(\cdot)$, as done in the first part of the proof of Theorem 3.2. But an analogous method, using a direct computation of the (conditional) characteristic function, can be used instead, as follows. Replace the expression $F(\cdot)$ defined above (3.12) by the following. For arbitrary integers m, n, k, let $u_i \le m$ and $v_p; p \le n$, be arbitrary integers. Let ν_i be vectors (with the dimension of $\xi(n)$), and replace the expression by

$$\begin{aligned} F(\psi(\cdot)) &= q(x(v_p), \langle \pi(u_i), \phi_j \rangle, b(u_i); i \le m, p \le n, j) \\ &\times \left[\exp\left[\sum_{l=m}^{m+k} \nu_l' [b(l+1) - b(l)] \right] - \exp\left[\sum_{l=m}^{m+k} |\nu_l|^2/2 \right] \right] \end{aligned}$$

We need to show that, for almost all ω,

$$\int F(\psi(\cdot)) Q^{\omega}(d\psi(\cdot)) = 0, \tag{5.6a}$$

which will imply that

$$\begin{aligned} &\int Q^{\omega}(d\psi(\cdot)) q(x(v_p), \langle \pi(u_i), \phi_j \rangle, b(u_i); i \le m, p \le n, j) e^{\sum_{l=m}^{m+k} \nu_l' [b(l+1) - b(l)]} \\ &\quad = \int Q^{\omega}(d\psi(\cdot)) q(x(v_p), \langle \pi(u_i), \phi_j \rangle, b(u_i); i \le m, p \le n, j) e^{\sum_{l=m}^{m+k} |\nu_l|^2/2}. \end{aligned} \tag{5.6b}$$

This, in turn, implies that (for almost all ω) the $\xi^\omega(l)$ are mutually independent, normally distributed with mean zero and with the covariance being the identity matrix, and are independent of the $X^\omega(\cdot)$ and of the "past" of the $\Pi^\omega(\cdot)$ process. The proof of (5.6a) is analogous to the arguments below (3.12), and the details are omitted.

We need to characterize $X^\omega(\cdot)$. In particular, we need to show that it is Markov with the transition function of $X(\cdot)$. A characteristic function and "weak law of large numbers type" argument can be used. The goal is to show that, for any integer m, any bounded and continuous real-valued function $q(\cdot)$, and any vector λ,

$$\int Q^\omega(d\psi(\cdot))q(x(u_j); u_j \leq m)\left[e^{\lambda' x(m+1)} - \int e^{i\lambda' v} P(dv, 1|x(m))\right] = 0, \tag{5.7}$$

where $P(dv, 1|x)$ is the one step transition function of $X(\cdot)$. Equation (5.7) says that $X^\omega(\cdot)$ is Markov with the transition function of $X(\cdot)$ since it says that the conditional expectation of $e^{i\lambda' X^\omega(m+1)}$, given the "past" and the current state, can be computed by use of the transition function.

Rewrite (5.7) as the weak sense limit of

$$\frac{1}{N}\sum_{n=1}^{N} q(X(n+u_j); u_j \leq m)\left[e^{i\lambda' X(n+m+1)} - \int e^{i\lambda' v} P(dv, 1|X(n+m))\right]. \tag{5.8}$$

Note that by the Markov property the expectation of the nth summand (conditioned on the data to time $n+m$) in (5.8) is zero. Now, use this last fact to show that the mean square value is $O(1/N)$. The rest of the details are as in the proofs of Theorems 3.1 and 3.2. □

5.2. A non-Markov signal process. Recall the model of section 4, where the underlying signal process, called $X^h(\cdot)$ there, was not Markov. We supposed that it converged weakly to the process of section 2 as $h \to 0$. We can do a similar analysis in the discrete time case, and it is worthwhile since the actual signal process will not usually be Markov and it is not a priori obvious that even small errors per step will not lead to large errors as the time interval goes to infinity. As noted in section 4, we could use a symbol other than h to index this process, but there is no loss of generality in using h. In lieu of the perturbed test function method introduced in section 4, we will follow the procedure of the previous subsection as closely as possible. We keep the assumptions on $\tilde{X}^h(\cdot)$ and the filter form (5.3), which uses instead of $Y(\cdot)$ the following modified observation process:

$$Y^h(n) - Y^h(n-1) = g(X^h(n)) + \xi(n).$$

The main new problem is the identification of the $X^\omega(\cdot)$. We concentrate on that and make the assumptions of the last subsection.

Some assumption on the convergence of $X^h(\cdot)$ to $X(\cdot)$ is needed. The procedure of the last subsection requires that we show that

$$\frac{1}{N}\sum_{n=1}^{N} q(X^h(n+u_j); u_j \leq m)\left[e^{i\lambda' X^h(n+m+1)} - \int e^{i\lambda' v} P(dv, 1|X^h(n+m))\right] \tag{5.9}$$

converges to zero in probability as $N \to \infty$. The convergence in (5.9) is not guaranteed by the weak convergence of $X^h(\cdot)$ to $X(\cdot)$, since the property (5.9) involves the "long range dependencies" of the $X^h(\cdot)$ processes. We will take an approach that is quite flexible and whose conditions are not stringent.

We make the following assumptions. Define the process $X(\cdot|x)$ to have the law of evolution of $X(\cdot)$, but with initial condition x.

ASSUMPTION 5.4. *For an arbitrary integer m and any bounded and continuous real-valued function $f(\cdot)$ of m arguments,*

$$\lim_h \sup_n \left[Ef(X^h(n+u_j); u_j \le m) - E\left[f(X(u_j|X^h(n)); u_j \le m) \right] \right] = 0. \tag{5.10}$$

ASSUMPTION 5.5. *For an arbitrary integer m and any bounded and continuous real-valued function $f(\cdot)$ of m arguments and any $\mu > 0$, there is $m_\mu < \infty$ such that*

$$\begin{aligned} \limsup_h \sup_{|n-\nu|\ge m_\mu} & E\left(\left[f(X^h(n+u_j); u_j \le m) - Ef(X^h(n+u_j); u_j \le m) \right] \right. \\ & \times \left. \left[f(X^h(\nu+u_j); u_j \le m) - Ef(X^h(\nu+u_j); u_j \le m) \right] \right) \le \mu. \end{aligned} \tag{5.11}$$

Condition (5.10) can be interpreted as follows: It is equivalent to saying that if for any sequence n_h, $X^h(n_h)$ converges weakly to a random variable $\hat{X}(0)$ as $h \to 0$, then $\{X^h(n_h + \cdot)\}$ converges weakly to $X(\cdot)$ with initial condition $\hat{X}(0)$. It is unrestrictive since the usual choices for $X^h(\cdot)$ are Markov with time independent transition functions. Condition (5.11) deals with long-range dependence. It basically says that if we take increments of the $X^h(\cdot)$ process which are separated by a large time interval which does not depend on h), then their correlation is small for small h. This does not appear to be restrictive. It is guaranteed by appropriate mixing conditions [9]. Actually, (5.11) is needed only for the types of functions which appear in (5.9). In the definition of $\Psi^h(\cdot)$ replace $X(\cdot), Y(\cdot)$ by $X^h(\cdot), Y^h(\cdot)$, resp. Redefine $\Pi^h(\cdot)$ by replacing $Y(\cdot)$ by $Y^h(\cdot)$ in (5.3). Modify the definitions of $\Psi^h_f(\cdot)$ and $Q^{h,N}(\cdot)$ in a similar manner.

THEOREM 5.2. *Assume that Assumptions* 5.1, 5.2, 5.3, 5.4, *and* 5.5 *hold. Then the conclusions of Theorem* 5.1 *hold with the above modified definitions of* $\Pi^h(\cdot)$ *and* $Q^{h,N}(\cdot)$.

Comments on the proof. All of the details are as in Theorem 5.1, except for the characterization of $X^\omega(\cdot)$ and we concentrate on this. Return to the proof of (5.7) via an evaluation of the limit of (5.9). For a bounded and continuous function $q(\cdot)$, define

$$q_1(x(u_j), x(m+1); u_j \le m) = q(x(u_j); u_j \le m) e^{i\lambda' x(m+1)},$$

$$q_2(x(u_j); u_j \le m) = q(x(u_j); u_j \le m) \int e^{i\lambda' v} P(dv, 1|x(m)).$$

We will show that

$$\left| \int Q^{N,h}(d\psi(\cdot)) \left[q_1(x(u_j), x(m+1); u_j \le m) - q_2(x(u_j); u_j \le m) \right] \right| \to 0 \tag{5.12}$$

in mean as $N \to \infty$ and $h \to 0$. This, together with the weak convergence, yields (5.7) for almost all ω. Evaluate the integral in (5.12) by first rewriting it as

$$\frac{1}{N} \sum_{n=1}^{N} \left[q_1(X^h(n+u_j), X^h(n+m+1); u_j \le m) - q_2(X^h(n+u_j); u_j \le m) \right]. \tag{5.13}$$

Now, split (5.13) into the difference of the two sums

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N} \big[&q_1(X^h(n+u_j), X^h(n+m+1); u_j \le m) \\ &-Eq_1(X^h(n+u_j), X^h(n+m+1); u_j \le m)\big]\end{aligned} \tag{5.14}$$

and

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N} \big[&q_2(X^h(n+u_j), X^h(n+m+1); u_j \le m) \\ &-Eq_1(X^h(n+u_j), X^h(n+m+1); u_j \le m)\big].\end{aligned} \tag{5.15}$$

By Assumption 5.5, (5.14) goes to zero in mean square as $N \to \infty$ and $h \to 0$. By Assumption 5.4, the term $Eq_1(X^h(n+u_j), X^h(n+m+1); u_j \le m)$ in (5.15) can be replaced by $Eq_1(X(u_j|X^h(n)), X(m+1|X^h(n)); u_j \le m)$ in the sense that the mean square limits as $N \to \infty$ and $h \to 0$ are the same. By the Markov property of $X(\cdot)$, for any initial condition

$$Eq_1(X(u_j), X(m+1); u_j \le m) = Eq_2(X(u_j); u_j \le m).$$

By the last two sentences, we can replace the expected value in (5.15) by

$$Eq_2(X(u_j|X^h(n)); u_j \le m)$$

without changing the limits. Using Assumption 5.4 again, we can replace $Eq_2(X(u_j|X^h(n));$ $u_j \le m)$ by $Eq_2(X^h(n+u_j); u_j \le m)$ in (5.15) in that the mean square limits are the same. Now, use Assumption 5.5 again to get that the mean square limit of (5.15) with this last replacement is zero. We have shown that the mean square limit of (5.13) is zero, as $N \to \infty$ and $h \to 0$, which implies (5.7). □

6. Discrete to continuous time: White or wideband observation noise.

The most appropriate justification of the classical continuous time filter is as an approximation to the discrete time model for small sampling intervals. Quite often, one has wide bandwidth rather than white noise and the signal is not a diffusion but only an approximation to a diffusion. One still might wish to use some sort of ideal filter, which is built on the assumption that the signal is a particular diffusion and that the observation noise is white. If the filter is to be used for a very long time interval and one is interested in pathwise prediction errors or other performance measures (see section 2 for an example), questions arise concerning the asymptotic quality, as the time, the bandwidth of the observation noise, the parameter in the diffusion approximation to the actual signal process, etc., all go to their respective limits. We will see that the previous robustness results continue to hold under reasonable conditions.

Asymptotic, large time results were obtained in [21] for mean value criteria. Depending on how the filter with wide bandwidth observation noise is constructed, there might be the so-called correction terms [21]. In the present context, we suppose that either the basic Bayes' rule formulas such as (2.7) or its later analogs are used or that the constructed filter is equivalent to these formulas, at least asymptotically. There is no need to introduce correction terms into the Bayes' rule formulas as written here.

The problem formulation. The signal process $X^h(\cdot)$ is assumed to satisfy the conditions used in Theorem 4.1 and is sampled at intervals of width $\Delta_h \to 0$. The observation noise is wide bandwidth in the sense to be made precise below. There is a process $B^h(\cdot)$ satisfying Assumptions 6.1, 6.2, and 6.3 below, such that the observation process, $\{Y^h(n\Delta_h); n \geq 0\}$ satisfies the equation

$$Y^h(n\Delta_h) - Y^h(n\Delta_h - \Delta_h) = \Delta_h g(X^h(n\Delta_h)) + B^h(n\Delta_h) - B^h(n\Delta_h - \Delta_h). \tag{6.1}$$

The assumptions on the process $B^h(\cdot)$ are as follows.

Assumption 6.1. *The process $B^h(\cdot)$ is independent of $X^h(\cdot)$. The set*

$$\{B^h(t + \cdot) - B^h(t); h > 0, t \geq 0\} \tag{6.2}$$

is tight and any weakly convergent subsequence of the form $\{B^{h_k}(t_k + \cdot) - B^{h_k}(t_k)\}_{k\geq 1}$, such that $h_k \to 0$ as $k \to \infty$ converges to the standard Wiener process $B(\cdot)$ as $k \to \infty$.

Assumption 6.2. *For every family of random variables of the form $\{\mu_n^{\Delta,h}, n \geq 1, h > 0, \Delta > 0\}$ which satisfies* (i)–(iii) *below, we have that for each $T > 0$ and $\epsilon > 0$,*

$$\lim_{\Delta\to 0} \limsup_h \sup_n P\left\{ \sup_{s\leq T} \left| \sum_{i=n}^{n+s/\Delta_h} \mu_i^{\Delta,h} \left[B^h(i\Delta_h + \Delta_h) - B^h(i\Delta_h)\right] \right| \geq \epsilon \right\} = 0. \tag{6.3}$$

(i) $\mu_n^{\Delta,h}$ *is bounded uniformly in n, Δ, and h.*

(ii) *For each fixed h and Δ, the family $\{\mu_n^{\Delta,h}, n < \infty\}$ is independent of $B^h(\cdot)$.*

(iii)

$$\lim_{\Delta\to 0} \limsup_h \sup_n P\left\{ \sup_{n\leq i\leq n+T/\Delta_h} \left|\mu_i^{\Delta,h}\right| \geq \delta \right\} = 0, \text{ for each } \delta > 0. \tag{6.4}$$

The requirement (6.3) is mainly a condition on how fast the conditional dependence of the "future," given the "past," goes to zero. Owing to the independence of $\{\mu_n^h, n < \infty\}$ and $B^h(\cdot)$, the condition is not stringent. Various sufficient conditions for (A6.2) to hold will be given at the end of the section.

Now for the final assumption.

Assumption 6.3. *For every positive integer m, real numbers $\rho > 0$, $u_j \leq \rho, j \leq m$, and real-valued bounded and continuous function $f(\cdot)$*

$$\begin{aligned} &Ef(B^h(t + u_j) - B^h(t); j)f(B^h(s + u_j) - B^h(s); j) \\ &\quad -Ef(B^h(t + u_j) - B^h(t); j)Ef(B^h(s + u_j) - B^h(s); j) \to 0 \end{aligned} \tag{6.5}$$

as $h \to 0$, uniformly in (s, t), such that $s > t + \rho + 1$.

This condition essentially says that the correlation between intervals that are separated in time by at least some constant goes to zero as $h \to 0$, uniformly in time.

The process $B^h(\cdot)$ could be white, but not necessarily Gaussian. In this case, the verification of the conditions in Assumptions 6.2 and 6.3 is usually trivial.

We define $Y^h(s) = B^h(s) = 0$ for $s < 0$. The filter is denoted by $\{\Pi^h(n\Delta_h)\}$, and we define the continuous time interpolations $Y^h(\cdot), B^h(\cdot), \Pi^h(\cdot)$, and $\tilde{X}^h(\cdot)$ so that they are constant on the intervals $[n\Delta_h, n\Delta_h + \Delta_h)$.

In what follows, we use the convention that if a limit of summation is not an integer, take the integer part. We represent the filter as in the previous sections, in terms of an auxiliary process $\tilde{X}^h(\cdot)$ which satisfies the assumptions imposed on it in section 2. At the sampling times, the filter takes the form

$$\langle \Pi^h(n\Delta_h), \phi\rangle = \frac{E_{\{\Pi^h(0), Y^h_{0,n\Delta_h}\}}\left[\phi(\tilde{X}^h(n\Delta_h))R(\tilde{X}^h_{0,n\Delta_h}, Y^h_{0,n\Delta_h})\right]}{E_{\{\Pi^h(0), Y^h_{0,n\Delta_h}\}}R(\tilde{X}^h_{0,n\Delta_h}, Y^h_{0,n\Delta_h})}, \tag{6.6}$$

where we use the definition

$$R(\tilde{X}^h_{0,n\Delta_h}, Y^h_{k\Delta_h,(n+k)\Delta_h}) = e^Z, \tag{6.7}$$

where Z is defined by

(6.8)
$$\sum_{i=1}^{n} g'(\tilde{X}^h(i\Delta_h))[Y^h(k\Delta_h + i\Delta_h) - Y^h(k\Delta_h + i\Delta_h - \Delta_h)] - \frac{\Delta_h}{2}\sum_{i=1}^{n}\left|g(\tilde{X}^h(i\Delta_h)\right|^2.$$

The definition (6.6), with (6.8) used, satisfies the semigroup property over the discrete time mesh: $\{n\Delta_h; n \geq 1\}$.

Instead of constant interpolation of $\Pi^h(\cdot)$ over the interval $[n\Delta_h, n\Delta_h + \Delta_h)$, one may use the following different definition for the approximate filter.

$$\langle \Pi^h(t), \phi\rangle = \frac{E_{\{\Pi^h(0), Y^h_{0,t}\}}\left[\phi(\tilde{X}^h(t))R(\tilde{X}^h_{0,t}, Y^h_{0,t})\right]}{E_{\{\Pi^h(0), Y^h_{0,t}\}}R(\tilde{X}^h_{0,t}, Y^h_{0,t})}. \tag{6.9}$$

The forms (6.6) and (6.9) are equal at the sampling times. Theorem 6.1 continues to hold for the form of filter in (6.9); however, for the sake of brevity we will only prove the constant interpolation case.

Define $\Psi^h(\cdot) = \big(X^h(\cdot), \Pi^h(\cdot), Y^h(\cdot), B^h(\cdot)\big)$, $\Psi^h_f(\cdot) = \big(X^h(\cdot), \Pi^h(\cdot)\big)$, and define $Q^{h,T}$, $\psi(\cdot)$ and $\psi_f(\cdot)$ analogously.

THEOREM 6.1. *Assume that Assumptions* 2.1, 2.2, 2.3, 4.1, 4.2, 6.1, 6.2, *and* 6.3 *hold. Then all the conclusions of Theorem* 4.1 *hold with the above modified definitions of* $\Pi^h(\cdot)$ *(see* (6.6)*) and* $Q^{h,T}(\cdot)$.

Comments on the proof. The proof is similar to those of Theorems 3.1, 3.2, and 4.1. The main differences concern the characterization of the limit $X^\omega(\cdot)$, the proof for (3.3c), the proof that $B^\omega(\cdot)$ is a standard Wiener process which is independent of $X^\omega(\cdot)$, and the proof of the representation of $\Pi^\omega(\cdot)$.

The characterization of $X^\omega(\cdot)$ as $X(\cdot)$, with the appropriate initial condition, is analogous to what was done in section 4, with only minor notational differences, which are due to the replacement of $B(\cdot)$ by $B^h(\cdot)$, and so we omit the details.

Proof of (3.3c). It is difficult to prove (3.3c) directly. We will actually prove the tightness of an approximation to the family $\{\Pi^h(t+\cdot); h > 0, t > 0\}$ and use the following useful fact. Let the sequence $\{Z_n(\cdot)\}$ have paths in some Skorohod space with metric denoted by $d(\cdot,\cdot)$. Suppose that for each positive real numbers ϵ_1, ϵ_2 and T_0, there is a process $Z_n^{\epsilon_1,\epsilon_2,T_0}(\cdot)$ such that

$$\limsup_n P\left\{\sup_{s\leq T_0} d(Z_n(s), Z_n^{\epsilon_1,\epsilon_2,T_0}(s)) \geq \epsilon_1\right\} \leq \epsilon_2, \tag{6.10}$$

and $\{Z_n^{\epsilon_1,\epsilon_2,T_0}(\cdot), n < \infty\}$ is tight for each $\epsilon_1, \epsilon_2, T_0$. Then $\{Z_n(\cdot)\}$ is tight. The approximations generally simplify the characterization of the weak sense limits as well.

Note that for our case the process index is the pair (h,t) instead of n. Pursuing this idea, let $\epsilon_1, \epsilon_2, T_0$ be as above. Then, for each h and t, we seek a process $\hat{\Pi}^h(t,\cdot)$ (dropping the affixes $\epsilon_1, \epsilon_2, T_0$) such that

$$\limsup_h \sup_t P\left\{ \sup_{s\le T_0} \left| \langle \Pi^h(t+s), \phi\rangle - < \hat{\Pi}^h(t,s), \phi > \right| \ge \epsilon_1 \right\} \le \epsilon_2 \tag{6.11}$$

and $\{\hat{\Pi}^h(t,\cdot); h>0, t\ge 0\}$ is tight. If we can find such processes, then (3.3c) holds. We now show (6.11). We can work with each $\phi(\cdot)$ separately due to the definition of the weak topology.

Keep in mind that we can have a different modification for each (h,t). We first consider the case where $t = n\Delta_h$ for some n. Fix $\phi(\cdot)$. Let $\Delta > 0$ be small but fixed. Assume that h is small enough such that $\Delta >> \Delta_h$. The idea is to break the sum

$$\sum_{j=1}^{s/\Delta_h} g'(\tilde{X}^h(j\Delta_h)) \left[Y^h(t+j\Delta_h) - Y^h(t+j\Delta_h-\Delta_h)\right] \tag{6.12}$$

in (6.8) as a sum of

$$\sum_{i=0}^{s/\Delta-1} g'(\tilde{X}^h(i\Delta)) \left[Y^h(t+i\Delta+\Delta) - Y^h(t+i\Delta)\right] \tag{6.13}$$

and the "error"

$$\begin{aligned} V^{h,\Delta}(t,s) = \sum_{i=0}^{s/\Delta-1} & \sum_{j=i\Delta/\Delta_h+1}^{(i+1)\Delta/\Delta_h \wedge s/\Delta_h} \\ & \left[g'(\tilde{X}^h(j\Delta_h)) - g'(\tilde{X}^h(i\Delta))\right] \left[Y^h(t+j\Delta_h) - Y^h(t+j\Delta_h-\Delta_h)\right]. \end{aligned} \tag{6.14}$$

Note that if we concatenate the sums in (6.14), we get

$$\sum_{i=n}^{n+s/\Delta_h} \mu_i^{h,\Delta} \left[Y^h(i\Delta_h+\Delta_h) - Y^h(i\Delta_h)\right]$$

for an obvious definition of $\mu_i^{h,\Delta}$, which clearly satisfies (i), (ii), and (iii) in Assumption 6.2, in view of weak convergence properties of $\tilde{X}^h(\cdot)$. The case where t/Δ_h is not an integer is dealt with in a completely analogous way and we omit the details. Now we use Assumption 6.2 for the family $\{\mu_i^{h,\Delta}; h, \Delta, i\}$. For an arbitrary small $\epsilon_3 > 0$, define $\tau^{h,\Delta,t} = \min\{s : |V^{h,\Delta}(t,s)| \ge \epsilon_3\}$. Then Assumption 6.2 yields that, for small $\epsilon > 0$ and small enough $\epsilon_3 > 0$,

$$\lim_{\Delta\to 0} \limsup_h \sup_t P\left\{ P_{\{\Pi^h(0), Y^h_{t,t+s}\}} \left\{ \tau^{h,\Delta,t} < T_0 \right\} \ge \epsilon \right\} = 0. \tag{6.15}$$

Define $\hat{R}^{h,\Delta}(\tilde{X}^h_{0,s}, Y^h_{t,t+s})$ to be $R(\tilde{X}^h_{0,s}, Y^h_{t,t+s})$ with (6.13) replacing (6.12). Now define $\hat{\Pi}^h(t,\cdot)$ by

$$\langle \hat{\Pi}^h(t,s), \phi\rangle = \frac{E_{\{\Pi^h(0), Y^h_{0,t}\}} \left[\phi(\tilde{X}^h(t)) \hat{R}^{h,\Delta}(\tilde{X}^h_{0,s}, Y^h_{0,s})\right]}{E_{\{\Pi^h(0), Y^h_{0,t}\}} \hat{R}^{h,\Delta}(\tilde{X}^h_{0,s}, Y^h_{0,s})}. \tag{6.16}$$

We have

$$\hat{R}^{h,\Delta}(\tilde{X}^h_{0,s}, Y^h_{t,t+s})e^{V^h(t,s)} = R(\tilde{X}^h_{0,s}, Y^h_{t,t+s})$$

and, except on as a (h, t-dependent) set of arbitrarily small measure, $\tau^{h,\Delta,t} \geq T_0$ for small h and Δ. Hence, we can suppose that, asymptotically, $|e^{V^h(t,s)} - 1| \leq \epsilon_4$, arbitrary. Thus, (6.11) holds.

The tightness of the set $\{\hat{\Pi}^h(t,\cdot); h, t\}$ for each $\Delta > 0$ is clear. This follows by noting that the exponential contains the sum (6.13), hence $\hat{\Pi}^h(t,\cdot)$ contains only simple discontinuities at deterministic times $i\Delta$. This proves (3.3c). Thus in view of the tightness assumptions on the set $\{X^h(t+\cdot), B^h(t+\cdot) - B^h(t); h > 0, t\}$, we have that (3.4) holds. We can extract a weakly convergent subsequence of the set in (3.4), indexed again by (h, t), and with limit denoted by $Q(\cdot)$, with sample values $Q^\omega(\cdot)$.

The independence of $B^\omega(\cdot)$ and $X^\omega(\cdot)$. We will only do a simple calculation to illustrate the general procedure. Let $f_i(\cdot)$ be bounded and continuous real-valued functions of their arguments, and with compact support. We wish to show that, for almost all ω and any nonnegative u, v,

$$\int Q^\omega(d\psi(\cdot))f_1(b(v))f_2(x(u)) = \int Q^\omega(d\psi(\cdot))f_1(b(v)) \int Q^\omega(d\psi(\cdot))f_2(x(u)).$$

Write the prelimit forms of the difference between the two sides:

$$\begin{aligned} &\frac{1}{T}\int_0^T f_1(B^h(t+v) - B^h(t))f_2(X^h(t+u))dt \\ &\quad -\frac{1}{T^2}\int_0^T f_1(B^h(t+v) - B^h(t))dt \int_0^T f_2(X^h(s+u))ds. \end{aligned} \tag{6.17}$$

The rest of the procedure is to square, take expectations, and use the fact that $B^h(\cdot)$ is independent of $X^h(\cdot)$ and Assumption 6.3 to show that the difference goes to zero in mean square as $h \to 0, T \to \infty$. Define $C = Ef_1(B(v))$. Let us first evaluate the expectation of the square of the first term; namely, evaluate

$$\frac{1}{T^2}\int_0^T\int_0^T Ef_1(B^h(t+v)-B^h(t))f_2(X^h(t+u))f_1(B^h(s+v)-B^h(s))f_2(X^h(s+u))dtds.$$

Now, use Assumption 6.3, the weak convergence of $B^h(t+\cdot) - B^h(t)$, and the independence of $B^h(\cdot)$ and $X^h(\cdot)$ to show that the limits are the same as those of

$$C^2\frac{1}{T^2}\int_0^T\int_0^T Ef_2(X^h(t+u))f_2(X^h(s+u))dtds.$$

A similar procedure shows that the limits of the expectation of the square of the second term and of the product of the first and second terms in (6.17) is also the same. Thus, the limit is zero, as desired. The same procedure works if the functions $f_i(\cdot)$ depend on the processes at an arbitrary finite number of times. Hence the independence of $X^\omega(\cdot)$ and $B^\omega(\cdot)$.

The Wiener property of $B^\omega(\cdot)$ for almost all ω. The proof is similar to what was done to prove the independence of the $X^\omega(\cdot)$ and $B^\omega(\cdot)$. We need to show characteristic function results such as the following. Let λ_q be arbitrary vectors (of the dimension of b). Show that, for almost all ω,

$$\begin{aligned}&\int Q^\omega(d\psi(\cdot))e^{i\lambda_1'(b(u+s)-b(u))}e^{i\lambda_2'b(u)}\\ &\quad= \int Q^\omega(d\psi(\cdot))e^{i\lambda_1'(b(u+s)-b(u))}\int Q^\omega(d\psi(\cdot))e^{i\lambda_2'b(u)}\\ &\quad= e^{-s|\lambda_1|^2/2}e^{-u|\lambda_2|^2/2}.\end{aligned}$$

To show this, work with the difference of the prelimit forms of the two sides:

$$\frac{1}{T}\int_0^T \left[e^{i\lambda_1'(B^h(t+u+s)-B^h(t+u))}e^{i\lambda_2 B^h(t+u)-B^h(t))}\right]dt - e^{-s|\lambda_1|^2/2}e^{-u|\lambda_2|^2/2}.$$

Square, take expectations, and use the weak convergence and Assumption 6.3, analogous to what was done to show the independence property.

Characterization of the process $\boldsymbol{\Pi^\omega(\cdot)}$. Given the tightness of $\{\Pi^h(\cdot)\}$ and the discussion in that section, the fact that (3.10) holds is essentially a property of the asymptotic smoothness of $\tilde{X}^h(\cdot)$. It can be proved by working with $\hat{\Pi}^h(t,\cdot)$ in lieu of $\Pi^h(t+\cdot)$, and then letting $\Delta \to 0$, and the details are omitted. □

Sufficient conditions for Assumption 6.2 to hold. Define $\xi_i^h = B^h(i\Delta_h + \Delta_h) - B^h(i\Delta_h)$. The simplest case satisfying (6.3) is where the ξ_i^h are martingale differences with variances bounded by some constant times Δ_h), i.e., where the observation noise is white but not necessarily Gaussian.

Correlated noise can be treated by the perturbed test function method [18] of the type introduced in section 4. Here we will construct a particular perturbation which serves our present purposes. Let E_n denote the expectation conditioned on $\{\mu_i^{h,\Delta}, \xi_i^h, i \le n\}$. Fix $T > 0$. Define

$$S_i^{h,\Delta,n} = \sum_{j=n}^{n+i} \mu_j^{h,\Delta}\xi_j^h,\ i \le T/\Delta_h.$$

Assume that

$$\sup_{i,h} E\left|\xi_i^h\right|^2 < \infty. \tag{6.18}$$

Define the “perturbation”

$$\delta S_i^{h,\Delta,n,T} = \sum_{j=n+i}^{n+T/\Delta_h-1} E_{n+i}\mu_{j+1}^{h,\Delta}\xi_{j+1}^h,\ i \le T/\Delta_h. \tag{6.19}$$

Suppose that

$$\sum_{j=n+i}^{n+T/\Delta_h} E\left|E_{n+i}\xi_{j+1}^h\right| = O(\Delta_h), \tag{6.20}$$

where $O(\Delta_h)$ is uniform in i, and that

$$\limsup_h \sup_n E\sum_{i=n}^{n+T/\Delta_h}|\xi_i^h|^2 + \left|\sum_{i=n}^{n+T/\Delta_h}\sum_{j=i+1}^{n+T/\Delta_h} E\xi_i^h\xi_j^h\right| < \infty. \tag{6.21}$$

These conditions are rather mild. They would be satisfied, for example, if $B^h(\cdot)$ were a sufficiently "nonanticipatively smoothed" Wiener process; e.g., $B^h(t) = \int_{-\infty}^t q_h(t-s)B(s)ds$ for an appropriate kernel $q_h(\cdot)$ which converges to a delta function as $h \to 0$.

Noting that the index i is in the upper limit in (6.8) and in the lower limit in (6.9), we have

$$E_{n+i}\left[S_{i+1}^{h,\Delta,n} - S_i^{h,\Delta,n}\right] = E_{n+i}\mu_{n+i+1}^{h,\Delta}\xi_{n+i+1}^h,$$

$$E_{n+i}\left[\delta S_{i+1}^{h,\Delta,n,T} - \delta S_i^{h,\Delta,n,T}\right] = -E_{n+i}\mu_{n+i+1}^{h,\Delta}\xi_{n+i+1}^h.$$

Thus the perturbed function

$$\tilde{S}_i^{h,\Delta,n,T} = S_i^{h,\Delta,n} + \delta S_i^{h,\Delta,n,T}$$

is a martingale, and martingale inequalities can be used to get (6.3). The desire to get this martingale property is the key motivation behind the way the perturbation $\delta S_i^{h,\Delta,n,T}$ was constructed. It was constructed specifically to effect the desired cancellation.

Recall that if M_i is a square integrable martingale, then

$$P\left\{\sup_{i\le N}|M_i| \ge \epsilon\right\} \le \frac{EM_N^2}{\epsilon^2}. \tag{6.22}$$

Condition (6.20) and the fact that the $\mu_i^{h,\Delta}$ can be taken to be arbitrarily small implies that, for any $\epsilon > 0$,

$$\lim_{\Delta\to 0}\limsup_h \sup_n P\left\{\sup_{i\le T/\Delta_h}\left|\delta S_i^{h,\Delta,n,T}\right| \ge \epsilon\right\} = 0. \tag{6.23}$$

This is a consequence of the facts that the μ-terms are bounded and small in probability and independent of the ξ terms, and that (6.20) implies that

$$\sum_{i\le T/\Delta_h} E\left|\delta S_i^{h,\Delta,n,T}\right| = O(1).$$

Since $\delta S_{T/\Delta_h}^{h,\Delta,n,T} = 0$,

$$E|\tilde{S}_{T/\Delta_h}^{h,\Delta,n,T}|^2 = E|S_{T/\Delta_h}^{h,\Delta,n}|^2.$$

Also, a direct computation using (6.21), (6.22), and the boundedness and smallness in probability of the μ-terms (and their independence of the ξ) yields

$$\lim_{\Delta\to 0}\limsup_h \sup_n E|S_{T/\Delta_h}^{h,\Delta,n}|^2 = 0.$$

The last two equations and (6.22) yield

$$\lim_{\Delta\to 0}\limsup_h \sup_n P\left\{\sup_{i\le T/\Delta_h}\left|\tilde{S}_i^{h,\Delta,n,T}\right| \ge \epsilon\right\} = 0. \tag{6.24}$$

This and (6.23) imply (6.3).

7. Uniqueness of the invariant measure of $(X(\cdot),\Pi(\cdot))$. In this section we address the question of uniqueness of the invariant measure for the pair $(X(\cdot),\Pi(\cdot))$. It is known from Kunita [12] that if $X(\cdot)$ is a time homogeneous Feller–Markov process with a compact state space, such that it admits a unique invariant measure and the corresponding stationary flow is purely nondeterministic, then the true optimal filtering process $\Pi(\cdot)$ taken by itself has a unique invariant measure. The results in [12], however, do not provide any information concerning the uniqueness of the stationary joint distribution of $(X(\cdot),\Pi(\cdot))$. Uniqueness for the joint problem is considerably harder to prove and to date the only known result, excepting the linear filtering case, is for a discrete time filtering problem where the signal is a finite state Markov chain (see also Stettner [28]). We provide in Theorem 7.1 a sufficient condition under which the uniqueness for the joint problem holds. The theorem roughly says that if the signal process has a unique invariant measure and the filter "forgets its initial condition," in the sense to be made precise below, then the pair $(X(\cdot),\Pi(\cdot))$ has at most one invariant measure. We refer the reader to [1, 2, 6, 7, 10, 24, 25] for some recent results on forgetting of the initial condition. Finally, we discuss some examples in which the sufficient condition can be seen to hold.

Let $(S,\mathcal{S})$ be a Polish space, $\mathcal{S}$ being the Borel sigma-field on S. Let $X(\cdot)$ be an S-valued Markov process with0 CADLAG paths and $Y(\cdot)$ be given via (2.2), where, as before, $B(\cdot)$ is independent of $X(\cdot)$. As in section 2, we will consider the process $(X(\cdot),Y(\cdot))$ on the path space, $D(S\times\mathbb{R}^m;0,\infty)$, where m denotes the dimension of the observation vector. Let P_μ be the probability measure on the path space of $(X(\cdot),Y(\cdot))$ under which $X(0)$ has the distribution μ. With an abuse of notation, we will use the notation P_x when μ is a point mass concentrated at $x\in S$. If the filter is initialized at $\Pi(0)=\nu$ (whether or not it is the distribution of $X(0)$), we denote the process by $\Pi^\nu(\cdot)$. Following the notation of section 2, $\Pi^\nu(\cdot)$ is defined by

$$\langle\Pi^\nu(t),\phi\rangle=\frac{E_{\nu,Y_{0,t}}\left[\phi(\tilde X(t))R(\tilde X_{0,t},Y_{0,t})\right]}{E_{\nu,Y_{0,t}}\left[R(\tilde X_{0,t},Y_{0,t})\right]},$$

where $\phi(\cdot)$ is a bounded and measurable real-valued function, and $\tilde X(\cdot)$ is as in section 2. Recall that $\Pi^\nu(0)=\nu$ is the distribution of $\tilde X(0)$.

Definition 7.1. *We say that the "filter forgets its initial condition" if for all $\nu_1,\nu_2\in\mathcal{M}(S)$, for all $x\in S$ and continuous bounded real valued functions $\phi(\cdot)$ on S:*

$$|\langle\Pi^{\nu_1}(t),\phi\rangle-\langle\Pi^{\nu_2}(t),\phi\rangle|$$

converges to zero in P_x-probability as $t\to\infty$.

Theorem 7.1. *Suppose that $X(t)$ has a unique invariant measure and the filter forgets its initial condition. Then the pair $(X(\cdot),\Pi(\cdot))$ has at most one invariant measure.*

Proof. Suppose ρ_1 and ρ_2 are two invariant measures for the pair $(X(\cdot),\Pi(\cdot))$. We will show that for all continuous and bounded functions f on $S\times\mathcal{M}(S)$

$$\int f(x,\alpha)\rho_1(dx,d\alpha)=\int f(x,\alpha)\rho_2(dx,d\alpha), \tag{7.1}$$

where (x,α) denotes a generic element of $S\times\mathcal{M}(S)$. Let μ denote the unique invariant measure of $X(\cdot)$ and let μ_1,μ_2 be regular conditional probability functions such that

$\rho_i(dx, d\alpha) = \mu_i(x, d\alpha)\mu(dx)$; $i = 1, 2$. The left side of (7.1) equals, for all $t > 0$,

$$\int \left[\int E_x \left[f(X(t), \Pi^\alpha(t)) \right] \mu_1(x, d\alpha) \right] \mu(dx),$$

while the right hand side of (7.1) equals, for all $t > 0$,

$$\int \left[\int E_x \left[f(X(t), \Pi^\beta(t)) \right] \mu_2(x, d\beta) \right] \mu(dx).$$

Thus

$$\begin{aligned} &\left| \int f(x,\alpha)\rho_1(dx,d\alpha) - \int f(x,\alpha)\rho_2(dx,d\alpha) \right| \\ &\quad \le \int \int \int E_x \left| f(X(t), \Pi^\alpha(t)) - f(X(t), \Pi^\beta(t)) \right| \mu_1(x, d\alpha)\mu_2(x, d\beta)\mu(dx). \end{aligned} \tag{7.2}$$

The assumption on forgetting the initial condition implies that for all $x \in S$ and $\alpha, \beta \in \mathcal{M}(S)$, as $t \to \infty$,

$$E_x \left| f(X(t), \Pi^\alpha(t)) - f(X(t), \Pi^\beta(t)) \right| \to 0.$$

Now an application of dominated convergence theorem shows that the right side of (7.2) converges to 0 as $t \to \infty$. This proves (7.1). □

Remark. Although we have stated and proved the above theorem for a signal evolving continuously in time and observations taken continuously, the result holds for discrete time models such as in section 5, where the signal can evolve either continuously or discretely in time, but the observations are taken at discrete time instants. (The paper [6] also allowed a point process observation model.) The proof, excepting some minor notational changes, remains unchanged.

We now give some examples, in the following two theorems, where the sufficient condition in the above theorem is seen to hold. The theorems below are fairly straightforward consequences of the results proved in Atar and Zeitouni [2] and LeGland [10], in view of which details of the proofs are omitted.

THEOREM 7.2. *Let the signal $X(\cdot)$ (indexed either by a continuous or discrete time parameter) be a Feller time homogeneous Markov process taking values in a compact Polish space S. Suppose that the observations are in discrete time and given by* (5.1)*, Suppose that there exists a finite measure μ on $(S, \mathcal{S})$ and an integer $n \geq 1$ such that for all $x \in S$:*

$$c_1\mu(d\chi) \le P(d\chi, n|x) \le c_2\mu(d\chi), \tag{7.2}$$

where $P(d\chi, n|x)$ is the n-step transition probability function for $X(\cdot)$ and c_1, c_2 are some finite positive constants. Then the process $\{X(k), \Pi(k), k < \infty\}$ has at most one stationary measure.

Remarks on the Proof. The proof is based on the properties of Hilbert's projective metric (see [4]) on the space of finite measures. It can be shown that for x, ν_1, ν_2 as in the beginning of the section, the Hilbert distance between $\Pi^{\nu_1}(m)$ and $\Pi^{\nu_2}(m)$ converges to zero a.s. P_x as $m \to \infty$. This fact is proved for $n = 1$ in [2] and for $n \geq 1$, in the special case of finite state Markov chains in [10]. The case of general compact state space follows by similar arguments. The convergence in the Hilbert metric implies the convergence in total variation norm which in turn implies the property of the filter forgetting its initial condition in the sense of Definition 7.1.

Since $X(\cdot)$ is Feller with a compact state space, it has at least one invariant measure. Finally, if μ_1, μ_2 are two such invariant measures, we have, using the properties of Hilbert metric and (7.2), that the Hilbert distance between $\int P(\cdot, n|x)\mu_1(dx)$ and $\int P(\cdot, n|x)\mu_2(dx)$ converges to zero as $n \to \infty$. This implies that $\mu_1 = \mu_2$ and therefore Theorem 7.1 can be applied.

Theorem 7.3. *Let the signal be given via* (2.1) *and suppose that it takes values in a compact Riemannian manifold. Suppose that* $p(\cdot)$ *and* $\sigma(\cdot)$ *are Lipschitz continuous and that the generator of the Markov process is strictly elliptic. Assume that the observation process is given via* (2.2) *with* $g(\cdot)$ *a twice continuously differentiable real-valued function. Then* $(X(\cdot), \Pi(\cdot))$ *has a unique stationary measure.*

Remarks on the Proof. Arguments as in Theorems 3.1 and 3.2 show that $(X(\cdot), \Pi(\cdot))$ has at least one invariant measure. Furthermore, upper and lower bounds on the transition probability kernel of a Markov diffusion process on a compact manifold (see [2]) yield that if μ_1 and μ_2 are two invariant measures for $X(\cdot)$, then the Hilbert distance between $\int P(\cdot, t|x)\mu_1(dx)$ and $\int P(\cdot, t|x)\mu_2(dx)$ converges to 0 as $t \to \infty$, where $P(\cdot, t|x)$ is the transition probability kernel of $X(\cdot)$. This proves the uniqueness of the invariant measure for $X(\cdot)$. Let $h(\cdot)$ denote the Hilbert distance between the measures. Next, the results in [2] show that, for the additive white noise corrupted observation case, where x, ν_1, ν_2 are as in the beginning of this section and satisfy $h(\nu_1, \nu_2) < \infty$, a.s. P_x, we have that $h(\Pi^{\nu_1}(t), \Pi^{\nu_2}(t))$ converges to zero a.s. P_x as $t \to \infty$. The constraint, $h(\nu_1, \nu_2) < \infty$, can be removed using techniques similar to those of [7, Theorem 3.2] on first proving that the P_x-probability of the event, $h(\Pi^{\nu_1}(t), \Pi^{\nu_2}(t)) < \infty$ for some t, is one.

REFERENCES

[1] R. Atar, *Exponential stability for nonlinear filtering of diffusion processes in a noncompact domain*, Ann. Probab., 26 (1998), pp. 1552–1574.

[2] R. Atar and O. Zeitouni, *Exponential stability for nonlinear filtering*, Ann. Inst. H. Poincaré Probab. Statist., 33 (1997), pp. 697–725.

[3] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[4] G. Birkhoff, *Lattice Theory*, 3rd ed., AMS, Providence, RI, 1967.

[5] G. Blankenship and G.C. Papanicolaou, *Stability and control of stochastic systems with wide-band noise disturbances*, I, SIAM J. Appl. Math., 34 (1978), pp. 437–476.

[6] A. Budhiraja and H. J. Kushner, *Robustness of nonlinear filters over the infinite time interval*, SIAM J. Control Optim., 36 (1998), pp. 1618–1637.

[7] A. Budhiraja and D. Ocone, *Exponential stability of discrete-time filters for bounded observation noise*, Systems Control Lett., 30 (1997), pp. 185–193.

[8] R. Elliot, *Stochastic Calculus and Applications*, Springer-Verlag, Berlin, New York, 1982.

[9] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.

[10] F. Le Gland and L. Mevel, *Exponential forgetting and geometric ergodicity in hidden Markov models*, Math. Control Signals Systems, to appear.

[11] G. Kallianpur, H. Fujisaki, and H. Kunita, *Stochastic differential equations for the nonlinear filtering problem*, Osaka J. Math., 9 (1972), pp. 19–40.

[12] H. Kunita, *Asymptotic behavior of the nonlinear filtering errors of a Markov process,* J. Multivariate Anal., 1 (1971), pp. 365–393.

[13] T. G. Kurtz, *Semigroups of conditional shifts and approximation of Markov processes*, Ann. Probab., 4 (1975), pp. 618–642.

[14] T. G. Kurtz, *Approximation of Population Processes*, CBMS-NSF Regional Conference Series in Applied Mathematics 36, SIAM, Philadelphia, PA, 1981.

[15] H. J. Kushner, *Dynamical equations for nonlinear filtering*, J. Differential Equations, 3 (1967), pp. 179–190.

[16] H. J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.

[17] H. J. Kushner, *Jump-diffusion approximations for ordinary differential equations with wide-band random right-hand sides*, SIAM J. Control Optim., 17 (1979), pp. 729–744.

[18] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.

[19] H. J. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Systems & Control: Foundations & Applications 3, Birkhäuser, Boston, MA, 1990.

[20] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, Berlin, New York, 1992.

[21] H. J. Kushner and H. Huang, *Approximation and limit results for nonlinear filters with wide bandwidth observation noise*, Stochastics Stochastics Rep., 16 (1986), pp. 65–96.

[22] H. J. Kushner and L. F. Martins, *Limit theorems for pathwise average cost per unit time problems for queues in heavy traffic*, Stochastics Stochastics Rep., 42 (1993), pp. 25–51.

[23] R. Liptser and A.N. Shiryaev, *Statistics of Random Processes*, Springer-Verlag, Berlin, New York, 1977.

[24] D. Ocone, *Asymptotic stability of Benes filters*, Stochastic Anal. Appl., to appear.

[25] D. Ocone and E. Pardoux, *Asymptotic stability of the optimal filter with respect to its initial condition*, SIAM J. Control Optim., 34 (1996), pp. 226–243.

[26] G. C. Papanicolaou, D. Stroock, and S. R. S. Varadhan, *Martingale approach to some limit theorems*, in Papers from the Duke Turbulence Conference, Duke Univ. Math. Ser. III, Duke University, Durham, NC, 1977, paper 6.

[27] R. Rishel, *Necessary and sufficient dynamic programming conditions for continuous time stochastic optimal control*, SIAM J. Control, 8 (1970), pp. 559–571.

[28] L. Stettner, *On invariant measures of filtering processes*, in Stochastic Differential Systems, Proceedings 4th Bad Honnef Conference, Lecture Notes in Control and Inform Sci. 126, K. Helmes, N. Christopeit, and M. Kohlmann, eds., Springer, Berlin, New York, 1989, pp. 279–292.